# Évaluation des méthodes d'intégration de données permettant d'améliorer la prédiction spatiale des épidémies de grippe

Marie Morvan

**AGROCAMPUS OUEST**

☐ **CFR Angers**

☑ **CFR Rennes**

| | |
|---|---|
| Année universitaire : 2015 - 2016<br>Spécialité : Agronomie<br>Spécialisation (et option éventuelle) :<br>Statistique Appliquée | **Mémoire de fin d'études**<br>☑ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage<br>☐ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage<br>☐ d'un autre établissement (étudiant arrivé en M2) |

# Evaluation des méthodes d'intégration de données permettant d'améliorer la prédiction spatiale des épidémies de grippe

Marie MORVAN

***Soutenu à Rennes, le 6 septembre 2016***

***Devant le jury composé de :***

Président :

Maître de stage : Clément TURBELIN

Enseignant référent : David CAUSEUR

# Fiche de confidentialité et de diffusion du mémoire

## Confidentialité

☐ Non ☐ Oui si oui : ☐ 1 an ☐ 5 ans ☐ 10 ans

Pendant toute la durée de confidentialité, aucune diffusion du mémoire n'est possible (1).
Date et signature du maître de stage (2) :

**A la fin de la période de confidentialité,** sa diffusion est soumise aux règles ci-dessous (droits d'auteur et autorisation de diffusion par l'enseignant à renseigner).

## Droits d'auteur

L'auteur(3) **Morvan Marie**
autorise la diffusion de son travail (immédiatement ou à la fin de la période de confidentialité)

☐ Oui ☐ Non

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☐ la diffusion papier et électronique du mémoire (joindre dans ce cas la fiche de conformité du mémoire numérique et le contrat de diffusion)

(Facultatif) ☐ accepte de placer son mémoire sous licence Creative commons CC-By-Nc-Nd (voir Guide du mémoire Chap 1.4 page 6)

Date et signature de l'auteur :

## Autorisation de diffusion par le responsable de spécialisation ou son représentant

L'enseignant juge le mémoire de qualité suffisante pour être diffusé (immédiatement ou à la fin de la période de confidentialité)

☐ Oui ☐ Non

Si non, seul le titre du mémoire apparaîtra dans les bases de données.
Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☐ la diffusion papier et électronique du mémoire

Date et signature de l'enseignant :

(1) L'administration, les enseignants et les différents services de documentation d'AGROCAMPUS OUEST s'engagent à respecter cette confidentialité.
(2) Signature et cachet de l'organisme
(3).Auteur = étudiant qui réalise son mémoire de fin d'études
(4) La référence bibliographique (= Nom de l'auteur, titre du mémoire, année de soutenance, diplôme, spécialité et spécialisation/Option)) sera signalée dans les bases de données documentaires sans le résumé

Présentation de l'organisme d'accueil du stage :

Le réseau Sentinelles est un système de surveillance épidémiologique reposant sur un réseau de médecins généralistes et pédiatres en France métropolitaine. Ce réseau est coordonné par l'équipe "Surveillance et Modélisation des maladies transmissibles" de l'Institut Pierre Louis d'Epidémiologie et de Santé Publique (UMR S 1136) de l'Inserm (Institut National Supérieur de la Recherche Médicale) et de l'Université Pierre et Marie Curie, en collaboration avec Santé Publique France. Ce système national de surveillance permet le recueil, l'analyse, la prévision et la redistribution en temps réel de données épidémiologiques issues de l'activité des médecins généralistes. Le réseau Sentinelles collecte de façon continue des informations sur 8 indicateurs de santé (sept maladies infectieuses et un indicateur non-infectieux). Les cartes, les séries chronologiques et les tableaux concernant ces indicateurs sont disponibles pour tout utilisateur. De plus, un bulletin d'information est diffusé chaque semaine.

Remerciements :

Un grand merci à Clément Turbelin et Pierre-Yves Boëlle qui ont encadré ce stage, pour leur confiance me permettant de mener ce projet, leur grande aide et leurs suggestions. Je les remercie d'avoir été disponibles malgré leurs nombreuses missions.

Merci à Cécile Souty d'avoir répondu à mes questions avec précision. Merci pour sa disponibilité et son aide précieuse concernant ce stage et la rédaction de ce mémoire, mais aussi pour les nombreux conseils concernant mon projet professionnel.

J'adresse également mes remerciements à toute l'équipe du réseau Sentinelles, pour m'avoir accueillie chaleureusement, Thierry Blanchon - responsable adjoint du réseau Sentinelles - Ana, Caroline, Thibaud, Yves, Victoire, Louise. Cette expérience professionnelle et personnelle a été très enrichissante et je remercie toutes les personnes qui m'ont aidée directement ou indirectement.

Liste des abréviations :

RS : Réseau Sentinelles
MG : Médecin généraliste
MS : Médecin Sentinelles
SG : Syndromes grippaux
PLSR : Régression des moindres carrés partiels (Partial Least Square Regression)
OLS : Méthode des moindres carrés ordinaire (Ordinary Least Squares)
WLS : Méthode des moindres carrés pondérés (Weighted Least Squares)
VL : Variable Latente

Liste des figures :



Liste des tableaux :



Liste des annexes :

# INTRODUCTION

La surveillance épidémiologique est définie par l'OMS comme « la collecte continue et systématique, l'analyse et l'interprétation de données médicales nécessaires pour la planification, la mise en place et l'évaluation des pratiques de santé publiques » (http://www.who.int/topics/public_health_surveillance/en/). Une telle surveillance est utilisée dans un contexte de veille sanitaire et permet notamment de rendre compte d'un changement de situation concernant l'état de santé d'une population ou l'apparition d'un risque et de mettre en place des actions de santé publique.

Pour les maladies qui ne font pas l'objet de déclarations obligatoires, il existe des systèmes d'information basés sur des réseaux de médecins généralistes (MG) volontaires et souvent bénévoles permettant le recueil de données sur des échantillons de population (Deckers et al., 2006). Ces réseaux de surveillance épidémiologique sont présents dans de nombreux pays européens, et permettent l'obtention de grandes bases de données (Souty et al., 2014). Cependant, une couverture imparfaite de la population à surveiller est un problème rencontré lors de l'utilisation de ce type de système. En effet, l'échantillon de médecins collectant les données ne peut pas être contrôlé, et varie d'une semaine à l'autre ou d'une zone à l'autre, et la répartition de ces médecins sur le territoire étudié n'est pas forcément homogène (Schlaud, 1999) (Annexe I).

Pour surveiller les maladies fréquentes dans la communauté, la mesure d'incidence, qui représente le nombre de nouveaux cas sur le territoire étudié, est couramment utilisée. Etudier spatialement le taux d'incidence d'une maladie, notamment infectieuse, permet d'observer sa diffusion sur le territoire (Carrat et Valleron, 1992).

La méthode d'interpolation spatiale de krigeage permet de prédire en tout point d'un territoire une variable qui n'est observée que pour un échantillon de points, en s'appuyant sur la structure de dépendance spatiale des observations (Matheron, 1963). Cette méthode d'interpolation permet l'estimation de la répartition spatiale des incidences d'une maladie sur un territoire. Cependant, cette estimation est très dépendante de la répartition des déclarations des médecins sur le territoire. En effet, en interpolation spatiale, si le nombre de zones renseignées est faible, l'estimation résultante de la répartition des cas risque d'être peu précise ou biaisée (Baillargeon, 2005).

Face à ce problème de manque de couverture spatiale rencontré dans les systèmes de surveillance en soin primaires, des stratégies peuvent être mises en place pour améliorer l'estimation de l'incidence d'une maladie. Les données auxiliaires peuvent notamment être utilisées en parallèle des systèmes traditionnels pour la surveillance épidémiologique comme les données de requêtes sur les moteurs de recherche comme Google (Pelat et al., 2009) ou les données de délivrances médicamenteuses. Par exemple,Vergu et al. (2006) ont montré que les volumes de délivrance de certaines classes de médicaments utilisées pour le traitement de syndromes grippaux peuvent prédire l'évolution de leurs incidences nationales durant les épidémies de grippe en France. Cependant, ces données auxiliaires sont aujourd'hui peu intégrées aux systèmes de surveillance au niveau spatial.

L'objectif de mon stage est d'évaluer les méthodes d'intégration de données auxiliaires permettant d'améliorer la prédiction spatiale des épidémies de grippe à partir de données collectées par un système de surveillance en soins primaires. Un deuxième objectif est d'intégrer la dimension temporelle dans l'interpolation spatiale. La problématique ici est la

suivante : l'utilisation de données auxiliaires permet-elle d'obtenir des prédictions spatiales plus robustes des incidences de syndromes grippaux ?

Afin de répondre à cette problématique, nous nous appuierons sur les données de surveillance épidémiologiques du réseau Sentinelles (RS). Depuis 1984, ce réseau collecte des données en temps réel sur 8 indicateurs santé, et permet l'analyse et la redistribution de données épidémiologiques. Parmi ces indicateurs, les syndromes grippaux (SG) font l'objet d'une surveillance importante en termes de période d'étude et de nombre de publications. Le krigeage est appliqué sur les données du réseau Sentinelles (Carrat et Valleron, 1992) et permet la construction de cartes de répartition des SG. Les données auxiliaires utilisées seront les données de délivrances médicamenteuses.

Après un point méthodologique expliquant les méthodes de krigeage, les analyses statistiques menées seront détaillées, puis les résultats seront exposés et discutés.

# 1. MATERIELS ET METHODES

## 1.1. LES DONNEES

### 1.1.1. LES DONNEES D'INCIDENCE DE SYNDROMES GRIPPAUX RECUEILLIES PAR LE RS

L'estimation de l'incidence des SG au RS repose sur les données transmises par des médecins généralistes libéraux bénévoles inscrits au RS, les médecins Sentinelles (MS). Ils fournissent des données en temps réel concernant les patients vus en consultation, via un site sécurisé dédié aux MS ou via un logiciel client jSentinel (Turbelin et Boëlle, 2010). Un médecin déclare un cas de SG quand un de ses patients présente les symptômes suivants : apparition brutale de fièvre supérieure à 39°C, accompagnée de myalgies et de signes respiratoires (Turbelin et al., 2013). Pour chaque cas décrit, des informations supplémentaires peuvent aussi être fournies sur le patient (ces informations ne permettent pas de l'identifier).

Le calcul de l'incidence et du taux d'incidence des SG a pour objectif d'estimer le nombre total de cas vus en consultation par l'ensemble des médecins généralistes français à partir des données transmises par les MS.

Les données recueillies correspondent au nombre de cas de SG vus par chaque MS pour une période donnée. Un prétraitement de ces données consiste à calculer la participation hebdomadaire de chaque médecin et le nombre de cas affectés à chaque semaine. Après ce prétraitement des données, l'incidence hebdomadaire est estimée en deux étapes : d'abord l'estimation du nombre moyen de cas par médecin à partir des données des médecins du réseau d'une zone géographique puis l'estimation du nombre total de cas en extrapolant l'information recueillie auprès des médecins du réseau à l'ensemble des médecins français de cette même zone. Les estimations d'incidences sont dans un premier temps effectuées par zone (département ou région), puis globalement au niveau national. Spatialement, l'estimation d'incidence d'une zone est attribuée à son centroïde. On travaillera par la suite sur le taux d'incidence de SG pour 100 000 habitants. Cela permet de comparer des zones ayant des populations différentes, ou des dates différentes, et de travailler sur l'ensemble du territoire français avec les données régionales/départementales. Les données disponibles pour chaque semaine sont illustrées par le *Tableau 1*.

*Définition et détection des épidémies de grippes*

Une épidémie est définie comme l'augmentation rapide de l'incidence, c'est-à-dire du nombre de nouveau cas de grippe sur un période donnée. L'épidémie est déclarée lorsque l'incidence dépasse

le seuil épidémique correspondant au nombre de cas auquel on s'attend en l'absence de circulation du virus grippal. Ce niveau de base prend en compte la saisonnalité de la grippe, les épidémies ayant généralement lieu entre octobre et avril. Il est estimé par une régression périodique appliquée sur les incidences observées dans le passé, hors épidémie (Costagliola et al., 1991).

**Tableau 1 : Données d'incidence des syndromes grippaux du réseau Sentinelles.**
Inc100 est le taux d'incidence de SG pour 100 000 habitants. Les variables x et y correspondent aux coordonnées en Lambert 93 du centroïde de la zone observée. « 201101 » correspond à la première semaine de l'année 2011. Zone.level est le niveau géographique étudié (départemental ou régional).

| x | y | Semaine | Zone | Zone.level | Inc100 |
|---|---|---|---|---|---|
| **881423.2** | 6558217 | 201101 | 1 | DEP | 1307 |
| **740407.2** | 6940171 | 201101 | 2 | DEP | 904 |
| **…** | … | … | … | … | … |
| **636534.6** | 6887342 | 201101 | 96 | DEP | 853 |

### 1.1.2. Les donnees auxiliaires de ventes de medicaments

Depuis 1999, le réseau Sentinelles et la société IMS-Health France entretiennent un partenariat de recherche non financier. IMS-Health fournit chaque lundi au RS une base de données contenant une estimation des volumes de délivrance de médicaments en ville, avec ou sans ordonnance, de la semaine précédente. Ces données concernent plus de 500 classes de médicaments, identifiées par leur code ATC4 (Anatomical Therapeutic Chemical) donné par l'EphMRA (European Pharmaceutical Marketing Research Association). Dans cette classification, les médicaments sont regroupés en classes selon l'organe ou le système sur lequel ils agissent, et/ou selon leurs propriétés chimiques, pharmacologiques et thérapeutiques (http://www.whocc.no/atc/structure_and_principles/).

Ces données sont disponibles depuis 2011 au niveau départemental.

L'estimation des volumes de délivrance est réalisée par la société IMS-Health France, et s'appuie sur un recueil d'information en temps réel dans un échantillon de 14 000 pharmacies environ, correspondant à 60% des pharmacies françaises, ce qui en fait des données robustes (https://www.ims-pharmastat.fr/pharmastat).

Face à un SG, la population peut avoir recours à l'automédication avant d'aller voir un médecin. Les volumes de délivrance de certaines classes de médicaments peuvent traduire l'évolution des SG en France. Une sélection a priori de 19 classes a été réalisée à partir d'une liste de médicaments susceptibles d'être prescrits en cas de SG par un groupe d'experts de l'OMS. La corrélation entre chaque classe de cette sélection et l'incidence des SG a d'ailleurs été montrée par Vergu et al. (2006).

Les données disponibles sont le nombre d'unité de chacune des classes médicamenteuses délivrées par semaine dans chaque zone considérée. Ce nombre d'unité est exprimé en taux pour 100 000 habitants. La description des classes médicamenteuses est disponible en annexe II.

### 1.1.3. Analyse de donnees a reference spatiale

Dans les données étudiées, un individu statistique correspond à un centroïde d'un département, donc à un point sur la carte française, pour lequel on dispose des coordonnées (latitude et longitude, selon la projection Lambert93) et des valeurs prises pour chaque variable. Dans cette situation, les observations ne sont pas indépendantes car deux localisations proches se ressemblent plus que deux localisations éloignées. On parle alors de données spatiales (Bivand et al., 2008).

On a donc sur un espace géographique délimité, un ensemble de points répartis de façon non régulière associés à une « mesure d'incidence ». On souhaite savoir ce qu'il en est sur l'ensemble du territoire, pour lequel on ne dispose pas de mesure. L'objectif de l'interpolation spatiale est de prédire la valeur d'incidence en tout point de la carte française à partir des valeurs connues, rapportées aux centroïdes des départements.

Finalement, les données disponibles sont représentées dans le *Tableau 2*. La période allant des semaines 201101 à 201601 sera étudiée, ajoutant une dimension temporelle à l'analyse.

**Tableau 2 : Données utilisées pour l'interpolation spatiale.**
Ce type de tableau est disponible pour toutes les semaines depuis 201101. Les variables A11G1 à R05F correspondent aux noms des classes médicamenteuses.

| x | y | Semaine | Zone | Zone.level | A11G1 | … | R05F | Inc100 |
|---|---|---|---|---|---|---|---|---|
| **881423.2** | 6558217 | 201101 | 1 | DEP | 168.469 | | 234.303 | 1307 |
| **740407.2** | 6940171 | 201101 | 2 | DEP | 160.774 | | 84.746 | 904 |
| **…** | … | … | … | … | … | | … | … |
| **636534.6** | 6887342 | 201101 | 96 | DEP | 201.217 | | 181.348 | 853 |

## 1.2. Principe de la methode – Le krigeage

### 1.2.1. Le krigeage utilise en routine par le reseau Sentinelles

Le krigeage est une méthode d'interpolation spatiale permettant d'estimer la valeur d'une mesure en un point de l'espace non observé à partir d'un échantillon de mesures.

Le krigeage permet de prévoir la valeur de la variable Z en un point non observé $s_p$ de coordonnées $(x_p, y_p)$ en utilisant les valeurs observées aux points $(s_1,\dots,s_n)$ selon : $z(s_p) = \sum_{i=1}^{n} \lambda_i . z(s_i) + \varepsilon$ .

Le prédicteur linéaire est définit par : $\hat{Z}(s_p) = \sum_{i=1}^{n} \lambda_i . Z(s_i)$ , avec $\sum_{i=1}^{n} \lambda_i = 1$.
Les poids $\lambda_i$ associés à chacune des valeurs observées dépendent de la distance entre les observations et de leur variabilité spatiale. L'analyse variographique détaillée ci-dessous permet de les calculer.

Sous l'hypothèse de stationnarité intrinsèque on a :
$E\big(Z(s+h) - Z(s)\big) = 0$
$Var\big(Z(s+h) - Z(s)\big) = E[(Z(s+h) - Z(s))^2] = 2\gamma(h), \qquad avec\ s\ et\ s+h \in D$
γ est appelée semi-variogramme et mesure la variabilité entre les valeurs en fonction de leur éloignement. γ ne dépend que de la distance euclidienne h entre deux points.

Le semi-variogramme expérimental est calculé à partir des données selon l'estimateur suivant : $\hat{\gamma}(\mathrm{h}) = \frac{1}{2N(\mathrm{h})} \sum_{i=1}^{N(\mathrm{h})} (Z(s_i + \mathrm{h}) - Z(s_i))^2$ . $Z(s_i)$ et $Z(s_i + h)$ sont des valeurs observées aux points $s_i$ et $s_i + h$ et N(h) est le nombre de paires de points séparés par une distance h. En pratique, pour un maillage irrégulier, des intervalles de distance contenant un ensemble de points sont considérés.

Une fonction continue est ensuite ajustée au semi-variogramme expérimental. Cette étape est la plus difficile et a un impact important sur le résultat final. Les 4 paramètres de forme de modèle, palier, portée et effet pépite détaillés en *figure 1* doivent être déterminés.

Les poids $\lambda_i$ sont ensuite calculés à partir du semi-variogramme théorique γ selon :

$$\begin{cases} \sum_{i=1}^{n} \lambda_i \gamma(h_{i,j}) + l = \gamma(h_{j,p}) \; pour \; j = 1, \dots, n \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases}$$

Avec $h_{i,j}$ la distance entre deux observations $s_i$ et $s_j$, $l$ multiplicateur lagrangien.

Avec $N = \begin{pmatrix} \gamma(0) & \dots & \gamma(h_{1,n}) & 1 \\ \vdots & \ddots & \vdots & 1 \\ \gamma(h_{n,1}) & \dots & \gamma(h_{n,n}) & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$, $b = \begin{pmatrix} \gamma(h_{1,p}) \\ \vdots \\ \gamma(h_{n,p}) \\ 1 \end{pmatrix}$ et $\lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ l \end{pmatrix}$ on a $\lambda = N^{-1} \cdot b$

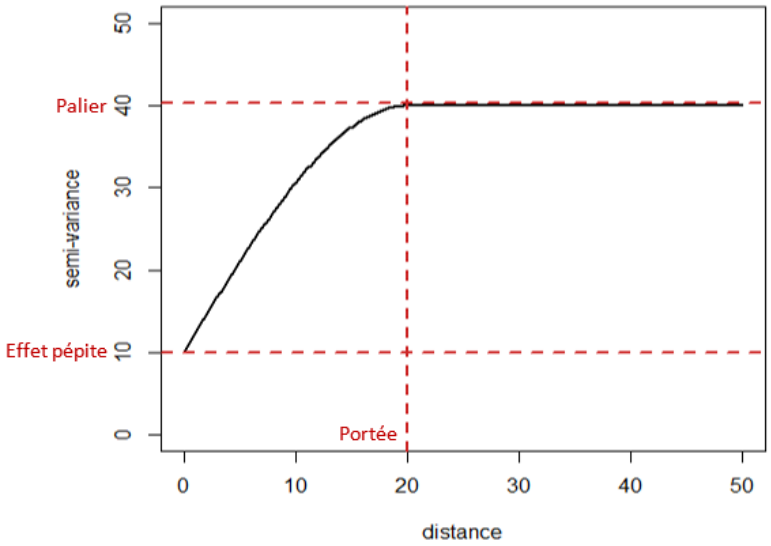



**Figure 1 : Exemple de modèle de variogramme**
Avec les paramètres : forme de modèle Sphérique, Palier = 40, portée = 20, effet pépite = 10.

En tant que méthode stochastique, le krigeage permet d'obtenir des erreurs d'estimation. C'est la seule méthode qui prend en compte la structure de dépendance spatiale des données. De plus, les prévisions obtenues sont non biaisées. Le krigeage ordinaire est utilisé par le réseau Sentinelles pour réaliser chaque semaine une carte montrant la répartition spatiale des SG. Cette carte est obtenue à partir des données de taux d'incidence départementaux. Cette méthode permet de prendre en compte la communication entre les observations de SG et la propagation des cas dans l'espace (Carrat et Valleron, 1992).

### 1.2.2. Les extensions du krigeage : le cokrigeage et le krigeage spatio-temporel

En cokrigeage, des variables auxiliaires sont intégrées pour améliorer l'estimation de la variable principale. La variable d'intérêt est une combinaison linéaire pondérée des observations de la variable à interpoler et des variables auxiliaires. L'estimateur de cokrigeage est : $\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i) + \sum_{j=1}^{L} \beta_j W(s_j)$ sous la contrainte $\sum_{i=1}^{N} \lambda_i = 1$ et $\sum_{j=1}^{L} \beta_j = 0$.

Un variogramme est calculé pour chaque variable, ainsi qu'un variogramme croisé qui décrit la structure de dépendance entre la variable principale et la variable auxiliaire. On a l'estimateur suivant : $\hat{\gamma}_{zw}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z_z(s_i) - Z_z(s_i + h)][Z_w(s_i) - Z_w(s_i + h)]$ . L'analyse variographique est effectuée de la même façon que précédemment : après avoir relevé ses propriétés, une fonction est ajustée au variogramme expérimental.

Le krigeage spatio-temporel permet de prédire en un point $s_p$ et au temps $t$ la variable $Z(s_p, t)$ observée aux points $(s_1, t_1)$, …, $(s_n, t_n)$ avec le prédicteur linéaire : $\hat{z}(s,t) = \sum_{i=1}^{n} \lambda_i z(s_i, t)$ sous la contrainte $\sum_{i=1}^{K} \lambda_i = 1$ en tout point $s_i$ et temps $t$.

### 1.3. ANALYSES STATISTIQUES

L'objectif principal des analyses réalisées est d'améliorer les estimations spatiales des SG en intégrant les données externes dans le krigeage déjà réalisé en routine. Les apports de l'inclusion d'une dimension temporelle dans l'analyse sont aussi étudiés. Dans un premier temps, le travail est purement spatial et les variables auxiliaires sont intégrées grâce au cokrigeage selon deux possibilités. D'une part, la variable la plus corrélée avec l'incidence est utilisée et d'autre part, une variable latente qui résume l'ensemble de nos variables auxiliaires est construite. Le krigeage spatio-temporel permet ensuite d'inclure la dimension temporelle à l'analyse. La sélection du meilleur modèle est une question importante, car elle ne peut pas être réalisée en minimisant les erreurs d'estimation, les vraies valeurs d'incidence n'étant jamais connues. Différentes propriétés et critères sont donc évalués pour vérifier un modèle.

Toutes les analyses ont été réalisées avec le logiciel R.

#### 1.3.1. CONSTRUCTION DU MODELE DE VARIOGRAMME

Le modèle de krigeage appliqué aux taux d'incidence est étudié en premier lieu, avant l'ajout des variables auxiliaires. Cela permet de connaître la structure de dépendance de la variable à interpoler, et d'avoir un ordre de grandeur des paramètres d'ajustement du variogramme théorique. Le krigeage sert aussi de base de comparaison pour évaluer les avantages du cokrigeage (Rossiter, 2007). De nombreux choix doivent être faits lors de la mise en place du krigeage, et conditionnent les résultats obtenus.

##### 1.3.1.1. CARACTERISTIQUES DES DONNEES A PRENDRE EN COMPTE

Le choix des intervalles de distances servant au calcul du semi-variogramme a un impact important sur la qualité d'ajustement (Bivand et al., 2008). Le premier intervalle de distance choisi doit être représentatif de la réalité du territoire et ne doit pas seulement prendre en compte les points exceptionnellement proches, et non représentatifs des données étudiées. La distance entre les deux points voisins les plus éloignés est utilisée pour construire le premier intervalle. Un premier intervalle trop large supprimerait l'effet de dépendance spatiale entre les points proches.

Le RS estime des valeurs d'incidence aux niveaux régional et départemental. Les données régionales sont précises et n'ont pas de valeurs manquantes, alors que les données départementales peuvent être plus bruitées et incomplètes. Les données utilisées pour le krigeage au RS sont les données départementales. Dans notre étude, un centroïde pour lequel aucune mesure n'est disponible est considéré comme un point à interpoler. Les analyses sont menées sur les données régionales d'une part et départementales d'autre part, sur le territoire de France métropolitaine hors Corse.

Un krigeage local est réalisé, car pour estimer la valeur en un point, seule une partie des observations voisines de ce point est utilisée. Actuellement, le krigeage réalisé au RS prend en compte les 9 voisins les plus proches lors de l'estimation de l'incidence en un point. Cela permet de diminuer les temps de calcul, et est plus représentatif du phénomène étudié. Cette hypothèse paraît plus raisonnable que d'imposer une dépendance spatiale très distante pour modéliser la répartition de syndromes grippaux.

#### 1.3.1.2. LES PARAMETRES DU VARIOGRAMME THEORIQUE

Plusieurs méthodes peuvent être utilisées pour estimer les paramètres du modèle de variogramme qui s'ajuste au variogramme expérimental. Ces méthodes comprennent l'ajustement visuel, l'utilisation d'algorithmes OLS/WLS, l'estimation par maximum de vraisemblance et l'utilisation de méthodes bayésiennes (Ribeiro Jr et Diggle, 2001). Ces deux dernières méthodes ne sont pas utilisées dans notre étude. De nombreux auteurs conseillent d'initialiser les paramètres du modèle de variogramme, et de vérifier l'ajustement visuellement (Bivand et al. (2008), Baillargeon (2005), Goulard et Voltz (1992)). Cependant, l'ajustement du modèle visuellement chaque semaine n'est pas compatible avec une surveillance continue. Le variogramme théorique est donc ajusté par la méthode des moindres carrés pondérés au variogramme expérimental, à partir de paramètres fixés par défaut.

L'algorithme WLS doit être initialisé avec des valeurs qui permettent la convergence. Le palier correspond à la variance maximale atteinte, en général entre les points les plus éloignés. Les parties les plus difficiles concernent la portée, distance à partir de laquelle la variance entre les points est maximale, qui a une grande influence sur le résultat final et est difficile à déterminer automatiquement, et la forme de modèle. Il existe un ensemble de formes de modèles disponibles, pour lesquelles le calcul de krigeage a été vérifié et est possible, c'est-à-dire mène à une matrice de covariance définie positive. Les formes les plus classiques adaptées à notre situation sont les modèles sphérique, exponentiel, gaussien, de Matérn, de Stein et de Bessel. La portée est actuellement fixée à 200 km pour le krigeage réalisé au RS. L'effet pépite, qui représente la variabilité à micro échelle, est actuellement fixé à 0 arbitrairement. Une méthode de sous-échantillonnage est mise en place pour estimer ce paramètre pour chaque semaine (Annexe III).

Une procédure automatique qui minimise l'écart entre le variogramme expérimental et le variogramme théorique est mise en place pour calculer les paramètres de portée et de forme simultanément pour une semaine donnée. En réalisant cette procédure pour toute la période d'étude, les valeurs de portée et de forme sont obtenues pour l'ensemble des semaines. S'il n'y a pas de différences importantes entre les portées estimées sur l'ensemble de la période, une valeur moyenne sera utilisée pour initialiser l'ajustement du variogramme.

### 1.3.2. AJOUT DES VARIABLES AUXILIAIRES

Pour chaque variable auxiliaire, les modèles de variogrammes et de variogramme croisé avec l'incidence sont ajustés simultanément. Ces modèles doivent mener à des systèmes de cokrigeage définis positifs. La solution la plus courante et la plus facile à mettre en œuvre est d'ajuster un modèle linéaire de co-régionalisation (Emery, 2012). Dans ce cas, tous les modèles ont la même forme et la même portée, mais peuvent avoir différents effets pépite et différents paliers. Pour cela, les paramètres du modèle de variogramme de la variable principale sont utilisés comme modèle de départ.

#### 1.3.2.1. CRITERES D'UTILISATION DE VARIABLES AUXILIAIRES

Deux méthodes principales sont couramment utilisées pour sélectionner les données auxiliaires (Rossiter, 2007). Théoriquement, la connaissance du processus spatial qui cause la distribution des variables principales et auxiliaires permet de guider le choix du modèle. Empiriquement, examiner les corrélations spatiales et les covariances spatiales entre les variables peut permettre de sélectionner les données auxiliaires.

Ces données sont structurées de la même façon que l'incidence, ce sont des données spatiales appartenant au même champ d'étude, ce qui est nécessaire pour l'inclusion dans le modèle de krigeage (Rossiter, 2007). Plus particulièrement, ces données auxiliaires sont recueillies aux mêmes points de mesure que la variable principale.

Le lien entre la variable principale et les possibles variables auxiliaires est d'abord étudié. Le critère d'inclusion le plus utilisé est la corrélation spatiale entre les variables (Bivand et al., 2008) qui correspond à la corrélation entre les variables pour une semaine donnée. La structure spatiale de la variable auxiliaire est comparée avec celle de la variable principale avant de réaliser le cokrigeage. Pour cela, les variogrammes des variables sont ajustés séparément et comparés. La covariance spatiale entre les variables de médicaments et l'incidence est représentée par le variogramme croisé.

#### 1.3.2.2. SELECTION DE VARIABLES AUXILIAIRES

Pour chaque variable auxiliaire ajoutée, un variogramme simple et un variogramme croisé avec les autres variables du modèle sont calculés et doivent être modélisés. Si les 16 variables auxiliaires sont ajoutées, 153 variogrammes devront être ajustés. Il n'est donc pas envisageable d'inclure toutes les classes de médicaments dans le modèle de cokrigeage. Une sélection parmi les 16 classes de médicaments doit être faite. On se limitera par la suite à du cokrigeage avec une unique variable auxiliaire. Une stratégie peut être d'utiliser la variable la plus corrélée spatialement à la variable principale pour le cokrigeage (Rossiter, 2007).

Une autre stratégie peut consister en la construction d'une variable latente à partir des données auxiliaires. En effet, il y a de la redondance dans les données médicamenteuses et certaines classes sont très corrélées. De plus, une combinaison de médicaments est généralement prescrite contre les SG, plutôt qu'une seule classe. Il peut donc y avoir intérêt à résumer l'information portée par les 16 classes pour avoir une variable de cokrigeage qui résume l'ensemble des données de délivrances médicamenteuses. Une régression des moindres carrés partiels (PLSR) permet d'obtenir une ou plusieurs variable(s) latente(s) (VL) combinaison linéaire de l'ensemble des variables médicamenteuses. La VL obtenue peut être utilisée comme variable auxiliaire dans le modèle de cokrigeage. Une approche similaire a été mise en place par Sampson et al. (2013). Une PLSR de l'incidence sur les 16 classes médicamenteuses est donc réalisée et les VL résultantes sont intégrées au krigeage. Une VL est calculée pour chaque semaine, et une autre sur des portions de 16 semaines. Enfin, une variable latente est construite pour chaque semaine en utilisant toutes les variables de classes médicamenteuses ainsi qu'une variable correspondant à l'incidence des SG de la semaine précédant la semaine considérée. Ces trois modèles de cokrigeage avec variable latente seront étudiés.

### 1.3.3. ETUDES DES SERIES TEMPORELLES

A l'issue du krigeage, les données interpolées peuvent être reportées sur une carte pour chaque semaine. Cette carte correspond à une grille régulière pour laquelle chaque cellule a une valeur d'incidence estimée. Pour un point donné de la carte, il est possible de collecter toutes les valeurs d'incidence hebdomadaire estimées pour tracer la série temporelle d'incidence pour la zone géographique considérée. L'évolution de l'incidence des SG en fonction du temps sur plusieurs épidémies peut être analysée au niveau de zones ciblées.

Une série d'incidence nationale est ré-estimée à partir des estimations issues des différents modèles de krigeage et cokrigeage, en sommant toutes les estimations dans chaque zone. Ces estimations issues du cokrigeage sont comparées aux estimations nationales du RS.

### 1.3.4. INTEGRATION DE LA DIMENSION TEMPORELLE : KRIGEAGE SPATIO-TEMPOREL

La grippe évolue dans l'espace mais aussi dans le temps. Or, d'une part une analyse purement spatiale est réalisée par le krigeage et d'autre part une analyse temporelle par l'étude des séries temporelles par zone. Etudier conjointement les deux dimensions permet d'avoir une vision de la dynamique spatio-temporelle de la maladie. De la même façon que deux observations proches dans l'espace ont tendance à se ressembler plus que deux observations éloignées, des données proches à la fois dans le temps et dans l'espace sont plus liées (Cressie et Read, 1989). Le krigeage spatio-temporel permet d'estimer les localisations spatio-temporelles inconnues grâce aux mesures situées proches sur le territoire et à la fois grâce à la dimension temporelle.

#### 1.3.4.1. AJUSTEMENT DES VARIOGRAMMES SPATIO-TEMPORELS

En pratique, un variogramme expérimental est calculé par unité de temps dans la période considérée. Un variogramme temporel doit être modélisé en plus du variogramme spatial (Gräler et al., 2016). Dans ce cas, le temps représente une troisième dimension et le variogramme est représenté non plus par une courbe, mais par une surface. L'ajustement est réalisé en minimisant la distance entre le modèle et la surface du variogramme expérimental, par la méthode des moindres carrés pondérés. Le krigeage spatio-temporel est réalisé sur une épidémie complète, permettant la prise en compte de la dynamique spatio-temporelle de la maladie pour cette saison.

#### 1.3.4.2. LA DIMENSION TEMPORELLE COMME VARIABLE AUXILIAIRE

Comme évoqué précédemment, les observations proches dans le temps sont liées. Une autre possibilité de prendre en compte la dimension temporelle est donc de considérer les observations de la variable d'intérêt pour une localisation temporelle proche comme variable auxiliaire (Pebesma et al., 2005). On peut par exemple réaliser le cokrigeage avec comme variable principale les mesures d'incidence de la semaine 201206, et comme variable auxiliaire les mesures d'incidence de la semaine 201205. Ce type de cokrigeage est mis en place de la même façon que le cokrigeage avec les classes médicamenteuses.

### 1.3.5. VALIDATION DES MODELES D'INTERPOLATION SPATIALE

Une des difficultés majeures de ce projet est d'évaluer la qualité du modèle. Le choix des critères de comparaison et de validation des modèles a été une partie importante de mon travail. En effet, la méthode classique de minimisation des erreurs d'estimation n'est pas adaptée à la situation étudiée ici, car on ne dispose pas de données de référence. La « vraie » valeur d'incidence pour un département à une semaine donnée n'est jamais connue, et l'estimation est attribuée arbitrairement au centroïde du département. Il est donc difficile de juger directement de la qualité des estimations issues du krigeage. La vérification du modèle peut se faire grâce à différentes propriétés visuelles des résultats, et aux connaissances concernant le phénomène modélisé. Ces propriétés sont attendues d'un « bon » modèle et doivent montrer que les résultats obtenus par cokrigeage sont mieux que les résultats de krigeage. Des indicateurs quantifiables et des tests de comparaison peuvent aussi être mis en place. Un ensemble d'outils de validation possibles est donc considéré.

Visuellement, la cohérence des cartes issues du krigeage est importante : représentations sans zones de discontinuités, sans changements abrupts, et sans variation à micro-échelle importante. Les séries temporelles permettent de vérifier la cohérence des

prédictions sur l'ensemble de la période d'étude, ainsi que leur adéquation avec l'évolution des cartes. De plus, le bruit d'une série temporelle renseigne sur la précision des estimations de la zone considérée. Le bruit correspond à des fluctuations irrégulières, des variations de faible intensité et de courte durée de nature aléatoire.

Des critères quantifiables peuvent aussi être représentés sur le champ d'étude. La représentation des zones où les écarts sont importants entre les modèles permet de déterminer les localisations où les estimations sont peu robustes. Cette représentation peut être comparée à la représentation de la variance de krigeage. La variance de krigeage est la variance de prédiction minimisée et représente l'incertitude associée à la valeur interpolée pour chaque point de la grille. La variance globale sur le territoire par saison pour chaque modèle, peut mettre en évidence des différences entre épidémie et entre modèles. Des tests simples permettent de détecter des différences significatives entre les résultats issus de différents modèles. Par exemple, la variance globale des différents modèles peut être comparée avec des tests de Student appariés.

Finalement, un modèle est considéré comme raisonnable si les cartes obtenues sont cohérentes, la variance de krigeage est faible et les séries temporelles peu bruitées. Les résultats obtenus avec les modèles de krigeage, de cokrigeage avec l'ensemble des variables des données IMS prises seules, et avec les variables latentes sont donc comparés. Cela permet de choisir la variable auxiliaire qui mène aux meilleurs résultats globalement (Rossiter, 2007).

# 2. RESULTATS

## 2.1. ANALYSE DU LIEN ENTRE LES VARIABLES AUXILIAIRES ET LA VARIABLE PRINCIPALE

### 2.1.1. ANALYSE DE LA CORRELATION TEMPORELLE

La série temporelle de l'incidence nationale (*figure 2*) met en évidence les 5 épidémies de 2011 à 2016. La durée de l'épidémie ainsi que son amplitude varie selon les saisons.

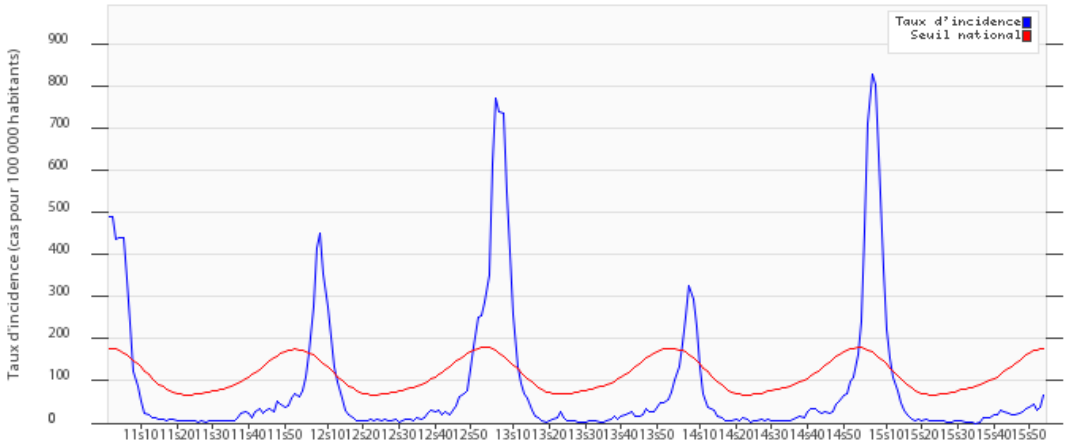



**Figure 2 : Série temporelle de l'incidence nationale calculée par le RS de 201101 à 201601**
La courbe bleue représente le taux d'incidence national, la courbe rouge représente le seuil épidémique. (Source Réseau Sentinelles)

La corrélation temporelle entre l'incidence des SG et les variables médicamenteuses au niveau national est comprise entre 0.4 et 0.85 avec une moyenne de 0.61. Les séries temporelles des classes médicamenteuses tracées en *figure 3* montrent une évolution de leurs délivrances liée à la série de l'incidence des SG. Au niveau départemental, la corrélation temporelle est beaucoup plus élevée pour les VL que pour les classes médicamenteuses.

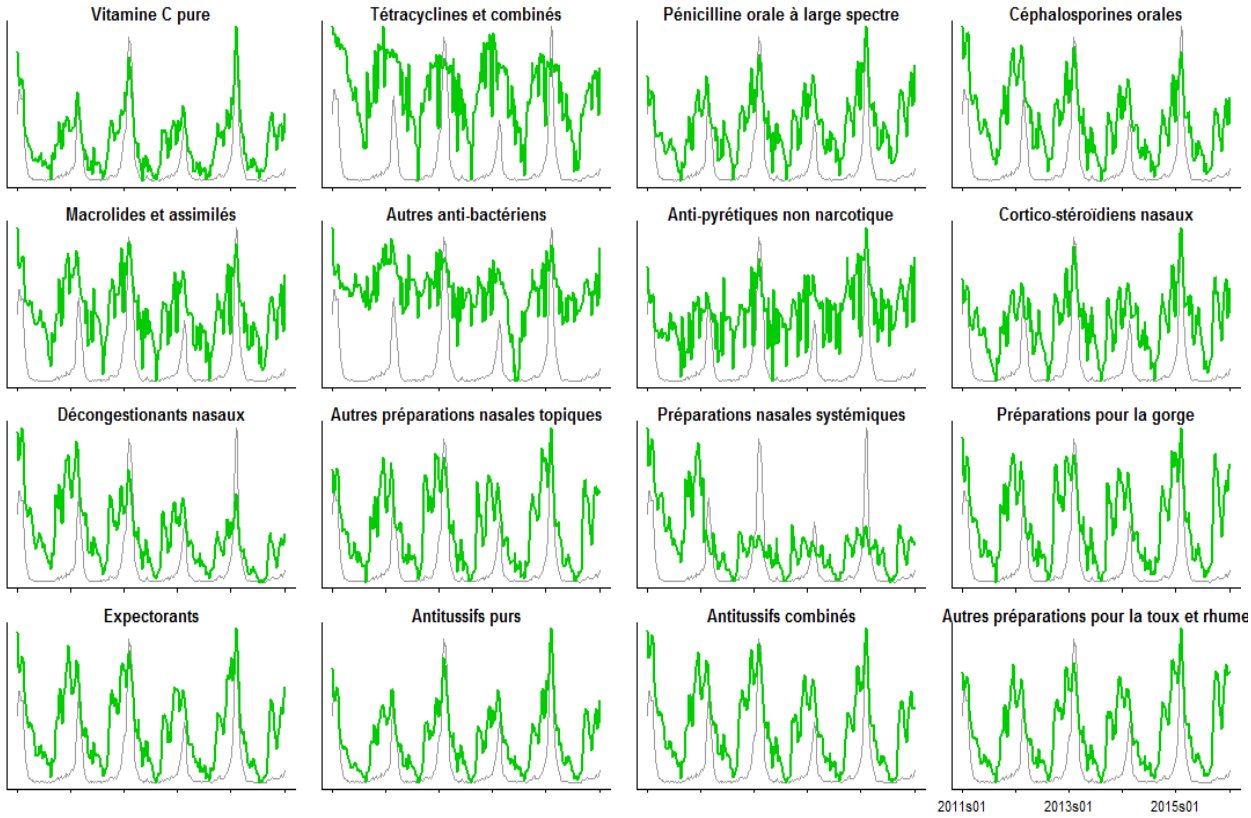



**Figure 3 : Séries temporelles des classes médicamenteuses de 201101 à 201601**
Les courbes vertes représentent les séries de ventes de médicaments, les courbes grises représentent le taux d'incidence national des syndromes grippaux. Les données sont représentées sur des échelles différentes.

### 2.1.2. Analyse de la correlation spatiale

Bien que la corrélation temporelle entre la variable d'incidence et les variables de ventes de médicaments soit élevée, la corrélation spatiale est globalement faible. Selon la classe médicamenteuse, la corrélation spatiale est significative pour 22 semaines en moyenne, au niveau départemental. Cela représente moins de 10% des semaines où la prise en compte d'une variable auxiliaire peut, en théorie (Baillargeon, 2005), apporter une amélioration dans l'estimation des SG. Cette faible corrélation peut être due à des disparités en termes d'utilisation des médicaments entre les différents départements. Pour le vérifier, les distributions d'une même classe de médicaments entre les différentes zones ont été comparées. Les moyennes sont significativement différentes entre les zones (départements ou régions) pour toutes les classes. Une standardisation des données médicamenteuses par zone ne permet pas d'améliorer la corrélation spatiale avec la variable d'incidence.

Dans le cas des variables latentes issues de la PLSR, la corrélation spatiale est améliorée par rapport aux classes seules, et est significative pour plus de 40% des semaines.

## 2.2. Selection des modeles de variogrammes

### 2.2.1. Selection des parametres du variogramme theorique

Les données régionales correspondent à des centroïdes peu nombreux et éloignés sur le territoire et l'ajustement du variogramme est très difficile car les points voisins sont très variables. Dans le cas des données départementales, les points proches sont en général plus liés que les points éloignés, ce qui s'observe par la tendance croissante du variogramme

expérimental (*figure 4*). Nous étudions donc seulement le krigeage des données départementales.

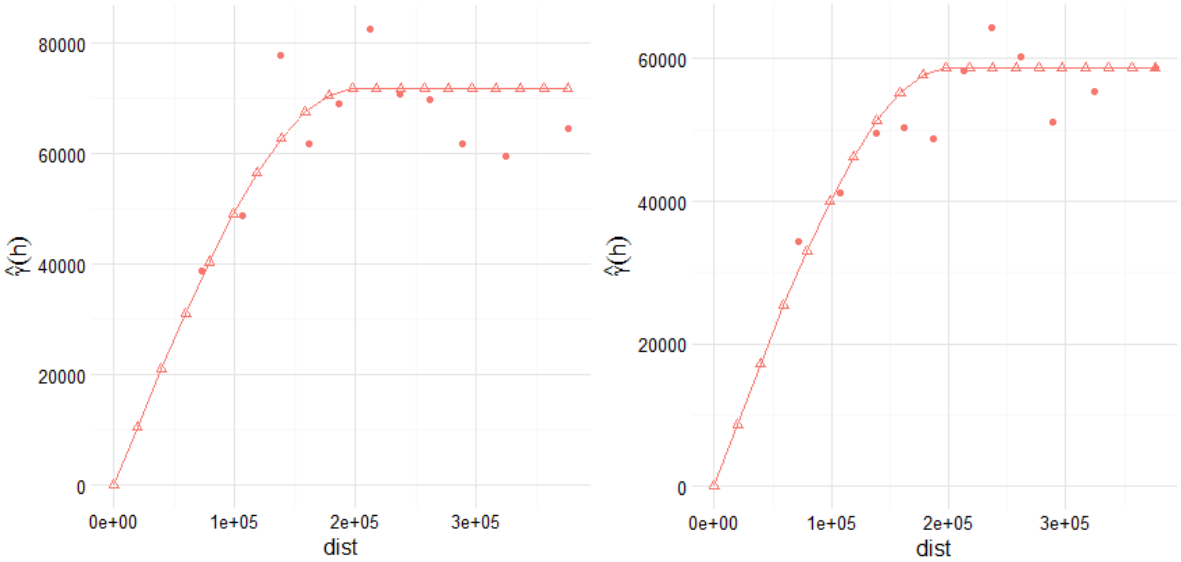



**Figure 4 : Variogrammes théoriques ajustés aux variogrammes expérimentaux de krigeage**
Les points représentent les variogrammes expérimentaux, les courbes les variogrammes théoriques. Variable d'incidence (à gauche) et variable R05C (à droite), pour les données départementales de la semaine 201210.

La sélection automatique des paramètres du variogramme de la variable d'incidence fait ressortir la forme de modèle Stein dans 62% des cas. Cependant, les valeurs de portée associées à ce paramètre sont souvent variables du fait de la non-convergence du modèle pour certaines semaines. Les semaines sans problème de convergences ont globalement une valeur de portée située entre 90 et 300 km. On choisit donc de conserver la valeur de portée de 200 km pour initialiser l'ajustement des variogrammes par la suite. En fixant ce paramètre, la sélection automatique est effectuée de nouveau et le modèle sphérique entraîne les erreurs les plus faibles dans plus de 77% des cas (199 semaines sur 258). La variance de krigeage est diminuée avec un modèle sphérique et une portée de 200 km, par rapport aux autres modèles. Le fait d'utiliser des paramètres différents pour chaque semaine ne permet pas d'amélioration des estimations. Il est donc préférable d'utiliser toujours les mêmes paramètres, qui sont adaptés au processus spatial étudié. Pour les semaines où il n'y a pas de convergence lors de l'ajustement automatique du modèle, ces paramètres par défaut permettent le krigeage, même si le variogramme théorique s'ajuste mal au variogramme expérimental.

L'étude de la variabilité des estimations d'incidences par sous-échantillonnage montre que la variabilité micro-échelle n'est pas nulle (Annexe III). Cependant, les tentatives d'ajustement du variogramme avec un paramètre d'effet pépite différent de la valeur nulle entraînent des problèmes de convergences ou des aberrations dans les modèles, comme des surestimations très importantes des variances, notamment en cokrigeage. On travaille donc avec un effet pépite fixé à 0.

Le premier intervalle de calcul du variogramme permet d'inclure 80% des distances entre un centroïde et son plus proche voisin, et est fixé à 75 km. Des intervalles de 25km sont ensuite utilisés. Ce choix est arbitraire, mais permet la prise en compte d'un nombre suffisant de points pour chaque intervalle de calcul (de 40 à 225 selon les semaines et selon les intervalles, les distances les plus élevées incluant un plus grand nombre de points dans le calcul), de limiter les aberrations lorsque trop peu de points sont pris en compte, et d'avoir des points de variogrammes représentatifs des points du territoire.

Le fait de prendre peu de voisins mène à des estimations à variance plus élevée que le krigeage local à 9 voisins. Aucune analyse n'a pu montrer que le fait d'augmenter le nombre de voisins lors du krigeage local améliore significativement les résultats, alors que les temps de calculs augmentent fortement. Le krigeage local avec 9 voisins paraît donc adapté à cette étude.

Le choix de certains paramètres est finalement arbitraire car aucune analyse n'a pu montrer que certaines valeurs sont plus adaptées. La connaissance du processus et des données, et l'optimisation des temps de calculs permettent de choisir ces critères. Un exemple de variogrammes théoriques de cokrigeage est illustré en *figure 5*.

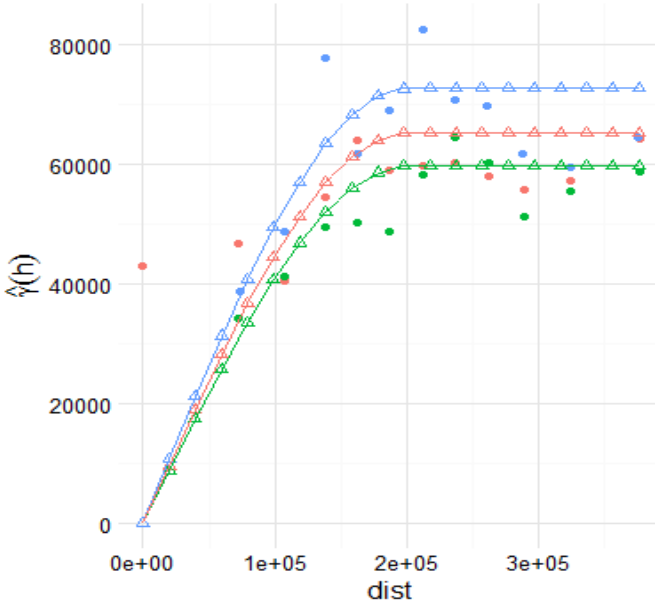



**Figure 5 : Variogrammes théoriques ajustés aux variogrammes expérimentaux (points) de cokrigeage**
Pour la semaine 201210. Le variogramme de l'incidence est représenté en bleu, le variogramme de la classe médicamenteuse R05C est représenté en vert. La courbe rouge représente le variogramme croisé.

### 2.2.2. SELECTION DES VARIABLES AUXILIAIRES SELON LES CRITERES D'INCLUSION

Les paramètres de forme de modèle et de portée sélectionnés pour le variogramme d'incidence sont visuellement adaptés aux variogrammes des variables de médicaments (*figure 4*).

Globalement la corrélation spatiale entre l'incidence et les données IMS est faible, et souvent ne justifie pas une utilisation du cokrigeage avec ces variables, a priori. Il est difficile d'affirmer qu'une des classes est plus adaptée qu'une autre à notre modèle, et de présélectionner les classes pour mener l'ensemble des analyses. La variable latente obtenue par régression PLS est plus corrélée spatialement et temporellement à l'incidence et peut entraîner des meilleurs résultats de cokrigeage.

Toutes les variables auxiliaires disponibles sont intégrées dans le krigeage une à une et l'ensemble des modèles de cokrigeage à une variable possibles est donc étudié. Le choix se fait donc a posteriori, en cherchant à sélectionner la variable auxiliaire qui permet d'aboutir aux meilleurs résultats en termes d'optimisation de l'ensemble de nos critères d'intérêt.

## 2.3. RESULTATS DU KRIGEAGE ET COKRIGEAGE

### 2.3.1. CARTES DE COKRIGEAGE

Les cartes obtenues par cokrigeage des données départementales sont globalement cohérentes car il n'y a pas de zone de discontinuités, de changements micro-échelle et leur évolution au cours du temps est cohérente. Ces cartes sont très semblables aux cartes réalisées avec les incidences seules (*figure 6*). Il n'y a pas de grosses différences entre les cartes obtenues avec les différentes variables auxiliaires. Nous ne pouvons donc pas faire de choix sur le critère visuel des résultats, mais cela indique que le cokrigeage fournit des résultats cohérents et exploitables. Représenter les points non mesurés souligne la présence de différences au niveau de zones sans mesures.

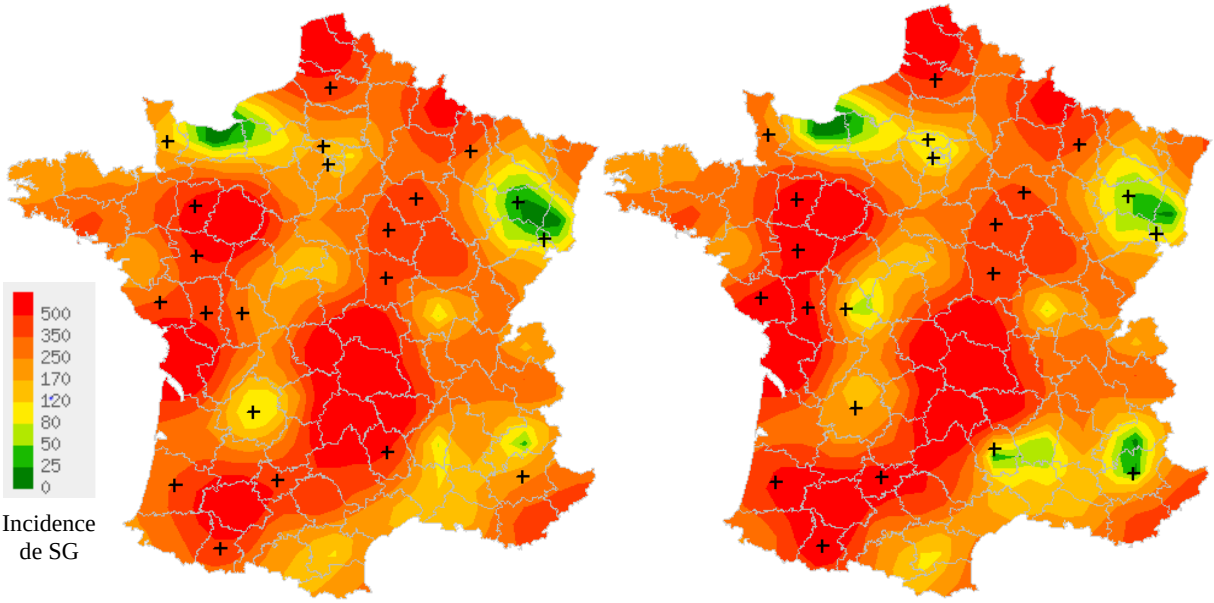



**Figure 6 : Cartes de krigeage et de cokrigeage avec la variable latente**
Les données utilisées sont les données départementales de la semaine 201210. La carte de krigeage est représentée à gauche, celle de cokrigeage est représentée à droite. Les points noirs correspondent aux centroïdes à données manquantes pour cette semaine.

### 2.3.2. EVALUATION DES DIFFERENTS CRITERES DE VALIDATION

La représentation des différences moyennes entre cokrigeage et krigeage (*figure 7*) pour chaque saison de 2011 à 2016 montre que la différence entre krigeage et cokrigeage est négligeable pour la majorité des zones. Cependant, pour les zones pauvres en mesures, dont l'estimation par krigeage est peu précise, la différence est importante systématiquement. Nous avons choisi de nous intéresser à ces zones-là. Il est raisonnable d'attendre que les estimations d'un « bon » modèle entraînent aux zones ciblées des propriétés retrouvées aux zones bien estimées en krigeage. Ces zones correspondent à des départements toujours mesurés, et ont des séries temporelles peu bruitées et une faible variance de krigeage. Les zones de différences sont moins nombreuses en 2015 qu'en 2011, ce qui s'explique par une meilleure couverture des médecins déclarants, et une diminution du nombre de départements non mesurés. Les différences entre cokrigeage et krigeage sont plus importante avec les variables auxiliaires de classes médicamenteuses qu'avec les variables latentes.

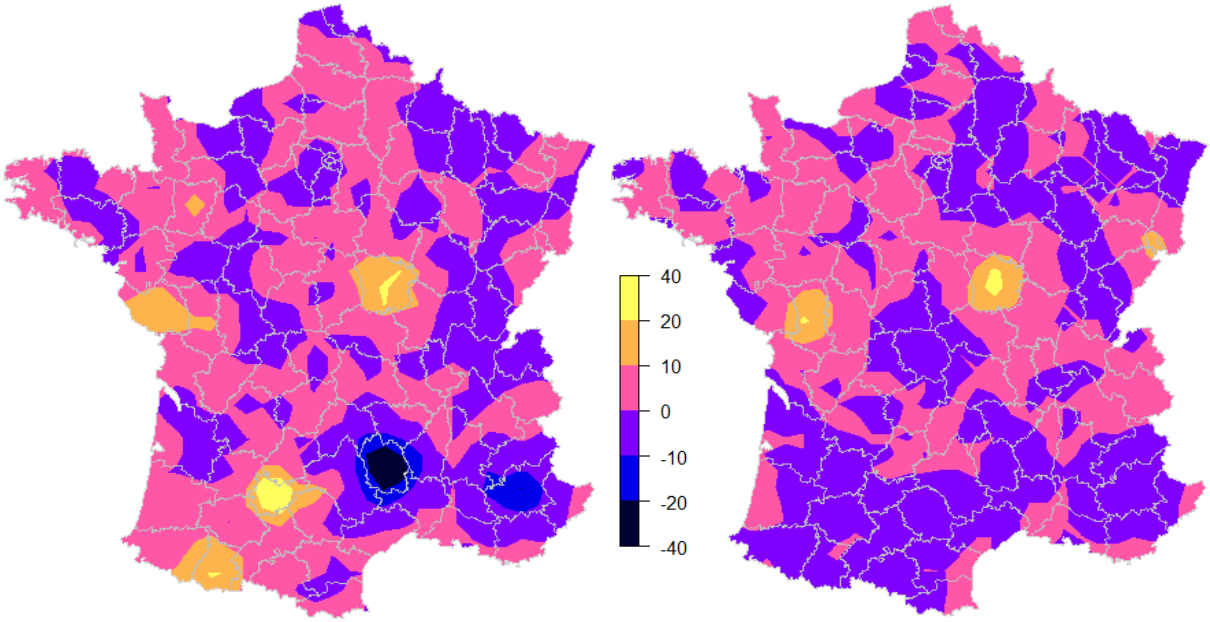



**Figure 7 : Cartes des différences moyennes entre les estimations de SG obtenues par cokrigeage avec variable latente et les estimations obtenues par krigeage pour l'année 2011 (à gauche) et 2015 (à droite)**

Le même constat est observé avec la représentation de la variance de krigeage (*figure 8).* Globalement, la variance est plus faible autour des centroïdes de départements qui correspondent à des points de mesure. Dans le cas du krigeage la variance est plus élevée dans les zones où il y a beaucoup de valeurs manquantes. Ces centroïdes à fortes variances en krigeage ne sont pas retrouvés dans le cas du cokrigeage, où la variance est du même ordre que la variance aux autres zones de mesures. Le cokrigeage permet donc de fournir des prédictions plus précises pour les zones non mesurées, même lorsque la variable auxiliaire est peu corrélée à l'incidence. Les zones pour lesquelles la différence entre krigeage et cokrigeage diminue avec le temps correspondent à des zones pour lesquelles la variance de prédiction diminue. Ce sont des zones pour lesquelles le nombre de MS déclarant augmente.

Pour ces zones à fortes différences, les séries temporelles des prédictions par krigeage sur l'incidence sont beaucoup plus bruitées pendant les épidémies que dans le cas du cokrigeage, où les séries sont plus lisses. Si au cours du temps, pour une zone donnée, la différence krigeage-cokrigeage diminue, la série temporelle de krigeage est aussi de moins en moins bruitée.

A posteriori, des variables auxiliaires permettent des résultats optimaux. La variance est significativement plus faible dans le cas du cokrigeage avec les classes J01X9 (antibactériens), R01B (préparations nasales) et R05F (préparations contre la toux et le rhume) seules. La variance est diminuée dans les zones creuses mais aussi sur l'ensemble du territoire. La classe permettant les variances les plus faibles est R05F. Avec les modèles incluant les variables latentes obtenues par PLS, les variances sont significativement plus faibles que pour les modèles avec les classes médicamenteuses. La variable latente calculée sur plusieurs semaines permet des estimations légèrement meilleures que la variable latente calculée pour chaque semaine. La variable latente incluant les mesures d'incidence de la semaine précédente entraîne des résultats de cokrigeage très similaires à ceux de krigeage en termes de variance et de valeurs estimées. Pour cette variable auxiliaire, le cokrigeage n'apporte pas d'amélioration significative des résultats.

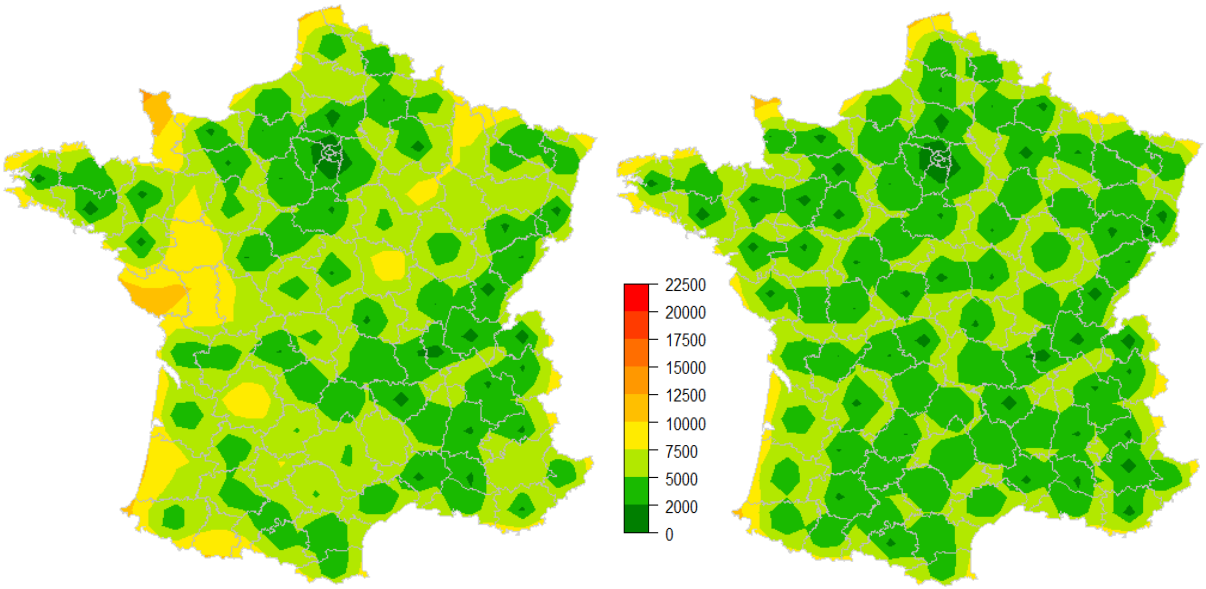



**Figure 8 : Carte de variance de krigeage moyenne (à gauche) et carte de variance moyenne de cokrigeage avec variable latente (à droite) pour l'année 2012**

Au cours du temps, le nombre de médecins Sentinelles déclarants augmente et est mieux réparti, il y a moins de départements non mesurés (Annexe IV). Le nombre de MS déclarant est le plus faible en 2013, année où les estimations sont les moins précises tous modèles confondus. A partir de 2015, l'échantillonnage des médecins est suffisamment bon pour que l'estimation par krigeage soit plus précise que les années précédentes. Pour cette année-là, il n'y a pas de différence de variance significative entre le krigeage et le cokrigeage avec la VL. La variance de krigeage est significativement plus faible que la variance de cokrigeage des classes J01X9, R05F et R01B. Le cokrigeage corrige donc certaines zones pauvres en mesures ; mais dans le cas où les données sont bien réparties sur le territoire, l'amélioration des estimations n'est pas significative.

## 2.4. ETUDE DE LA DIMENSION TEMPORELLE

### 2.4.1. CARTES DE KRIGEAGE SPATIO-TEMPOREL

Le krigeage spatio-temporel est très difficile à mettre en place car le nombre de paramètres à déterminer est élevé. L'ajustement se fait principalement visuellement et les tentatives d'automatisation ne permettent pas la convergence du modèle, avec des temps de calculs très longs. Ces problèmes sont très courants en krigeage spatio-temporel, ce qui en fait une méthode peu utilisée (Wikle, 2015). De plus, il n'y a aucune possibilité de gestion des données manquantes qui sont remplacées par 0 par nécessité de travailler sur un maillage spatio-temporel complet. Une étude ponctuelle de l'épidémie de 2012 a été réalisée. Les variogrammes spatio-temporels expérimental et théorique de cette période sont représentés en *figure 9*.

Le krigeage spatio-temporel permet de bien visualiser l'évolution temporelle de l'épidémie et la migration des SG. Il n'y a pas de grosse différence par rapport aux cartes de krigeage estimées pour chaque semaine. Les cartes obtenues sont globalement cohérentes et il n'y a pas de changement anormal entre deux semaines consécutives (*figure 10*).

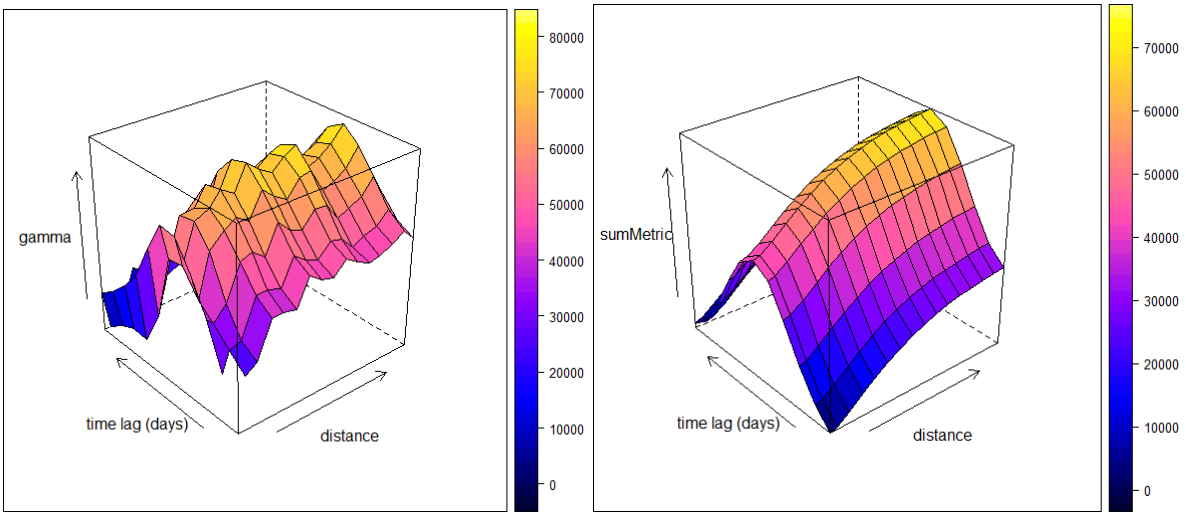



**Figure 9 : Variogrammes spatio-temporel calculés sur la période 201202 - 201212**
La carte de gauche représente le variogramme expérimental, la carte de droite le variogramme théorique

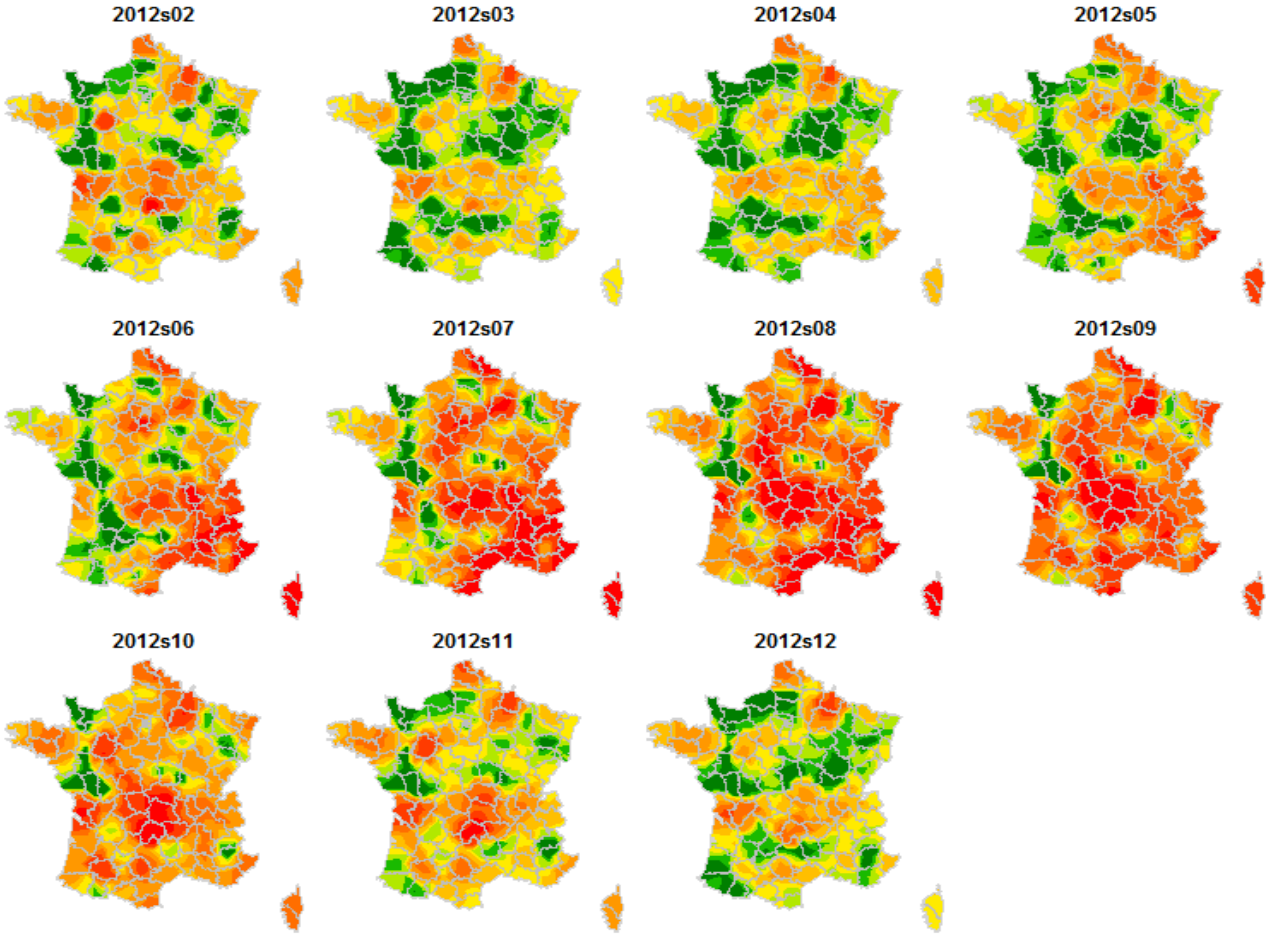



**Figure 10 : Cartes obtenues par krigeage spatio-temporel pour l'épidémie 2012**

L'autocorrélation de la variable d'incidence entre les valeurs pour une semaine donnée et les valeurs de la semaine précédente est généralement élevée, et significative pour 96% des semaines. Pour le cokrigeage avec l'incidence de la semaine précédente comme variable auxiliaire, les cartes obtenues sont cohérentes et il y a des modifications légères par rapport au krigeage. Cependant, la variance de krigeage est plus élevée, d'autant plus que les zones pauvres en mesures sont souvent retrouvées d'une semaine à l'autre. Globalement, cette méthode n'améliore pas les estimations de SG. Cela pourrait être utile pour améliorer le

krigeage dans le cas où très peu de données seraient disponibles pour une semaine particulière.

### 2.4.2. SERIES TEMPORELLES

Pour les zones complètes en mesure, la différence entre les séries temporelles de krigeage et de cokrigeage est moindre (*figure 11*). Le cokrigeage avec variable latente entraîne des séries légèrement moins bruitées que celles issues du krigeage et du cokrigeage avec la classe R05F. Globalement, pour ces zones, l'apport du cokrigeage ne semble pas important.

Pour les zones comportant de nombreuses valeurs manquantes, la série est complétée (*figure 12*) grâce aux mesures disponibles pour les zones voisines. Comme évoqué précédemment, les différences entre modèles sont plus importantes. Pour ces zones, le bruit de la série de krigeage est plus important que pour le cokrigeage et diminue en même temps que la différence entre cokrigeage et krigeage se rapproche de 0. Cela est visible pour les zones qui commencent à être renseignées au cours de la période d'étude (*figure 12*). La série du cokrigeage avec la VL construite par semaine est plus proche de celle de krigeage que la série du cokrigeage de la classe R05F. La série de cokrigeage avec la VL construite sur 16 semaines est plus bruitée et plus de différences avec le krigeage sont visibles par rapport aux deux autres séries de cokrigeage.

Lorsque l'on recalcule les incidences nationales à partir des estimations d'incidences départementales par cokrigeage, on obtient une série très proche de la série d'incidence nationale publiée. De plus, les résultats des séries temporelles confirment les observations précédentes, et montrent que la variable latente calculée pour chaque semaine est plus adaptée que la variable latente calculée sur plusieurs semaines. En effet, cette variable auxiliaire permet des estimations plus précises spatialement avec une variance de krigeage diminuée, et temporellement avec une série moins bruitée et plus proche de celle de krigeage que les séries des autres modèles de cokrigeage.

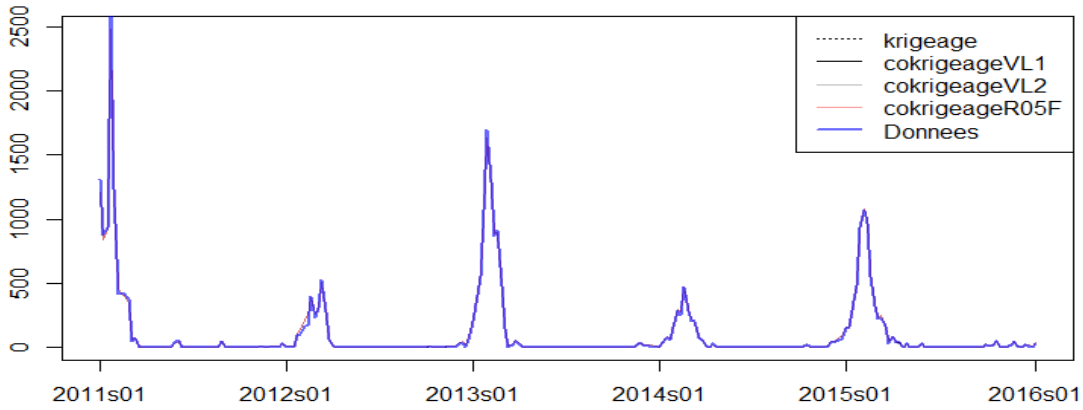



**Figure 11 : Incidences estimées de 201101 à 201601 pour le département de l'Ain**
par krigeage (pointillé noir), par cokrigeage avec la variable latente construite par semaine (noir), par cokrigeage avec la variable latente construite sur des périodes de 16 semaines (gris), par cokrigeage de la classe R05F (rouge). Les données de départ sont représentées par la courbe bleue.

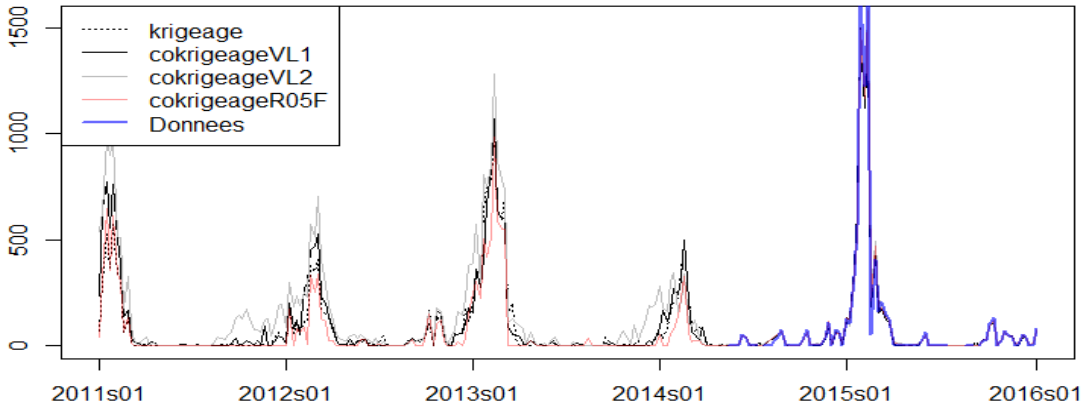



**Figure 12 : Incidences estimées de 201101 à 201601 pour le département du Tarn-et-Garonne** par krigeage (pointillé noir), par cokrigeage avec la variable latente construite par semaine (noir), par cokrigeage avec la variable latente construite sur des périodes de 16 semaines (gris), par cokrigeage avec la classe R05F(rouge). Les données de départ sont représentées par la courbe bleue.

# 3. DISCUSSION

## 3.1. LES CHOIX METHODOLOGIQUES

Pour cette étude, le choix a été fait d'utiliser les données auxiliaires de volumes de délivrance médicamenteuse ayant déjà été étudiées par le réseau Sentinelles. D'autres données auxiliaires sont disponibles pour prédire la grippe, comme par exemple les données de requête sur les moteurs de recherches. Ces données ont d'ailleurs été étudiées au RS par Pelat et al. (2009). Cependant, plusieurs études montrent des problèmes de surestimation liés notamment à un effet médiatique (Lazer et al., 2014).

Le projet initial était d'utiliser les données d'incidence régionales plutôt que départementales car elles sont plus robustes, notamment grâce à une meilleure couverture au niveau régional que départemental. Les données de médicaments auxiliaires devaient être utilisées au niveau départemental pour avoir un maillage plus fin d'observations sur le territoire. Cette façon de procéder au cokrigeage est conseillée par de nombreux auteurs (Bivand et al. (2008), Rossiter (2007)). Cependant, les cartes obtenues à l'issue du krigeage des incidences régionales sont moins précises et la variance d'estimation est importante dans la majorité des zones du territoire. Les cartes montrent des changements abrupts et une variabilité micro échelle importante. Il n'est donc pour l'instant pas envisageable de réaliser des cartes précises avec les données d'incidences régionales. Par ailleurs, les médecins déclarants sont de mieux en mieux répartis sur le territoire et le nombre de départements non mesurés diminue de 2011 à 2015, ce qui permet d'augmenter la qualité du krigeage des données départementales.

Toutes les analyses réalisées n'ont pas été détaillées dans ce rapport, notamment le krigeage « block » réalisé sur des données surfaciques, où chaque mesure correspond à la surface départementale, et non plus un point. D'autres méthodes peuvent être développées sur les données spatiales dont nous disposons, comme les méthodes bayésiennes par exemple, et l'étude de données surfaciques. Les choix d'utiliser les méthodes détaillées dans ce mémoire ont été guidés notamment par le fonctionnement du réseau Sentinelles. Un des principaux obstacles au choix de certaines méthodes et de certains paramètres a été les temps de calculs, qui augmentent rapidement lorsque l'on travaille sur une grille de plus de 1600 cellules, pour 262 semaines avec un nombre de modèles à comparer important. Les critères de validation utilisés sont souvent visuels ce qui est une limite importante, ils ont été choisis parmi un

ensemble de critères possibles attendus d'un modèle fournissant des estimations robustes. D'autres méthodes de validation auraient pu être utilisées comme par exemple le bootstrap paramétrique par exemple (Iranpanah et al., 2011) qui permet la validation du modèle de variogramme.

Une distribution normale de la variable principale n'est pas nécessaire en krigeage mais conseillée. De nombreuses contraintes ont poussé à ne pas travailler avec la transformation logarithmique de la variable d'incidence. En effet, cela entraîne des aberrations dans les estimations de krigeage au niveau local (Roth, 1998). De plus, l'incidence est souvent nulle, et la transformation logarithmique n'est pas adaptée. Enfin, cela nécessite de transformer toutes les variables auxiliaires.

Beaucoup de paramètres d'ajustement du variogramme fixés par le RS n'ont pas été modifiés. Cette étude a montré que la forme de modèle sphérique et la portée sont des paramètres adaptés aux données du réseau. Cependant, pour l'effet pépite, la valeur par défaut n'est pas adaptée aux données. Le ré-échantillonnage montre que la variabilité à micro-échelle n'est pas nulle. Malgré la volonté de modifier ce paramètre et de l'adapter aux données, l'ajustement est très difficile et les variogrammes sont souvent surestimés. Une étude approfondie de l'ajustement de ce paramètre pourrait être menée.

### 3.2. RETOUR SUR LES RESULTATS

Au départ, nous souhaitions faire de la sélection de variables a priori pour savoir quelles variables auxiliaires utiliser pour le cokrigeage, et ne pas tester tous les modèles pour diminuer les temps de calculs. Cependant, les critères d'inclusion des variables auxiliaires ne permettent pas de faire ressortir une structure particulière parmi l'ensemble des variables et de leurs combinaisons disponibles. On ne peut donc pas présélectionner les classes médicamenteuses les plus liées à l'incidence et tous les modèles sont évalués. Cette méthode est aussi mise en place par Rossiter (2007), dont les modèles se distinguent après analyses des résultats. Les résultats nous permettent bien de faire ressortir les modèles permettant les meilleures estimations, même si une amélioration de l'échantillonnage des médecins déclarants fait diminuer la différence entre le krigeage et le cokrigeage. La plupart des auteurs recommandent de se baser principalement sur la corrélation spatiale entre variables lors de la construction du modèle. Cependant, dans notre cas, des modèles de cokrigeage incluant certaines variables peu corrélées à la variable principale ont mené à des résultats meilleurs que ceux obtenus par krigeage. Il ne semble donc pas raisonnable d'utiliser la corrélation spatiale comme un critère unique pour sélectionner les variables auxiliaires, mais plutôt comme un guide lorsque de nombreuses variables sont disponibles.

De plus, dans notre cas, il apparaît logique que le modèle incluant la variable latente comme variable auxiliaire entraîne de meilleurs résultats. En effet, au-delà de la meilleure corrélation entre cette variable et l'incidence, en général une combinaison de médicaments issus de différentes classes médicamenteuses est prescrite en cas de SG, et non un seul médicament. Les poids attribués à chaque classe pour l'ensemble des PLSR réalisées sur chaque semaine de la période d'étude ont des variances différentes. Les classes dont les poids sont les plus stables sont J01X9, R01B et R05F, qui sont aussi les classes qui entraînent les meilleurs résultats de cokrigeage après les VL.

La représentativité des données du RS a été étudiée par Souty et al. (2014), qui a montré que les données sont représentatives pour un ensemble de critères mais pas tous. Des

données corrigées ont donc été calculées. Il pourrait être envisagé d'utiliser ces données pour le krigeage.

Les cartes obtenues par cokrigeage sont comparables aux cartes publiées obtenues par krigeage. Les résultats sont cohérents et les données auxiliaires ne modifient pas complètement les données d'incidence, mais entraînent des « corrections » pour les zones mal estimées dans le cas du krigeage. Ces corrections ponctuelles sont évoquées par Chiles et Delfiner (2009). C'est plutôt bon signe : les données du réseau Sentinelles sont cohérentes à la base, le cokrigeage permet juste d'améliorer les estimations pour les zones creuses.

La qualité des modèles n'est pas évaluable avec les méthodes classiques de différence entre observations et estimations. On n'évalue pas non plus la qualité d'ajustement du variogramme seul, mais le résultat final de krigeage. Différents critères sont optimisés ici, visuels et quantitatifs. Les résultats pour les différents critères se recoupent, ce qui renforce les conclusions et confirme que ces propriétés permettent d'évaluer les résultats. On voit bien ici la complémentarité de ces différentes représentations des résultats, qui permettent de choisir la variable auxiliaire.

L'avantage de la méthode de krigeage spatio-temporel est que chaque semaine est prédite en fonction des autres. Cette méthode étant difficile à mettre en place, son utilisation régulière n'est pas envisagée pour l'instant. Cependant, l'étude ponctuelle de certaines épidémies par krigeage spatio-temporel permet d'avoir une vision d'ensemble de l'épidémie.

## CONCLUSION

L'objectif de cette étude était d'améliorer la robustesse des estimations spatiales de syndromes grippaux. L'intégration des données de ventes de médicaments dans les modèles d'interpolation spatiale de l'incidence des syndromes grippaux a été possible grâce à la méthode de cokrigeage. Différents critères visuels et quantifiables et la représentation spatiale et temporelle des résultats mettent en évidence une amélioration significative des estimations grâce à plusieurs classes médicamenteuses. Les meilleurs résultats sont obtenus avec un modèle incluant l'ensemble des classes grâce à une variable latente construite par régression PLS. Les estimations sont améliorées au niveau de zones ciblées pauvres en mesures. Les données auxiliaires permettent donc de compléter les estimations réalisées par krigeage, sans changer la simplicité de mise en place ni les temps de calculs. L'apport des données auxiliaires est significatif lorsque la couverture spatiale des médecins déclarant n'est pas optimale.

De plus, l'utilisation de paramètres variographiques adaptés permet une amélioration des résultats de krigeage. Enfin, la prise en compte des dimensions spatiales et temporelles simultanément est possible en krigeage spatio-temporel et a permis l'étude ponctuelle d'épidémie.

## BIBLIOGRAPHIE

Baillargeon, S. (2005). Le krigeage: revue de la théorie et application. Mémoire, Université de Laval, Québec.

Bivand, R., Pebesma, E.,Gomez-Rubio, V. (2008). Applied spatial data analysis with R. éd.: Springer.

Carrat, F.,Valleron, A. J. (1992). Epidemiologic mapping using the "kriging" method: application to an influenza-like illness epidemic in France. *Am J Epidemiol*, 135(11), 1293-1300.

Chiles, J.-P.,Delfiner, P. (2009). Geostatistics: modeling spatial uncertainty. éd.: John Wiley & Sons. ISBN 0470317833.

Costagliola, D., Flahault, A., Galinec, D., Garnerin, P., et al. (1991). A routine tool for detection and assessment of epidemics of influenza-like syndromes in France. *American Journal of Public Health*, 81(1), 97-99.

Cressie, N.,Read, T. R. C. (1989). Spatial Data Analysis of Regional Counts. *Biometrical Journal*, 31(6), 699-719.

Deckers, J. G., Paget, W. J., Schellevis, F. G.,Fleming, D. M. (2006). European primary care surveillance networks: their structure and operation. *Fam Pract*, 23(2), 151-158.

Emery, X. (2012). Cokriging random fields with means related by known linear combinations. *Computers & Geosciences*, 38(1), 136-144.

Goulard, M.,Voltz, M. (1992). Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, 24(3), 269-286.

Gräler, B., Pebesma, E.,Heuvelink, G. (2016). Spatio-temporal interpolation using gstat.

Iranpanah, N., Mohammadzadeh, M.,Taylor, C. C. (2011). A comparison of block and semi-parametric bootstrap methods for variance estimation in spatial statistics. *Computational Statistics & Data Analysis*, 55(1), 578-587.

Lazer, D., Kennedy, R., King, G.,Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. *Science*, 343(6176), 1203-1205.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, 58(8), 1246-1266.

Pebesma, E. J., Duin, R. N. M.,Burrough, P. A. (2005). Mapping sea bird densities over the North Sea: spatially aggregated estimates and temporal changes. *Environmetrics*, 16(6), 573-587.

Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A., et al. (2009). More diseases tracked by using Google Trends. *Emerg Infect Dis*, 15(8), 1327-1328.

Ribeiro Jr, P. J.,Diggle, P. J. (2001). geoR: a package for geostatistical analysis. *R news*, 1(2), 14-18.

Rossiter, D. G. (2007). {Technical Note: Co-kriging with the gstat package of the R environment for statistical computing}. éd. Enschede, Netherlands: International Institute for Geo-information Science & Earth Observation (ITC), 81 p.

Roth, C. (1998). Is Lognormal Kriging Suitable for Local Estimation? *Mathematical Geology*, 30(8), 999-1009.

Sampson, P. D., Richards, M., Szpiro, A. A., Bergen, S., et al. (2013). A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM(2.5)

concentrations in epidemiology. *Atmospheric environment (Oxford, England : 1994)*, 75, 383-392.

Schlaud, M. (1999). Comparison and harmonisation of Denominator data for primary health care research in countries of the European Community: the European Denominator project. éd.: IOS Press. ISBN 0967335523.

Souty, C., Turbelin, C., Blanchon, T., Hanslik, T., et al. (2014). Improving disease incidence estimates in primary care surveillance systems. *Popul Health Metr*, 12, 19.

Turbelin, C.,Boëlle, P.-Y. (2010). Improving general practice based epidemiologic surveillance using desktop clients: the French Sentinel Network experience. *Studies in Health Technology and Informatics*, 160(Pt 1), 442-446.

Turbelin, C., Souty, C., Pelat, C., Hanslik, T., et al. (2013). Age Distribution of Influenza Like Illness Cases during Post-Pandemic A(H3N2): Comparison with the Twelve Previous Seasons, in France. *PLoS One*, 8(6), e65919.

Vergu, E., Grais, R. F., Sarter, H., Fagot, J. P., et al. (2006). Medication sales and syndromic surveillance, France. *Emerg Infect Dis*, 12(3), 416-421.

Wikle, C. K. (2015). Modern perspectives on statistics for spatio-temporal data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1), 86-98.

Sitographie

IMS Health Pharmastat, Pharmastat, https://www.ims-pharmastat.fr/pharmastat (consulté le 10/08/2016)

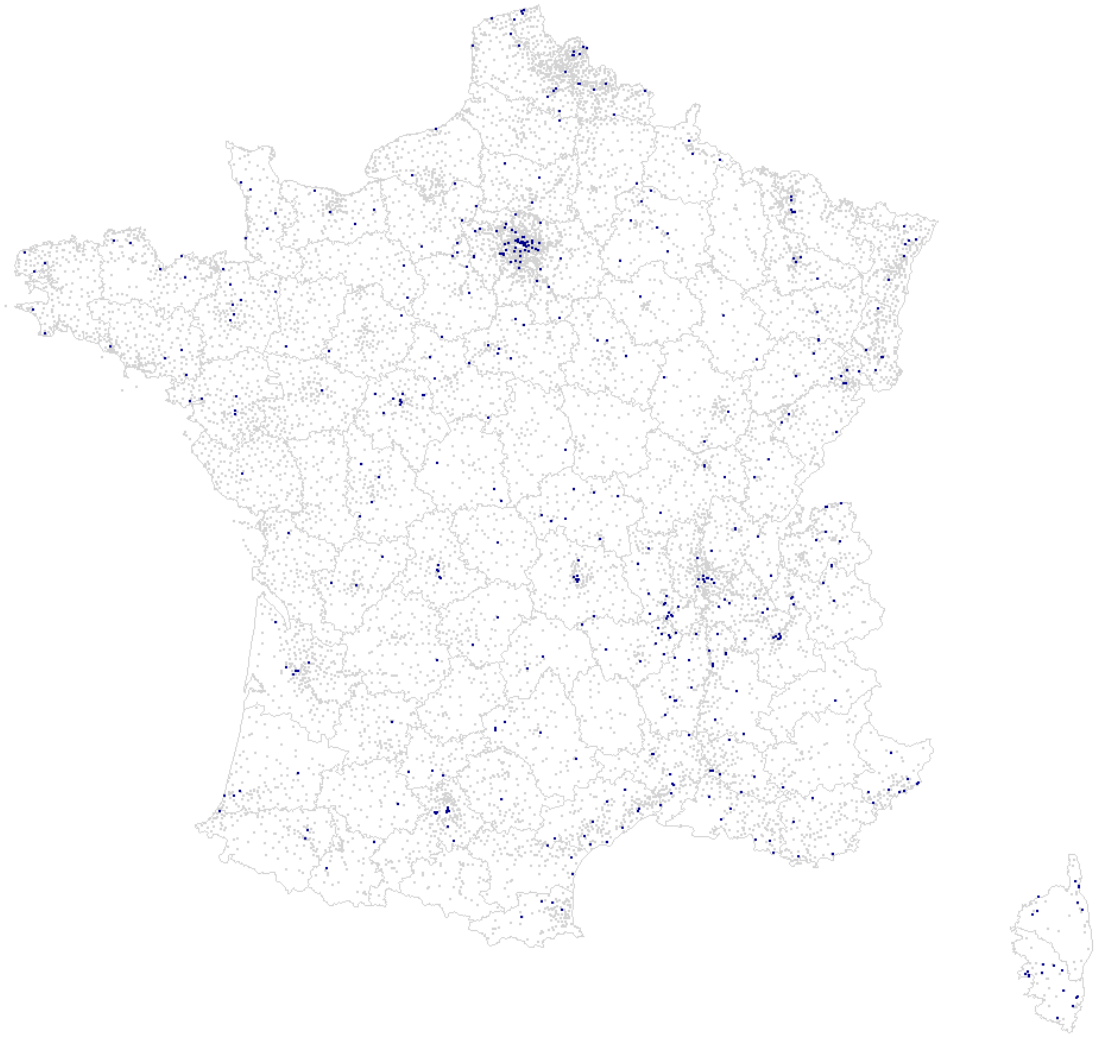
Réseau Sentinelles, INSERM, UPMC Bases de données, https://websenti.u707.jussieu.fr /sentiweb/?page =database (consulté le 10/08/2016)

Réseau Sentinelles, INSERM, UPMC, Bilans annuels, https://websenti.u707.jussieu.fr /sentiweb/?page=bilan (consulté le 10/08/2016)

World Health Organization, Public health surveillance, http://www.who.int/topics/public_ health_surveillance/en/ (consulté le 10/08/2016)

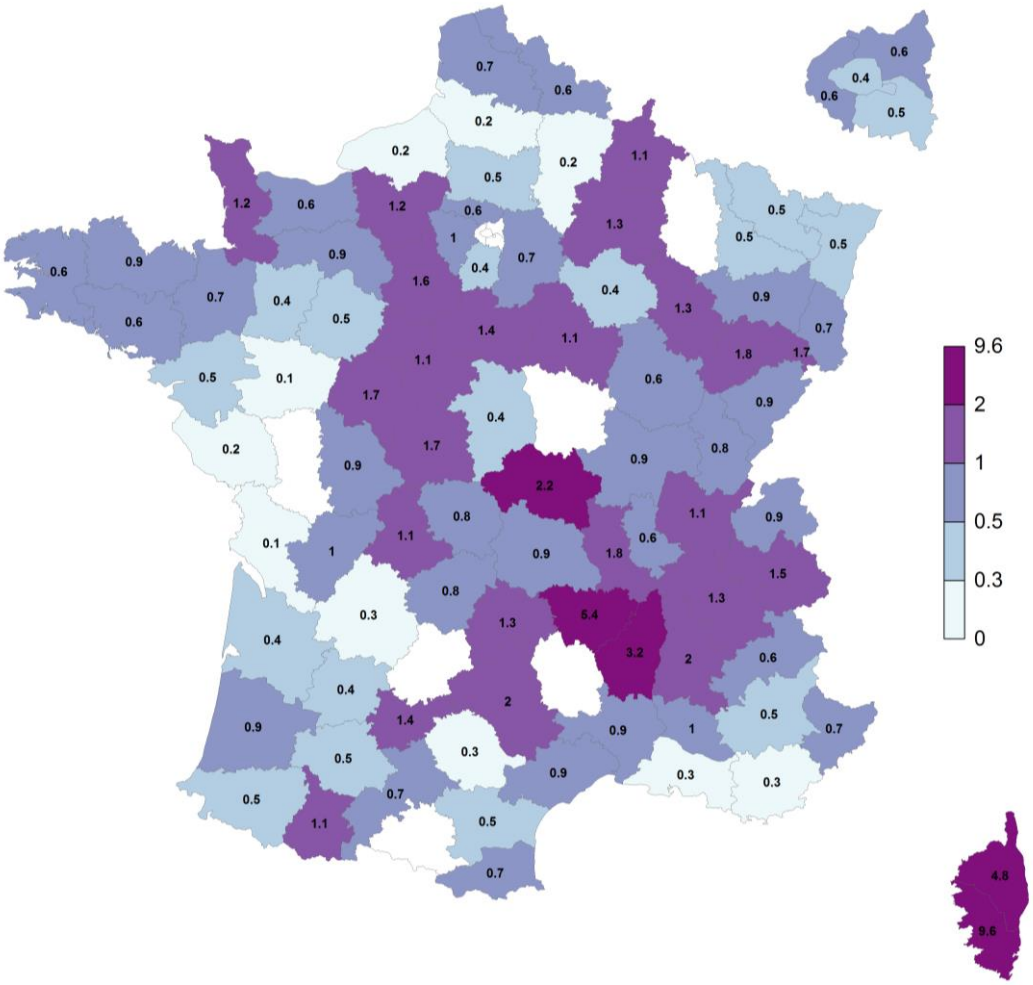
WHO Collaborating Centre for Drug Statistics Methodology, ATC – Structure and principles (2011) http://www.whocc.no/atc/structure_and_principles/ (consulté le 10/08/2016)

## **Annexe I :** Les médecins généralistes Sentinelles déclarants en 2015




Localisation des MG Sentinelles (points bleu) ayant participé en 2015 à la surveillance continue en regard de l'ensemble des médecins généralistes libéraux (points gris) en France métropolitaine au 1er janvier 2016 (Source : Bilan annuel 2015 du Réseau Sentinelles).

**Annexe I (Suite):**




Proportions (en %) des MG Sentinelles ayant participé à la surveillance continue en 2015 par rapport à l'ensemble des MG libéraux en exercice dans le département concerné en France métropolitaine (les départements en blanc correspondent aux départements où aucun MG Sentinelles n'a participé) (Source : Bilan annuel 2015 du Réseau Sentinelles)

**<u>Annexe II :</u> Description des variables auxiliaires médicamenteuses**

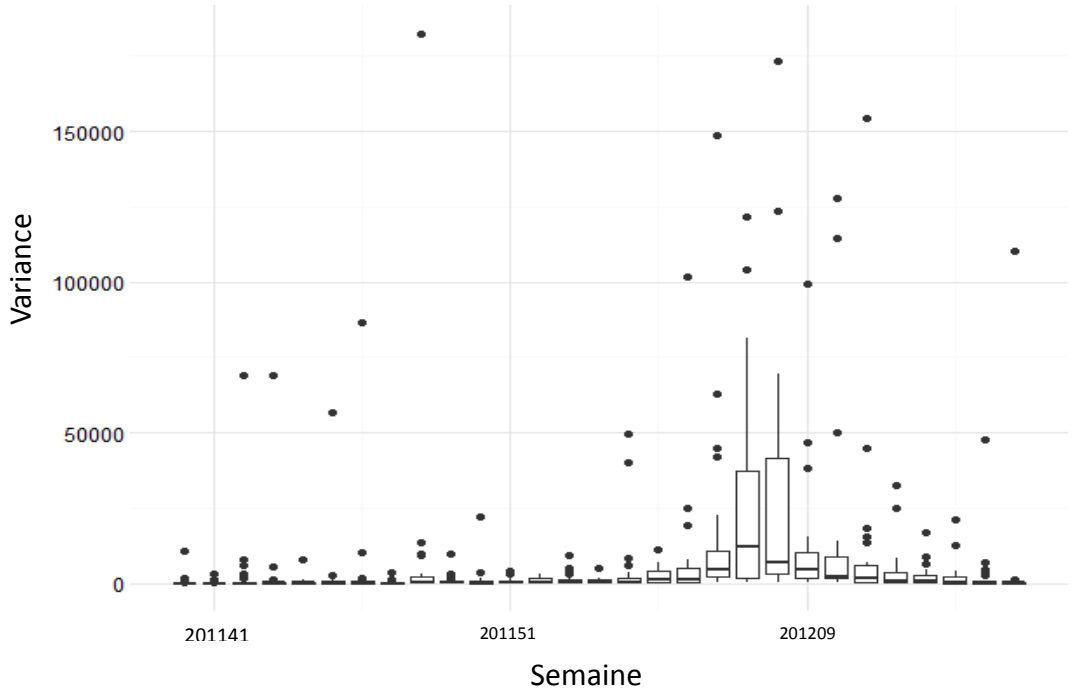
| Classe | Description |
|---|---|
| **A11G1** | Vitamine C pure |
| **J01A** | Tétracyclines et combinés |
| **J01C1** | Pénicilline orale à large spectre |
| **J01D1** | Céphalosporines orales |
| **J01F** | Macrolides et assimilés |
| **J01X9** | Autres antibactériens |
| **N02B** | Antipyrétiques non narcotiques |
| **R01A1** | Cortico-stéroïdiens nasaux |
| **R01A7** | Décongestionnants nasaux |
| **R01A9** | Autres préparations nasales topiques |
| **R01B** | Préparations nasales systémiques |
| **R02A** | Préparations pour la gorge |
| **R05C** | Expectorants |
| **R05D1** | Antitussifs purs |
| **R05D2** | Antitussifs combinés |
| **R05F** | Autres préparations pour la toux et le rhume |

**<u>Annexe III :</u> Ré-échantillonnage des médecins par zone pour déterminer l'effet pépite.**

Pour rappel, les données dont dispose le RS correspondent au nombre de cas déclarés par chaque médecin sur une période donnée. Un prétraitement des données permet d'obtenir le nombre de cas déclaré par médecin pour chaque semaine.

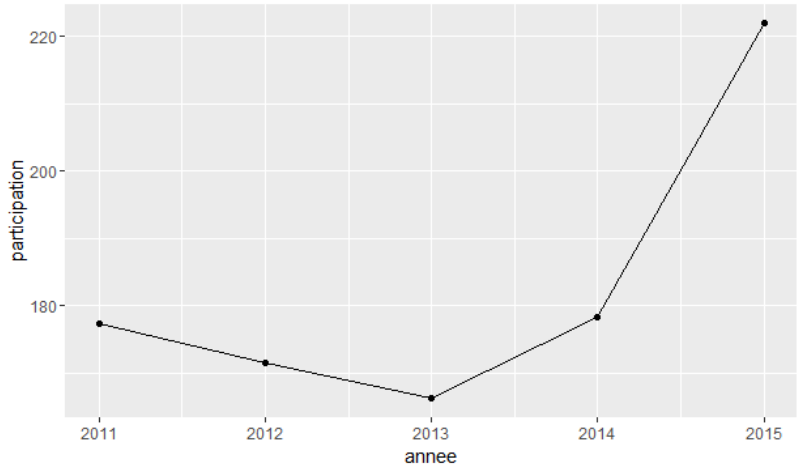
L'analyse est menée sur les données d'incidence départementales. L'objectif est de déterminer la valeur de l'effet pépite, c'est à dire la variabilité à micro-échelle. Pour cela, la variance des points de mesure est nécessaire, or une seule mesure est disponible par centroïde. Sur les données prétraitées, un sous-échantillonnage des médecins par département est donc réalisé pour estimer la variance des estimations d'incidence.

Par saison, les départements qui ont eu 2 déclarations ou plus chaque semaine sont sélectionnés. Cela représente 22 à 33 départements selon les saisons.
Les déclarations sont sous-échantillonnées à 75% pour chaque département sélectionné et les incidences départementales sont estimées à partir de ces données sous-échantillonnées. Cette opération est répétée 5000 fois par saison.
On obtient donc pour chaque département sélectionné et chaque semaine, 5000 estimations d'incidences basées sur les données sous-échantillonnées de façon aléatoire. Cela permet d'obtenir une variance des estimations d'incidence pour ces départements.



Variance des incidences de syndromes grippaux pour les départements sélectionnés

**<u>Annexe IV :</u> Participation hebdomadaire moyenne des médecins Sentinelles au cours du temps**




Participation hebdomadaire moyenne (en équivalent temps plein) des MG Sentinelles à la surveillance continue de 2011 à 2015.

Participation hebdomadaire moyenne (en équivalent temps plein) des MG Sentinelles à la surveillance continue en 2013, 2014 et 2015 par région française métropolitaine (Source : Bilan annuel 2015 du réseau Sentinelles)

| Région | | Participation hebdomadaire moyenne (ETP) en 2015 | Participation hebdomadaire moyenne (ETP) en 2014 | Participation hebdomadaire moyenne (ETP) en 2013 |
|---|---|---|---|---|
| **1** | Alsace | 5,9 | 5,2 | 5,1 |
| **2** | Aquitaine | 6,7 | 5,5 | 4,1 |
| **3** | Auvergne | 12,2 | 7,8 | 9,4 |
| **4** | Basse-Normandie | 7,5 | 4,9 | 2,3 |
| **5** | Bourgogne | 3,8 | 2,7 | 2,4 |
| **6** | Bretagne | 11,9 | 11,8 | 11,5 |
| **7** | Centre | 15,1 | 10,8 | 9,8 |
| **8** | Champagne-Ardenne | 6,0 | 3,2 | 3,0 |
| **9** | Corse | 10,7 | 9,1 | 6,0 |
| **10** | Franche-Comté | 5,6 | 5,1 | 5,5 |
| **11** | Haute-Normandie | 3,4 | 3,6 | 3,9 |
| **12** | Languedoc-Roussillon | 10,5 | 11,7 | 11,1 |
| **13** | Limousin | 4,7 | 4,3 | 3,5 |
| **14** | Lorraine | 5,7 | 4,6 | 5,1 |
| **15** | Midi-Pyrénées | 12,1 | 9,5 | 7,7 |
| **16** | Nord-Pas-de-Calais | 11,0 | 7,6 | 5,2 |

| | | | | |
|---|---|---|---|---|
| **17** | Pays de la Loire | 6,6 | 5,6 | 4,7 |
| **18** | Picardie | 2,9 | 2,8 | 2,3 |
| **19** | Poitou-Charentes | 2,8 | 2,6 | 2,9 |
| **20** | Provence-Alpes-Côte-D'azur | 16,4 | 12,3 | 10,7 |
| **21** | Ile-de-France | 23,9 | 18,1 | 17,2 |
| **22** | Rhône-Alpes | 36,3 | 29,6 | 32,8 |
| **France métropolitaine** | | **222,0** | **178,3** | **166,2** |

| | |
|---|---|
| | Diplôme : Ingénieur<br>Spécialité : Agronomie<br>Spécialisation / option : Statistique Appliquée<br>Enseignant référent : David CAUSEUR |
| Auteur(s) : Marie MORVAN<br>Date de naissance* : 31/12/1992 | Organisme d'accueil : Université Pierre et Marie Curie<br>Adresse : 27 rue Chaligny, 75012 Paris<br>Maître de stage : Clément TURBELIN, Pierre-Yves BOELLE |
| Nb pages : 23 Annexe(s) : 4 | |
| Année de soutenance : 2016 | |

Titre français : Evaluation des méthodes d'intégration de données permettant d'améliorer la prédiction spatiale des épidémies de grippe

Titre anglais : Evaluation of data integration methods allowing the improvement of spatial prediction of influenza epidemics

Résumé (1600 caractères maximum) :
La surveillance épidémiologique des syndromes grippaux est basée sur des réseaux de médecins généralistes bénévoles et permet l'estimation de la répartition des cas sur le territoire. Cependant, la répartition variable des médecins déclarants dans le temps et dans l'espace entraîne de l'imprécision dans l'interpolation spatiale des syndromes grippaux. Ce travail a pour objectif d'améliorer ces estimations spatiales en introduisant des données auxiliaires de vente de médicaments dans l'interpolation spatiale réalisée par krigeage. Après une analyse du lien entre l'incidence des syndromes grippaux et les délivrances médicamenteuses, ces données sont intégrées au modèle grâce à la méthode de cokrigeage. Une variable latente résumant l'ensemble des classes médicamenteuses est construite par PLS et intégrée au krigeage de l'incidence. La dynamique spatio-temporelle de l'épidémie est aussi étudiée grâce au krigeage spatio-temporel. Différents critères visuels et quantifiables et la représentation spatiale et temporelle des résultats de cokrigeage mettent en évidence une amélioration significative des estimations grâce à plusieurs classes médicamenteuses. Les meilleurs résultats sont obtenus avec la variable auxiliaire construite par PLS. Les améliorations concernent des zones ciblées pauvres en mesures pour les épidémies où la couverture spatiale des médecins déclarants n'est pas optimale. Le krigeage spatio-temporel est complexe à mettre en place mais les résultats montrent que son utilisation est possible pour étudier la dynamique d'une épidémie sur le territoire.

Abstract (1600 caractères maximum) :
Public health surveillance is based on computer networks of volunteer practitioners and allow the mapping of influenza-like illness (ILI) on a territory. However, the practitioners' declarations varying in space and time leads inaccuracy in the spatial interpolation of ILI. This works aims to improve the accuracy of these spatial predictions by introducing medication sales auxiliary variables in the spatial interpolation method called kriging. After analyzing the links between the medication sales and ILI incidence, the auxiliary variables are integrated by cokriging to the interpolation. A latent variable representing the medication sales is built using PLSR, and used as an auxiliary variable. The study of the spatio-temporal dynamic of an epidemic is done by spatio-temporal kriging. Several visual and quantitative criteria and the spatial and temporal representation of the results highlight a significant improvement of the predictions thanks to several medication classes. The best results are obtained with the latent variable. The improvements affect specific zones, where very few declarations are available in space and time. Spatio-temporal kriging is difficult to implement but the results show that this method allow the study of an epidemic on the territory.

Mots-clés : surveillance épidémiologique, syndromes grippaux, délivrances médicamenteuses, interpolation spatiale, cokrigeage

Key Words: public health surveillance, influenza-like illness, medication sales, spatial interpolation, cokriging

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*