Estimations intégratives des incidences régionales des syndromes grippaux en France par utilisation de données de délivrances médicamenteuses

Cyril Esnault

AGROCAMPUS OUEST

☐ **CFR Angers** ☑ **CFR Rennes**

Année universitaire : 2013 - 2014

Spécialité : Agronomie

Spécialisation : Statistique Appliquée

Mémoire de Fin d'Études

☑ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

☐ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

☐ d'un autre établissement (étudiant arrivé en M2)

Estimations intégratives des incidences régionales des syndromes grippaux en France par utilisation de données de délivrances médicamenteuses

Cyril ESNAULT

Soutenu à Rennes, le 3 septembre 2014

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)].
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).
Date et signature du maître de stage[(2)] :

Droits d'auteur :

L'auteur[(3)] autorise la diffusion de son travail

☑ 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)

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

☐ 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

Organisme d'accueil du stage : le réseau Sentinelles

Stage de fin d'étude au pôle « Système d'information et Biostatistique » du réseau Sentinelles.

Créé en 1984 à l'initiative conjointe de l'INSERM (Institut National Supérieur de la Recherche Médicale) et de l'UPMC (Université Pierre et Marie-Curie), le réseau Sentinelles est coordonné par l'équipe "Surveillance et Modélisation des maladies transmissibles" de l'IPLESP (Institut Pierre-Louis d'Epidémiologie et de Santé Publique, UMR S 1136), en collaboration avec l'InVS (Institut de Veille Sanitaire).

Remerciements

Un grand merci à Clément Turbelin – ingénieur de recherche au RS - et à Daniel Lévy-Bruhl – épidémiologiste à l'InVS - pour leur grande sympathie et leur confiance en mon égard en me permettant de m'approprier pleinement ce projet.

Merci à Clément Turbelin qui, malgré ses très nombreuses missions, a toujours su être disponible et avoir un regard avisé sur ma démarche.

Un grand merci aussi à Cécile Souty – ingénieur d'étude et doctorante au RS – pour sa présence aux réunions et ses remarques pertinentes concernant la rédaction de ce mémoire.

Un merci tout particulier à Pierre-Yves Boëlle – PU-PH à l'hôpital Saint-Antoine et responsable de l'équipe 1 de l'IPLESP – pour avoir été un réel soutien intellectuel, ainsi que pour sa grande disponibilité, que ce soit concernant ce travail ou mon projet professionnel.

Enfin, je ne peux pas clore cette page sans remercier plus généralement toute l'équipe du RS, avec en premier lieu Thomas Hanslik. – responsable du RS – Thierry Blanchon – responsable adjoint –, Esmeralda – secrétaire –, Yves Dorléans – assistant ingénieur -, ainsi que mes collègues – Caroline, Mathilde et Victoire. Leur générosité, leur accueil, et leur gentillesse m'ont particulièrement touché.

Liste des abréviations

RS : Réseau Sentinelles
MG : Médecin généraliste
MS : Médecin Sentinelles
SG : Syndromes grippaux
Inc100 : Taux d'incidence des SG pour 100 000 habitants
Val100 : Taux de délivrance médicamenteuse pour 100 000 habitants
YW : « YearWeek » ou « AnnéeSemaine », il s'agit de la nomenclature des semaines au sein du RS (ex : 201005 correspond à la 5ère semaine de l'année 2010)
CAH : Classification ascendante hiérarchique
ACP : Analyse en composantes principales
RMSEP : Racine carré de l'erreur quadratique moyenne de prédiction (root mean squared error)
BIC : Bayesian information criterion
IC : Intervalle de confiance

Sommaire

Introduction 1

I. Présentation du cadre du projet : Matériels et méthodes 3

1. Les données utilisées 3

1.1. Les données d'incidence des syndromes grippaux du réseau Sentinelles 3

1.2. Les données médicamenteuses : une base de donnée externe 5

2. La démarche globale entreprise 5

2.1. Présélection de classes médicamenteuses en lien avec les SG 7

2.1.1. La démarche de la présélection 7

2.1.2. Les méthodes de classification employées 7

2.2. Comparaison et sélection des meilleurs modèles de prédiction 9

2.2.1. Les types de modèles considérés 9

2.2.1.1. Le modèle de régression périodique log-linéaire 9

2.2.1.2. Le modèle de régression périodique log-non linéaire 11

2.2.1.3. Le modèle de régression PLS-Poisson 11

2.2.2. Les approches considérées 11

2.2.2.1. L'approche de type « modèle unique » 11

2.2.2.2. L'approche de type « modèles glissants à paramètres fixés optimisés » ... 13

2.2.2.3. L'approche de type « modèles glissants à paramètres non fixés optimisés » 13

2.2.3. Les critères de comparaisons de modèles 13

2.3. « Désagrégation » et « extrapolation » spatiales 15

2.3.1. Le principe de désagrégation spatiale 15

2.3.2. Le principe d'extrapolation spatiale 15

II. Résultats 17

1. La présélection de classes 17

1.1. Les classes présélectionnées sur la période 2004-2014 17

1.2. Les classes présélectionnées sur la période 2010-2014 17

2. La comparaison et sélection des meilleurs modèles de prédiction 19

2.1. Le meilleur modèle de prédiction 19

2.2. Affinement du meilleur modèle sélectionné 19

2.2.1. Le meilleur modèle prédictif des incidences nationales 19

2.2.2. Le meilleur modèle prédictif des incidences en Rhône-Alpes 21

2.3. Analyse du meilleur modèle sélectionné 21
2.3.1. Analyse du modèle prédictif des incidences nationales 21
2.3.2. Analyse du modèle prédictif des incidences en Rhône-Alpes 21
3. La désagrégation et l'extrapolation spatiales 23
3.1. Les nouvelles estimations des incidences régionales 23
3.2. Validations des nouvelles estimations des incidences régionales 25
3.2.1. Vérifications des hypothèses inhérentes à la désagrégation et à l'extrapolation 25
3.2.2. *Bootstrap* des nouvelles estimations des incidences régionales 27
III. Discussion 29
1. Retour sur les choix méthodologiques entrepris 29
2. Retour sur les résultats 31
Conclusions 35
Bibliographie 36
Annexes 38

Table des illustrations

Figure 1. Taux d'incidence des syndromes grippaux entre 201046 et 201345 0

Figure 2. Proportions régionales des médecins Sentinelles participant à la surveillance continue en 2013 par rapport à l'ensemble des MG en exercice dans la région concernée (en %). Source : Bilan Sentinelles 2013 2

Figure 3. Présentation de la démarche globale entreprise 4

Figure 4. Nombre de classes de médicaments sans données manquantes dans la base de données en fonction de la période d'étude considérée 6

Figure 5. Organisation du jeu de données afin de réaliser la présélection des classes médicamenteuses en lien avec les SG. Exemple de la période longue sur 2004-2014 6

Figure 6. Organigramme de sélection du meilleur modèle et de la meilleure approche pour la prévision des incidences des SG 8

Figure 7. Présentation de l'approche de type "modèle unique" 10

Figure 8. Présentation de l'approche de type "modèles glissants à paramètres fixés optimisés" 12

Figure 9. Présentation de l'approche de type "modèles glissants à paramètres non fixés optimisés". *Exemple de la prédiction de la semaine 201139 en se basant sur les 'j'=3 dernières semaines* 12

Figure 10. Présentation des hypothèses de la désagrégation et de l'extrapolation spatiales 14

Figure 11. Résultats de la présélection des classes sur la période longue (en haut) et la période courte (en bas) 16

Figure 12. Résultats de la comparaison d'approches et de modèles basées sur la qualité de prédictions des incidences nationales 18

Figure 13. Prédictions des taux d'incidence nationaux sur 201049-201348, à partir du meilleur modèle ('f' = 45 et 'n' = 3) 20

Figure 14. Prédictions des taux d'incidence en Rhône-Alpes sur 201051-201350, à partir du meilleur modèle ('f' = 50 et 'n' = 2) 20

Figure 15. Prédictions des taux d'incidence des SG dans les différentes régions par désagrégation (bleu) et extrapolation (rouge) spatiales sur 201046-201345 22

Figure 16. Pourcentage de fois que chaque classe soit présente dans les modèles nationaux (haut) ou dans les modèles en Rhône-Alpes (bas) 24

Figure 17. Corrélations entre les délivrances nationales (gauche) ou en Rhône-Alpes (droite) avec les délivrances dans les différentes régions, pour les classes présentes dans les modèles 24

Figure 18. *Bootstrap* des taux d'incidence obtenus par désagrégation spatiale sur 201046-201345. 26

Figure 19. *Bootstrap* des taux d'incidence obtenus par extrapolation spatiale sur 201051-201350. 26

Figure 20. Extrait de la caractérisation des 50 classes présélectionnées sur la période longue, par la fonction 'catdes' 30

Figure 21. Extraits de la caractérisation des 73 classes présélectionnées sur la période courte (en haut), et uniquement des 17 classes du panel d'experts (en bas), par la fonction 'catdes' . 30

Figure 22. Ecarts absolus moyens de prédictions des inc100 régionaux entre les méthodes de désagrégation et d'extrapolation spatiales, sur 201051-201345 32

Figure 23. Estimations des inc100 nationaux sur 201051-201350, par le RS (trait pointillé noir), par extrapolation spatiale (trait rouge), par le GROG (trait vert), et par recalcul à partir des inc100 régionaux issues de la désagrégation spatiale (trait bleu) 34

Tableau 1. Participation hebdomadaire moyenne (en ETP) des médecins Sentinelles à la surveillance continue en 2013 par région française métropolitaine, et évolution par rapport à 2012 et 2011. Source : Bilan Sentinelles 2013..... 2

Table des annexes

Annexe I. Tracés de la log-linéarité entre les taux d'incidence des SG et les taux de délivrance des classes. 38

Annexe II. Descriptions des classes présélectionnées sur 2004-2014 et valeurs propres de la CAH post-ACP (sortie "R"). 39

Annexe III. Descriptions des classes présélectionnées sur 2010-2014 (hors panel d'experts) et valeurs propres de la CAH post-ACP (sortie "R"). 39

Annexe IV. Les paramètres optimaux minimisant les erreurs de prédictions des inc100 nationaux sur 201039-201338, pour l'approche du type "modèle unique" 39

Annexe V. Les paramètres optimaux minimisant les erreurs de prédictions des inc100 nationaux sur 201039-201338, pour l'approche du type "modèles glissants à paramètres fixés optimisés" 39

Annexe VI. Comparaisons des erreurs de prédictions par les approches du type « modèles glissants », entre ceux "à paramètres fixés optimisés" et ceux "à paramètres non fixés optimisés ", sur 201139-201338 39

Annexe VII. Affinements au niveau national et en Rhône-Alpes du meilleur modèle : Obtention des paramètres 'f' et 'n' minimisant les erreurs de prédictions des inc100 39

Annexe VIII. Analyses du meilleur modèle de prédiction des incidences nationales : Résidus studentisés, distance de Cook et résidus partiels 39

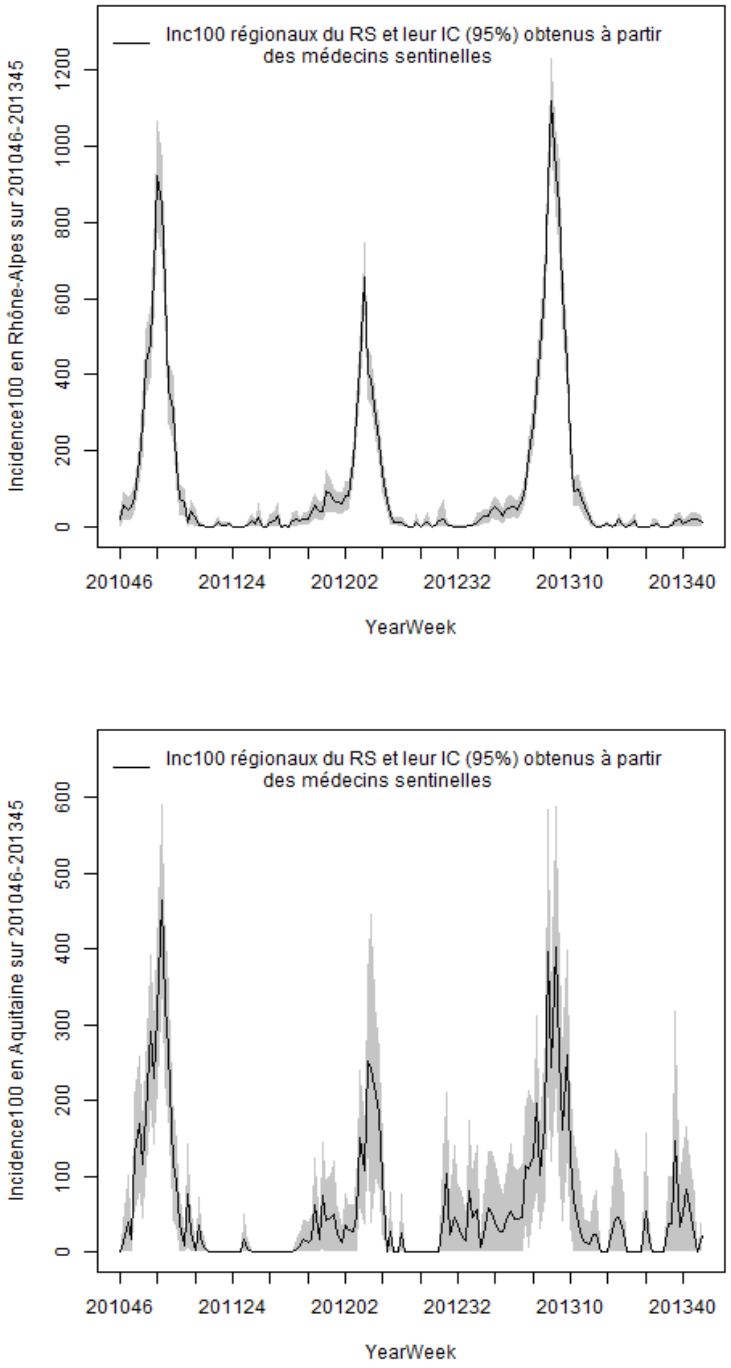
Annexe IX. Analyses du meilleur modèle de prédiction des incidences en Rhône-Alpes : Résidus studentisés, distance de Cook et résidus partiels 39

Annexe X. Prédictions des nouvelles incidences régionales par désagrégation et extrapolation spatiales 39

Annexe XI. Basic bootstrap des prédictions régionales des inc100 issues de la désagrégation (gauche) et de l'extrapolation (droite) spatiales 39

Annexe XII. La minimisation des erreurs de prédictions des inc100 nationaux par des modèles PLS-Poisson en utilisant toute la base de données médicamenteuses 39

Annexe XIII. Comparaisons des erreurs d'ajustement par les approches du type « modèles glissants », entre ceux "à paramètres fixés optimisés" et ceux "à paramètres non fixés optimisés ", sur 201139-201338 39



<u>Figure 1</u>. Taux d'incidence des syndromes grippaux entre 201046 et 201345
En haut : En région Rhône-Alpes (peu de fluctuations, stabilité « forte »).
En bas : En région Aquitaine (nombreuses fluctuations, stabilité « faible »).

'201046' : la 46ème semaine de l'année 2010
'201345' : la 45ème semaine de l'année 2013

Introduction

La **surveillance épidémiologique** se définit comme « *la collecte, l'analyse et la diffusion systématique des données sanitaires pour la planification, l'exécution et l'évaluation des programmes de santé publique* » (Thacker et al. 1988), et est en ce sens essentielle à toute politique de prévention et de lutte contre les maladies.

A l'exception des maladies à déclarations obligatoires qui font l'objet d'un recueil exhaustif des données par les agences régionales de santé, la surveillance épidémiologique se base plus généralement sur des échantillons de population. Pour cela, dans de nombreux pays, les réseaux basés en médecine générale permettent le recueil de données de surveillance de divers indicateurs de santé, à l'instar du réseau Sentinelles (RS) en France (source RS). Parmi les indicateurs surveillés par le RS, les syndromes grippaux (SG) sont ceux qui font l'objet de la plus longue période d'étude, ainsi que du plus grand nombre de publications chaque année. Les incidences des SG sont estimées et publiées hebdomadairement, aux niveaux national et régional, permettant ainsi d'avoir des informations sur la répartition spatiale de la maladie sur le territoire français. Les estimations des incidences s'obtiennent à partir des déclarations en temps réel de médecins généralistes (MG) bénévoles et volontaires, appelés « médecins Sentinelles » (MS). Les incidences étant dépendantes du nombre de MS participant durant une semaine donnée, la stabilité des estimations peut fluctuer dans le temps - certaines semaines ayant plus ou moins de participation par les MS -, et selon les régions, certaines possédant plus ou moins de MS (Figure 1).

Face à ce constat, il y a la nécessité d'améliorer les estimations des incidences régionales afin de les rendre plus stables et plus robustes dans le temps et l'espace. Un premier point de levier consiste à agir sur le recrutement de MS. Pour cela, divers moyens d'action sont envisagés, notamment celui de recruter les MS de manière plus stratégique, par le développement de méthodes d'optimisation spatiale du recrutement (Scarpino et al. 2012). De même, l'augmentation des effectifs de MS dans chaque région est un travail assez laborieux, nécessitant un contact au cas par cas avec les médecins. Des manœuvres facilitant le processus de recrutement doivent donc être développées, en collaborant par exemple avec des éditeurs de logiciels métiers, à l'instar de l'Angleterre où le GPRD (General Practitioners Research Database) constitue la plus grande base de données de patients au monde (5% de la population britannique couverte en 2006), par le recueil d'informations sur un même logiciel métier partagé par un réseau de MG (Valleron 2012).

D'autres stratégies ont également été abordées, visant cette fois non pas à agir sur le recrutement de MS, mais à lier la base de données Sentinelles avec d'autres sources d'information externes. Ainsi pouvons-nous citer la base de données de délivrances médicamenteuses en pharmacies où un premier travail a permis de mettre en évidence qu'un lien entre les incidences et les délivrances médicamenteuses pouvait être établi (Vergu et al. 2006). De même, les données issues des statistiques de requêtes auprès du moteur de recherche Internet « Google » pour certains mots clés - par exemple « grippe » - ont déjà fait l'objet d'études probantes au sein du RS (Pelat et al. 2009).

La mission qu'il m'a été confié au sein du RS est de développer une nouvelle méthode améliorant les estimations des incidences régionales des SG, par l'utilisation des données de délivrances médicamenteuses comme source d'information externe. Cette nouvelle méthode d'estimation n'est pas spécifique au seul RS et se veut généralisable à tout indicateur de santé concerné par la surveillance épidémiologique. Ce mémoire de fin d'étude s'attèle donc à présenter en premier lieu la démarche et les méthodes employées, puis, les résultats obtenus. Enfin, les choix méthodologiques et la pertinence des résultats seront discutés.

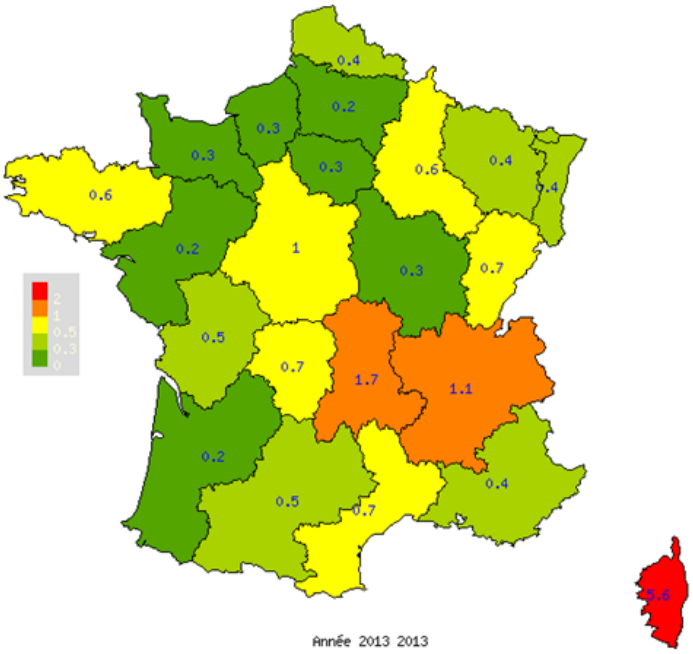


Figure 2. Proportions régionales des médecins Sentinelles participant à la surveillance continue en 2013 par rapport à l'ensemble des MG en exercice dans la région concernée (en %). Source : Bilan Sentinelles 2013

Tableau 1. Participation hebdomadaire moyenne (en ETP) des médecins Sentinelles à la surveillance continue en 2013 par région française métropolitaine, et évolution par rapport à 2012 et 2011. Source : Bilan Sentinelles 2013

Région		ETP(*) en 2013	ETP(*) en 2012	ETP(*) en 2011
1	Alsace	5,1	4,5	5,0
2	Aquitaine	4,1	4,9	6,6
3	Auvergne	9,4	10,4	9,2
4	Basse-Normandie	2,3	2,2	2,5
5	Bourgogne	2,4	2,1	3,1
6	Bretagne	11,5	11,9	14,4
7	Centre	9,8	9,3	7,2
8	Champagne-Ardenne	3,0	3,4	3,0
9	Corse	6,0	7,7	7,5
10	Franche-Comté	5,5	6,5	7,6
11	Haute-Normandie	3,9	3,2	3,0
12	Languedoc-Roussillon	11,1	11,0	10,1
13	Limousin	3,5	3,5	3,0
14	Lorraine	5,1	5,3	5,8
15	Midi-Pyrénées	7,7	6,4	4,3
16	Nord-Pas-de-Calais	5,2	2,9	4,3
17	Pays de la Loire	4,7	4,3	4,4
18	Picardie	2,3	2,4	3,4
19	Poitou-Charentes	2,9	2,5	3,2
20	Provence-Alpes-Côte-D'azur	10,7	13,3	14,5
21	Ile-de-France	17,2	19,9	22,6
22	Rhône-Alpes	32,8	33,9	32,9
France métropolitaine		**166,2**	**171.5**	**177,4**

(*)L'Equivalent Temps Plein est une mesure de la production, proportionnelle à l'activité du médecin

I. Présentation du cadre du projet : Matériels et méthodes

1. Les données utilisées

1.1. Les données d'incidence des syndromes grippaux du réseau Sentinelles

La surveillance des SG est basée sur le recueil en temps réel des informations issues de l'activité de plus de 1300 MS - soit 2,2% des médecins généralistes en France - dont environ 300 actifs hebdomadaires, constituant ainsi une base de 30 ans de données (source RS). Un MS considérera un cas de SG au sein de sa patientèle si celui-ci présente une fièvre supérieure à 39°C, accompagnée de myalgies et de signes respiratoires (source RS). La déclaration par le MS s'effectue alors par internet ou via un logiciel dédié (« jsentinel »).

Cette surveillance permet de détecter, d'alerter précocement et de prévoir la survenue d'épidémies de grippe. On définit l'épidémie comme l'augmentation rapide de l'incidence, à savoir le nombre de nouveau cas sur une période de temps donnée. La grippe se caractérise par un cycle annuel (ou saisonnier), les épidémies ayant lieu sur la période hivernale entre novembre et avril (source InVS). L'épidémie au sein du RS est déclarée par le dépassement d'un seuil épidémique, qui correspond au nombre de cas de SG auquel on s'attend en l'absence de circulation du virus grippal. Ce seuil est modélisé par régression périodique sur les incidences hebdomadaires observées dans le passé (Serfling 1963; Costagliola et al. 1991), hors périodes de fortes activités (inférieures à un taux fixé à 279 cas pour 100 000 habitants).

Le calcul hebdomadaire des incidences est réalisé à partir du nombre de cas *'C'* vu par chaque MS durant la semaine t, et de la participation *'P'* (proportion de la semaine couverte par les déclarations du médecin) de chaque MS à la semaine t. Le nombre moyen de cas *'Cm'* vu par MS dans une zone donnée *'z'* (ex : région) à la semaine t est alors estimé et l'incidence dans cette zone est obtenue par extrapolation à l'ensemble des médecins généralistes *'MG'* de la zone (Souty et al. 2014):

$$INC_z(t) = MG_z(t) * Cm_z(t) = MG_z(t) * \frac{\sum_{i=1}^{MS} C_{i,z}(t)}{\sum_{i=1}^{MS} P_{i,z}(t)}$$

Les incidences sont alors publiées hebdomadairement sous la forme de taux d'incidence, c'est-à-dire en nombre de cas pour 100 000 habitants. Notons que pour parler du '*taux d'incidence des cas de syndromes grippaux'*, nous parlerons plus brièvement de '*taux d'incidence'*.

En se basant sur les proportions régionales de MS ayant participé à la surveillance continue en 2013 (Figure 2), sur la participation hebdomadaire moyenne des MS par région entre 2011 et 2013 (en ETP, équivalent temps plein, Tableau 1), et au vu des graphiques des incidences hebdomadaires estimées dans chaque région (exemples en Figure 1), nous pouvons remarquer que la stabilité des estimations des incidences des SG n'est pas la même pour toutes les régions. Ainsi, nous pouvons identifier parmi les 22 régions étudiées (les régions métropolitaines, hors Outre-Mer) 4 régions de « confiance » : les régions PACA (Provence-Alpes-Côte d'azur), Auvergne, Corse et Rhône-Alpes. Pour ces régions, nous considérerons que les incidences du RS (*ie*, estimées à partir des MS) sont suffisamment stables et proches de la réalité. En revanche, du fait des nombreuses fluctuations observées dans les incidences pour les autres régions, nous considérerons que nous ne pouvons pas nous appuyer sur ces estimations.

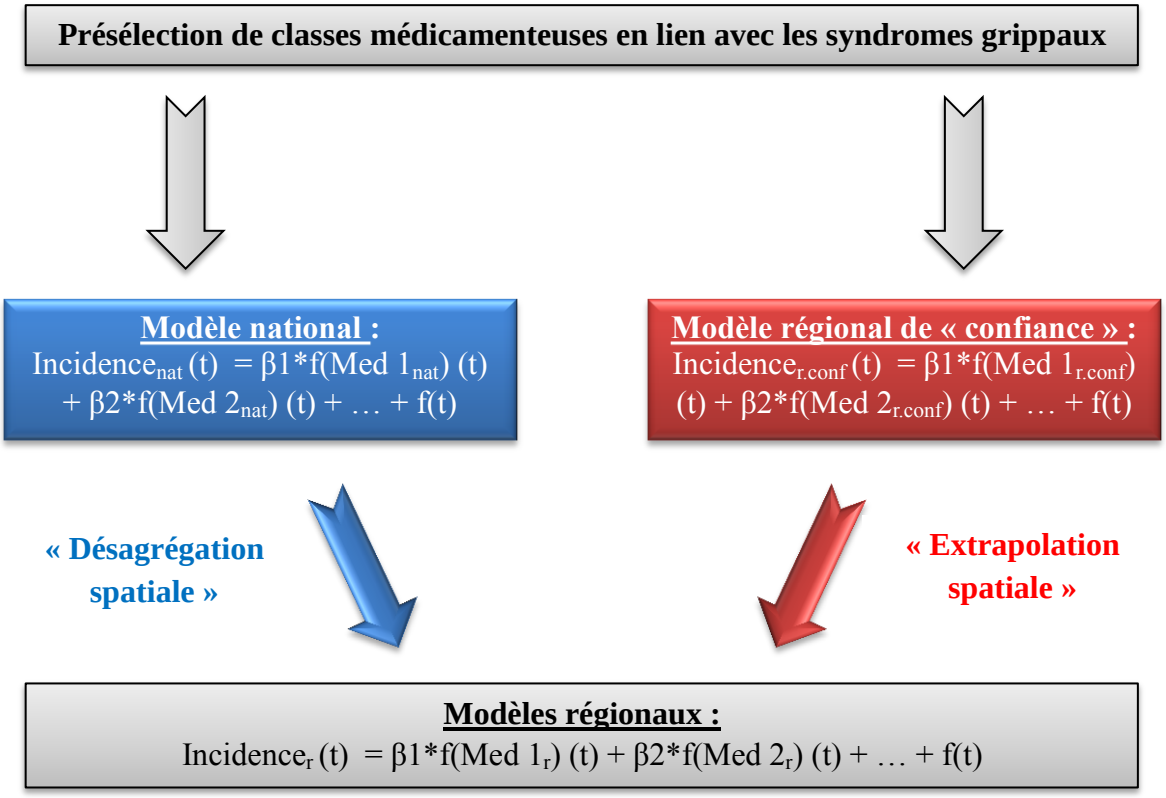


Figure 3. Présentation de la démarche globale entreprise

1.2. Les données médicamenteuses : une base de donnée externe

La base de données médicamenteuse est une source de données externe qui provient d'IMS France, filiale du Groupe IMS Health. Cette dernière est présente dans 135 pays depuis plus de 50 ans et est le leader mondial des données pharmaceutiques (source IMS Health). Cette base fournit les estimations des délivrances hebdomadaires en France de près de 500 « classes » de médicaments et sur plus de 10 ans de données, que ce soit au niveau régional qu'au niveau national. Pour cela, les estimations s'appuient sur le recueil en temps réel des délivrances médicamenteuses dans plus de 14 000 pharmacies (soit plus de 60% des pharmacies en France), ce qui en fait une source d'information très robuste (source IMS Pharmastat).

Les « classes » médicamenteuses considérées désignent des regroupements de médicaments utilisées selon certains critères pharmacologiques. Ainsi, certaines classes regroupent des médicaments contenant le même principe actif, d'autres des médicaments ayant la même action visée, d'autres encore des médicaments agissant sur des récepteurs semblables. Ces classes médicamenteuses correspondent au système de classification anatomique de produits pharmaceutiques développé par l'EphMRA (european pharmaceutical market research association).

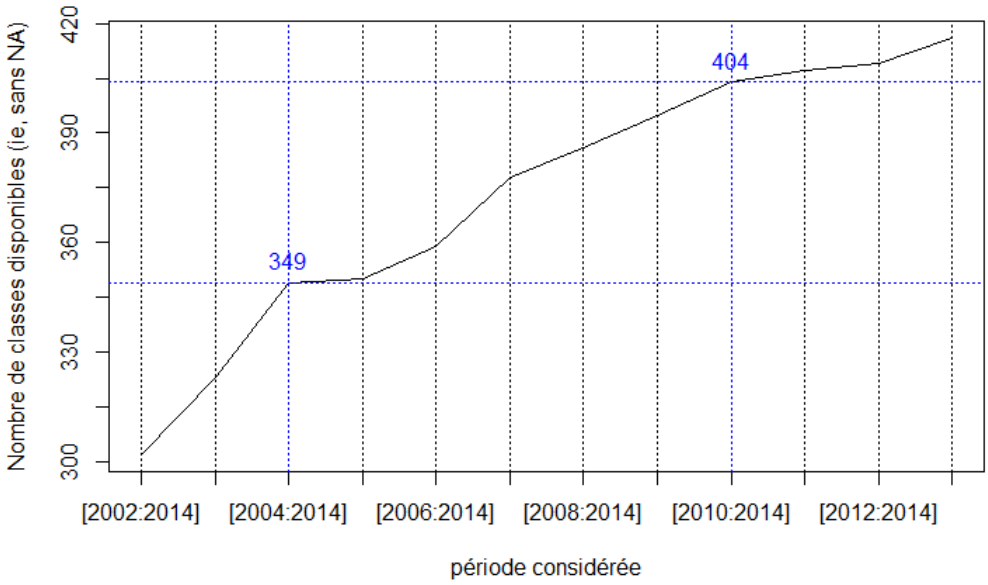
Les données utilisées seront le nombre d'unités de chaque classe médicamenteuse délivrée par semaine, exprimé en taux pour 100 000 habitants de la zone géographique considérée (nationale ou régionale).

2. La démarche globale entreprise

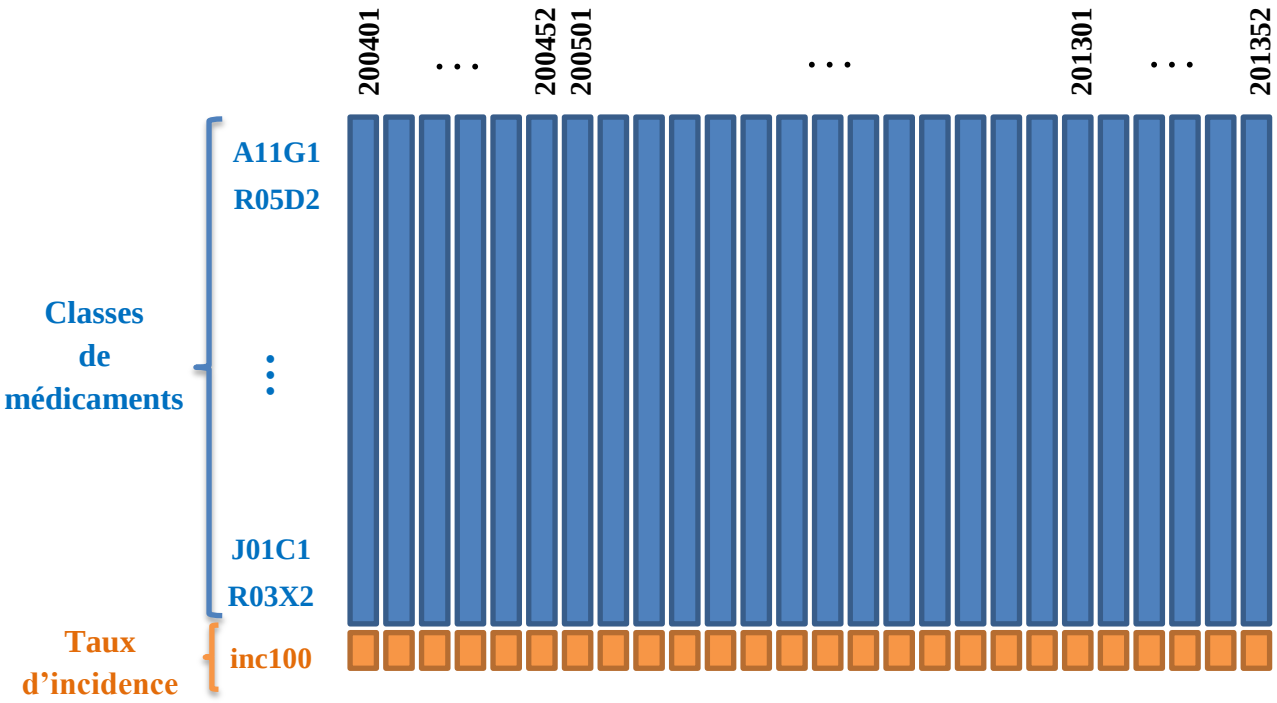
La Figure 3 présente la démarche globale de ce projet. L'objectif principal est d'estimer les incidences hebdomadaires régionales à partir des délivrances médicamenteuses et de fonctions temporelles. Pour cela, deux axes seront entrepris : l'un consistera en la modélisation des incidences nationales par les délivrances médicamenteuses, afin de « **désagréger** » ce modèle au niveau régional et ainsi obtenir une estimation des incidences hebdomadaires dans chaque région ; l'autre consistera cette fois en la modélisation des incidences dans les « régions de confiance » par les délivrances médicamenteuses, afin « **d'extrapoler** » ce modèle dans les autres régions et d'en estimer les incidences associées. La modélisation des incidences se fera en fonction de classes médicamenteuses au préalable **présélectionnées** comme étant liées aux SG (I.2.1.).

Aussi, ces deux axes nécessitent une modélisation la mieux prédictive des incidences nationales (resp. dans les régions de confiance) pour la désagrégation (resp. pour l'extrapolation) en fonction des classes médicamenteuses. Pour ce faire, une **étape de sélection de modèles et d'approches**, dont la présentation est faite en I.2.2., sera donc réalisée. Les concepts de « désagrégation spatiale » et « d'extrapolation spatiale », eux, seront présentés au I.2.3.

Une fois les nouvelles incidences régionales hebdomadaires obtenues à partir des modèles, nous tenterons de « **valider** » ces incidences (II.3.2.). Toutefois, ces validations prendront plus une allure de vérifications d'hypothèses : en effet, les « vraies » incidences n'étant pas connues, il sera relativement difficile d'émettre un jugement sur les nouvelles estimations des incidences régionales.



<u>Figure 4</u>. Nombre de classes de médicaments sans données manquantes dans la base de données en fonction de la période d'étude considérée



<u>Figure 5</u>. Organisation du jeu de données afin de réaliser la présélection des classes médicamenteuses en lien avec les SG. Exemple de la période longue sur 2004-2014

2.1. Présélection de classes médicamenteuses en lien avec les SG

L'objectif de la présélection des classes médicamenteuses est d'identifier un ensemble de classes qui soit le plus possible en lien avec les SG pour nos modèles. En effet, la modélisation des incidences par les données médicamenteuses nécessitera une règle d'inclusion particulière des classes dans les modèles - en se basant sur la corrélation -, ainsi qu'un ajustement des modèles sur des périodes pouvant être très petites (présentation au I.3.). Cette présélection vise donc à éviter l'inclusion dans les modèles de classes qui, sur une courte période, seront très corrélées à l'incidence mais qui seront peu informatives des SG. Notons que cette présélection de classes ne sera réalisée qu'au niveau national.

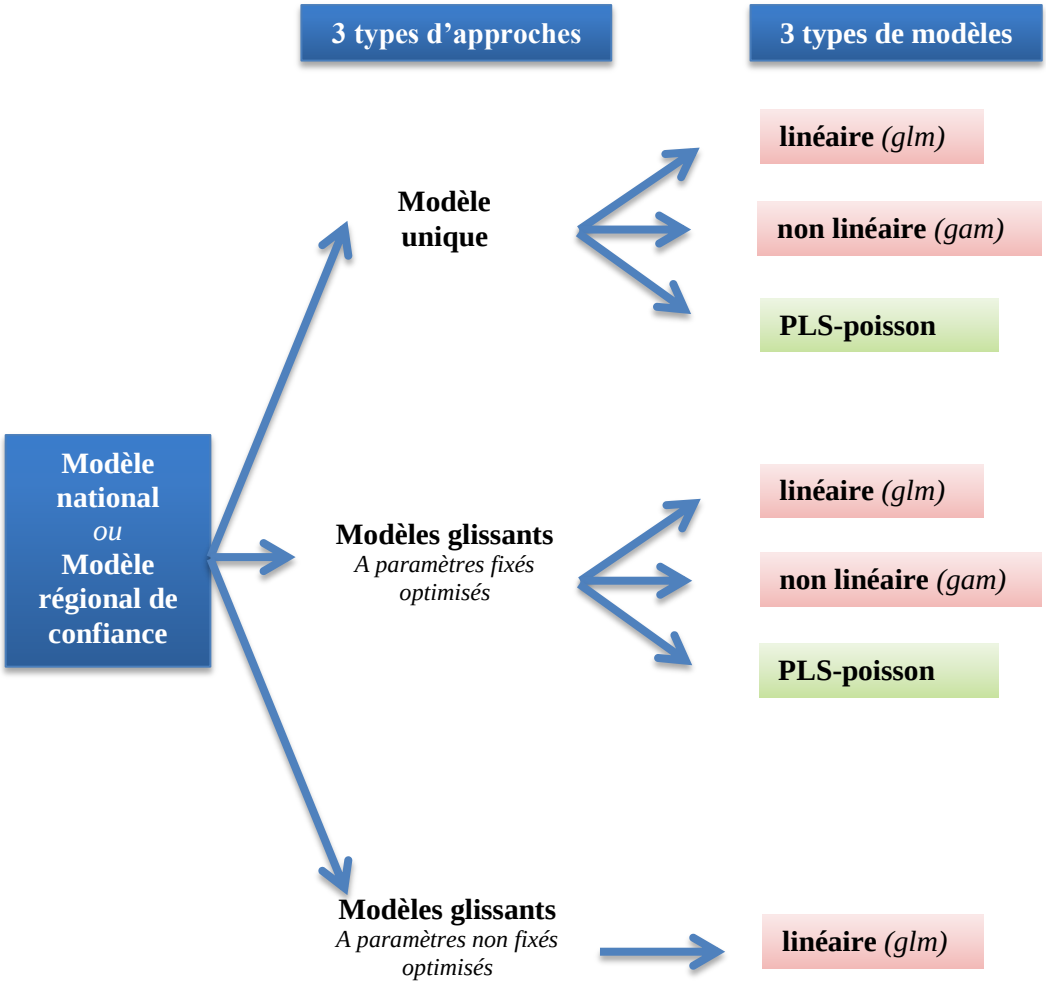
2.1.1. La démarche de la présélection

Les périodes sur lesquelles réaliser la présélection de classes sont obtenues à partir de la Figure 4. Celle-ci permet de maximiser au mieux le nombre de classes médicamenteuses disponibles (*ie*, sans données manquantes), ainsi que la longueur de la période d'étude. D'après ce graphique, on peut mettre en évidence 2 périodes intéressantes : la première, longue, s'étale sur 2004-2014, et peut ainsi prendre en compte une grande palette de variation des incidences ; la deuxième, plus courte, s'étale sur 2010-2014 et est par conséquent plus proche de l'actuel, et notamment de comment les médicaments sont délivrés de nos jours. Une présélection de classes sur ces 2 périodes sera donc réalisée.

Parmi les classes présélectionnées devront figurer celles sélectionnées par un panel d'experts du WHOCC (centre collaborateur de l'OMS pour la méthodologie sur les statistiques pharmaceutiques) comme étant prescrites contre la grippe (Vergu et al. 2006). Le descriptif de ces classes, au nombre de 17 (hors vaccins), est en Annexe II. Nous avons également envisagé que d'autres classes non spécifiquement prescrites contre la grippe pourraient avoir un lien avec les SG. Ainsi, une présélection de classes sera réalisée par des méthodes de classification, en s'inspirant de ce qui s'est fait pour la surveillance des gastro-entérites (Pelat et al. 2010). Pour ce faire, le jeu de donnée sera organisé afin d'avoir en individus statistiques l'ensemble des classes disponibles de la base, et en variables les semaines. De plus, les incidences des SG seront rajoutées parmi les individus (Figure 5). L'intersection d'une ligne et d'une colonne correspondra alors aux taux de délivrance nationaux (pour les classes) ou aux taux d'incidence nationaux (pour les SG) de la semaine considérée. Par conséquent, les méthodes de classification permettront de regrouper dans un même groupe l'incidence et les classes liées à celle-ci, permettant la présélection.

2.1.2. Les méthodes de classification employées

Diverses méthodes de classification seront utilisées. En premier lieu figurera la classification ascendante hiérarchique (CAH), où le critère d'agrégation utilisé sera l'indice de Ward permettant ainsi une meilleure homogénéité des groupes. Le niveau de coupure, lui, sera arbitraire : l'arbre de classification sera coupé de manière à avoir un groupe contenant l'incidence et un nombre de classes ni trop faible ni trop important (entre 10 et 60 classes). Une CAH post-ACP (Analyse en Composantes Principales) sera également effectuée afin de ne pas prendre en compte l'information contenue sur les dernières dimensions, celle-ci pouvant être considérée comme du « bruit » (Husson et al. 2010). La CAH sera alors réalisée sur les premières composantes principales uniquement, le choix de la méthode d'agrégation (Ward) et du nombre d'axe étant fondés sur l'inertie. Enfin, une autre méthode de type partitionnement par moyennes mobiles (K-means) sera également envisagée, qu'elle soit réalisée directement ou en consolidation d'une CAH par centres de classes initialisés.



<u>Figure 6</u>. Organigramme de sélection du meilleur modèle et de la meilleure approche pour la prévision des incidences des SG

Encadré rouge : modèle de régression par la méthode des moindres carrés
Encadré vert : modèle de régression des moindres carrés partiels

2.2. Comparaison et sélection des meilleurs modèles de prédiction

La recherche du meilleur modèle de prédiction des incidences nationales (afin de désagréger au régional) ou du meilleur modèle dans les régions de confiance (afin d'extrapoler dans les autres régions) sera réalisée en comparant 3 types d'approches et 3 types de modèles. Il est important de noter que l'ensemble de cette étape de recherche du meilleur modèle sera uniquement réalisée au national (pour désagréger par la suite) et non au niveau des régions de confiance pour l'extrapolation. Le résultat de cette recherche (approche et type de modèle) sera alors réutilisé pour le deuxième axe basé sur l'extrapolation à partir des régions de confiance. La Figure 6 présente la démarche entreprise pour la sélection du meilleur modèle prédictif des incidences.

2.2.1. Les types de modèles considérés

2.2.1.1. Le modèle de régression périodique log-linéaire

Le modèle log-linéaire établie la relation entre les taux d'incidence *'inc100'* (au log) de la semaine 't' et les taux de délivrance médicamenteuse (pour 100 000 habitants). S'agissant de données de comptage entières, il est fait l'hypothèse que les incidences suivent une loi de poisson, et par conséquent que la moyenne des incidences équivaut à la variance. Le lien logarithmique, lui, assure la positivité des incidences prédites.

$$log(inc100(t)) = \mu + \beta_1 * Med_1(t) + \beta_2 * Med_2(t) + \ldots\ldots + \beta_n * Med_n(t) + \beta cos_1 * cos(2*\pi*t/52{,}18) + \beta sin_1 * sin(2*\pi*t/52{,}18) + \beta cos_2 * cos(4*\pi*t/52{,}18) + \beta sin_2 * sin(4*\pi*t/52{,}18) + \beta cos_3 * cos(8*\pi*t/52{,}18) + \beta sin_3 * sin(8*\pi*t/52{,}18) + \epsilon(t)$$

avec μ une constante, 'Med i' le taux de délivrance de la classe de médicaments i et ϵ(t) l'erreur résiduelle (avec '52,18' correspondant au nombre moyen de semaines par an).

Ce modèle inclut 6 composantes temporelles et les taux de délivrance de 'n' classes de médicaments. Les composantes temporelles se décomposent en 2 composantes annuelles (βcos_1 et βsin_1), 2 composantes bisannuelles (βcos_2 et βsin_2) et 2 composantes trimestrielles (βcos_3 et βsin_3). Les 'n' classes de médicaments sont un sous-groupe de l'ensemble des classes présélectionnées en lien avec les SG. Il sera donc nécessaire de déterminer le nombre 'n' de classes qu'il faudra inclure dans les modèles pour optimiser la justesse des prédictions. De plus, l'ensemble des combinaisons de 'n' classes n'étant pas réalisable sur l'ensemble de notre étude, il sera donc inclus dans les modèles les 'n' classes les plus corrélées à l'incidence sur la période d'ajustement considérée. Ainsi, si nous prenons notre modèle avec 'n' égale 2, cela signifiera que seront inclues dans ce modèle les 2 classes les plus corrélées à l'incidence sur la période d'ajustement considérée, en plus des composantes temporelles.

Dans ce contexte, la gestion de la multicolinéarité est primordiale. En effet, la règle d'inclusion, qui stipule que sont présents dans les modèles les 'n' classes les plus corrélées à l'incidence, peut impliquer de la redondance d'information. De même, la présence de 6 composantes temporelles peut se révéler redondante, sachant que, les classes étant présélectionnées en amont sous critère de ressemblance avec les incidences (I.2.1), ces classes sont elles-mêmes porteuses d'information sur la saisonnalité des épidémies. Par conséquent, afin de prendre en compte cette multicolinéarité, une méthode de sélection pas-à-pas ascendante/descendante ('stepwise'), sous critère de minimisation du 'BIC', sera réalisée systématiquement. Enfin, les individus statistiques étant les semaines, on note que pour ce modèle il sera nécessaire d'ajuster au minimum sur 'n' + 6 (les composantes temporelles) + 1 (la constante) semaines.

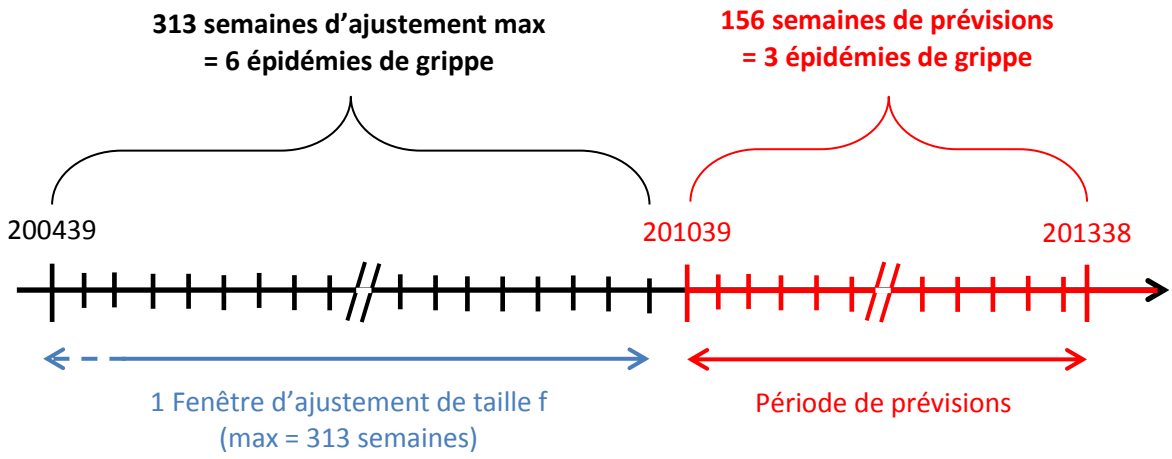


Figure 7. Présentation de l'approche de type "modèle unique"

2.2.1.2. Le modèle de régression périodique log-non linéaire

A partir des graphiques représentant le logarithme des taux d'incidence en fonction des taux de délivrance médicamenteuse (Annexe I), on constate que la linéarité ne semble pas toujours être vérifiée. Afin de prendre en compte une possible non-linéarité, il est laissé le choix au modèle log-non linéaire de considérer une relation linéaire ou non linéaire (splines) entre les 'n' classes du modèle et les incidences (au log).

$$log(inc100(t)) = \mu + \beta_1*(Med_{\,1}(t) \text{ ou } s(Med_{\,1}(t), df)) + \ldots\ldots + \beta_n*(Med_{\,n}(t) \text{ ou } s(Med_{\,n}(t), df)) + \beta cos_1*cos(2*\pi*t/52{,}18) + \beta sin_1*sin(2*\pi*t/52{,}18) + \beta cos_2*cos(4*\pi*t/52{,}18) + \beta sin_2*sin(4*\pi*t/52{,}18) + \beta cos_3*cos(8*\pi*t/52{,}18) + \beta sin_3*sin(8*\pi*t/52{,}18) + \epsilon(t)$$

Le choix du degré de liberté 'df' des fonctions splines est un paramètre à optimiser, afin de déterminer le niveau de lissage. On retrouve dans notre modèle les 6 composantes temporelles et les taux de délivrance de 'n' classes de médicaments les plus corrélées à l'incidence. Comme pour le modèle log-linéaire, l'incidence est supposée suivre une loi de Poisson et la gestion de la multicolinéarité est réalisée par sélection 'stepwise'.

2.2.1.3. Le modèle de régression PLS-Poisson

Une autre manière de capturer le lien entre les médicaments et les incidences des SG est de réaliser un modèle de régression des moindres carrés partiels pour modèles linéaires généralisés (ici, modèles de « Poisson »), à l'aide du package *plsRglm* (Bertrand et al. 2014). La multicolinéarité est ici prise en compte par la réduction de dimensions.

$$log(inc100(t)) = \mu + \beta_1 * nt_1(t) + \beta_2 * nt_2(t) + \beta_3 * nt_3(t) + \beta_4 * nt_4(t) + \ldots + \epsilon(t)$$

Les taux d'incidence 'inc100' sont ici fonctions de 'nt' variables latentes, toutes combinaisons linéaires de l'ensemble 'N' des classes de médicaments présélectionnées et des fonctions temporelles. Ainsi avons-nous pour chaque variable latente i :

$$nt_i(t) = \beta_{1i} * Med_{\,1}(t) + \beta_{2i} * Med_{\,2}(t) + \ldots + \beta_{Ni} * Med_{\,N}(t) + \beta cos_1*cos(2*\pi*t/52{,}18) + \beta sin_1*sin(2*\pi*t/52{,}18) + \beta cos_2*cos(4*\pi*t/52{,}18) + \beta sin_2*sin(4*\pi*t/52{,}18) + \beta cos_3*cos(8*\pi*t/52{,}18) + \beta sin_3*sin(8*\pi*t/52{,}18)$$

2.2.2. Les approches considérées

2.2.2.1. L'approche de type « modèle unique »

L'approche de type « modèle unique » consiste en la prévision des incidences sur 156 semaines (3 épidémies) entre la 39ème semaine de l'année 2010 et la 38ème semaine de l'année 2013, par un seul modèle ajusté sur les 'f' dernières semaines (Figure 7). La taille de la fenêtre 'f' - c'est-à-dire le nombre de semaines inclus dans le modèle pour l'ajuster - est donc à optimiser. De plus, pour les modèles log-linéaires et log-non linéaires, il faut également optimiser le nombre 'n' de classes de médicaments à inclure dans le modèle ; tandis que pour le modèle PLS-poisson, il faut optimiser le nombre 'nt' de valeurs latentes à inclure. Enfin, les modèles log-non linéaires nécessitent également l'optimisation du degré de liberté 'df'.

Par conséquent, il s'agira de **déterminer la meilleure combinaison de paramètres** (couple optimal **(f,n)** pour le modèle log-linéaire, **(f,nt)** pour le modèle PLS-Poisson et **(f, n, df)** pour le modèle log-non linéaire) en termes de prédiction des incidences sur les 156 semaines étudiées. Pour la compréhension dans la dénomination de cette approche, le terme « unique » vient du fait qu'on détermine au final un seul modèle pour la prédiction des 156 semaines, *a contrario* des autres approches.

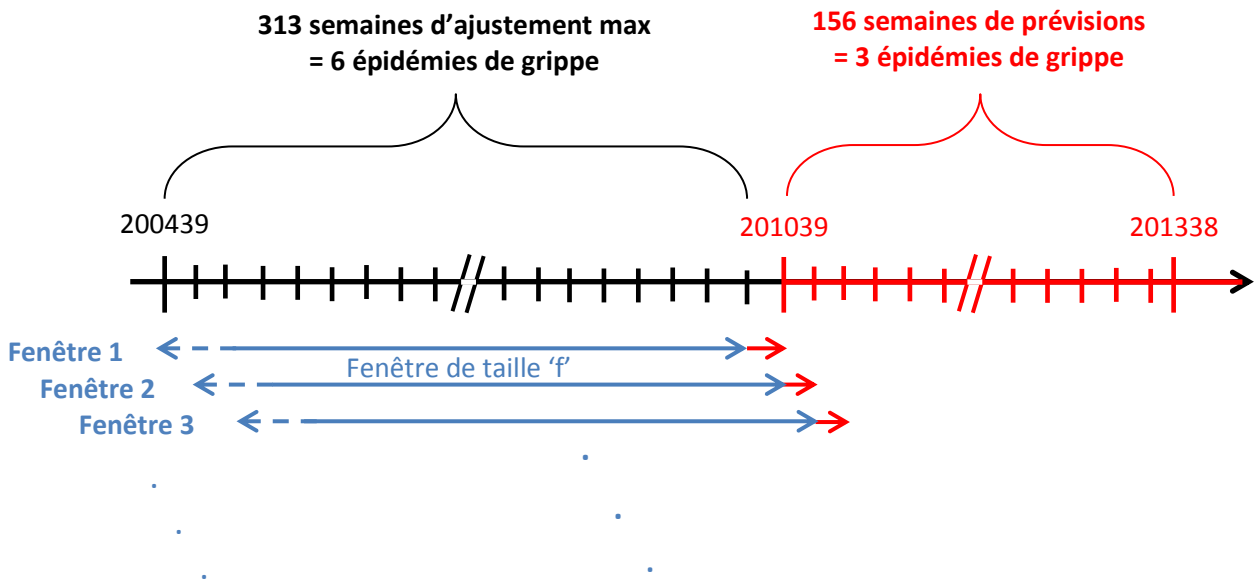


Figure 8. Présentation de l'approche de type "modèles glissants à paramètres fixés optimisés"

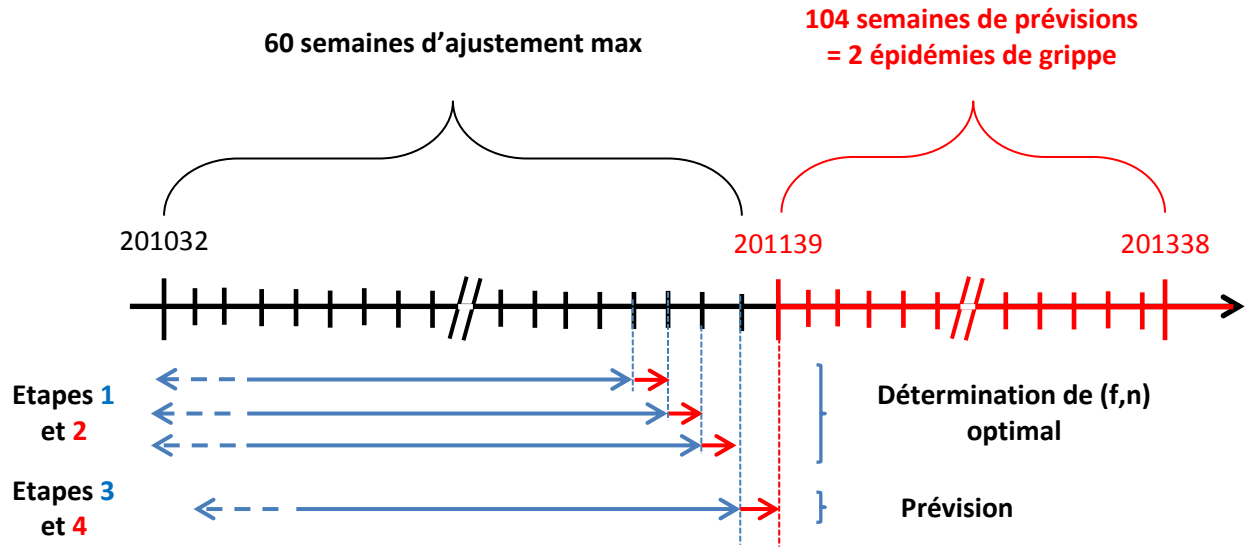


Figure 9. Présentation de l'approche de type "modèles glissants à paramètres non fixés optimisés". *Exemple de la prédiction de la semaine 201139 en se basant sur les 'j'=3 dernières semaines*

Etape 1 : Ajustements jusqu'en 201135, 201136, 201137 en fonction de 'f' et 'n'
Etape 2 : Prédictions en 201136, 201137 et 201138 ; détermination du couple (f,n) optimal sur ces 'j' = 3 semaines
Etapes 3 et 4 : Ajustement jusqu'en 201138 avec le couple (f,n) optimal et prédiction en 201139

2.2.2.2. L'approche de type « modèles glissants à paramètres fixés optimisés »

Afin de permettre l'évolution du modèle au cours du temps, l'approche de type « modèles glissants » consiste à réaliser chaque semaine un réajustement du modèle sur une fenêtre de taille 'f' pour prédire l'incidence (Figure 8). La prédiction des 156 semaines s'accompagne ainsi de l'ajustement de 156 modèles, conférant à cette approche un caractère non pas « figé » mais au contraire « glissant ». Il faudra cependant déterminer la combinaison optimale de paramètres ((f,n), (f,nt) ou (f,n,df) selon le type de modèle) qui prédit au mieux les incidences de la période de prévision. Cette approche est dite « à paramètres fixés » car la combinaison optimale recherchée est la même pour les 156 modèles.

2.2.2.3. L'approche de type « modèles glissants à paramètres non fixés optimisés »

Cette approche reprend le caractère glissant de l'approche précédente, où chaque incidence prédite est issue d'un modèle réajusté sur une taille 'f' de semaines et incluant un nombre 'n' de classes (ou, selon le type de modèle, un nombre 'nt ' de variables latentes). Toutefois, cette fois-ci les paramètres optimaux ne sont pas fixés mais, au contraire, re-déterminés chaque semaine en se basant sur les 'j' dernières semaines.

Pour mieux comprendre cette approche, la Figure 9 prend comme exemple la prédiction du taux d'incidence de la semaine 39 de l'année 2011 (201139), en se basant sur les 'j'=3 dernières semaines. Dans cet exemple, nous pouvons voir qu'il y a 4 étapes différentes : les deux premières consistent en l'application de l'approche précédente, mais non pas pour la prévision de 156 semaines mais de 'j' semaines. Ainsi, dans le détail, on ajuste 'j'=3 modèles (étape 1) jusqu'en 201135, 201136 et 201137, chacun en fonction de la taille 'f' et du nombre 'n'. Puis, on prédit la semaine suivante pour chacun de ces modèles, à savoir les semaines 201136, 201137 et 201138 (étape 2) et on en déduit ainsi le couple (f,n) optimal qui prédit au mieux les incidences de ces 'j'=3 semaines. La prédiction de la semaine 201139 est alors obtenue en ajustant le modèle jusqu'à la semaine 201138 avec les paramètres du couple optimal déterminé (étapes 3 et 4).

Cette approche, qui permet aux paramètres optimaux d'évoluer au cour du temps, nécessite cependant de **fixer le 'j' optimal** et donc de savoir sur combien de semaines passées doit se baser la prédiction des incidences hebdomadaires. Il est à noter que cette approche n'a été testée qu'avec des modèles log-linéaires (Figure 6), et sur seulement 104 semaines (2 épidémies), car celle-ci est bien trop couteuse en ressources avec des modèles log-non linéaires et de type PLS-Poisson.

2.2.3. Les critères de comparaisons de modèles

La sélection du meilleur modèle de prédiction des incidences repose sur la comparaison de modèles ajustés sur des données d'apprentissage et testés sur des données de validations (« test set »), en se basant sur deux critères de comparaisons. Le premier, et le plus important, est la Racine de l'Ecart Quadratique Moyen de Prédiction (ou RMSEP) qui correspond à l'écart entre les *valeurs* des incidences prédites et des incidences réelles : $\sqrt{\sum_i^n (y_i - \hat{y}_i)/n}$. Un deuxième critère a également été étudié, surtout d'un point de vue informationnel : le coefficient de corrélation linéaire de Pearson. Celui-ci correspond à l'écart entre les *variations* des incidences prédites et des incidences réelles : $\rho_{Y,\hat{Y}} = cov(Y, \hat{Y}) / \sigma_Y * \sigma_{\hat{Y}}$. Ainsi, la maximisation de ce dernier et la minimisation du premier permettent de sélectionner le modèle qui prédit le mieux les incidences réelles.

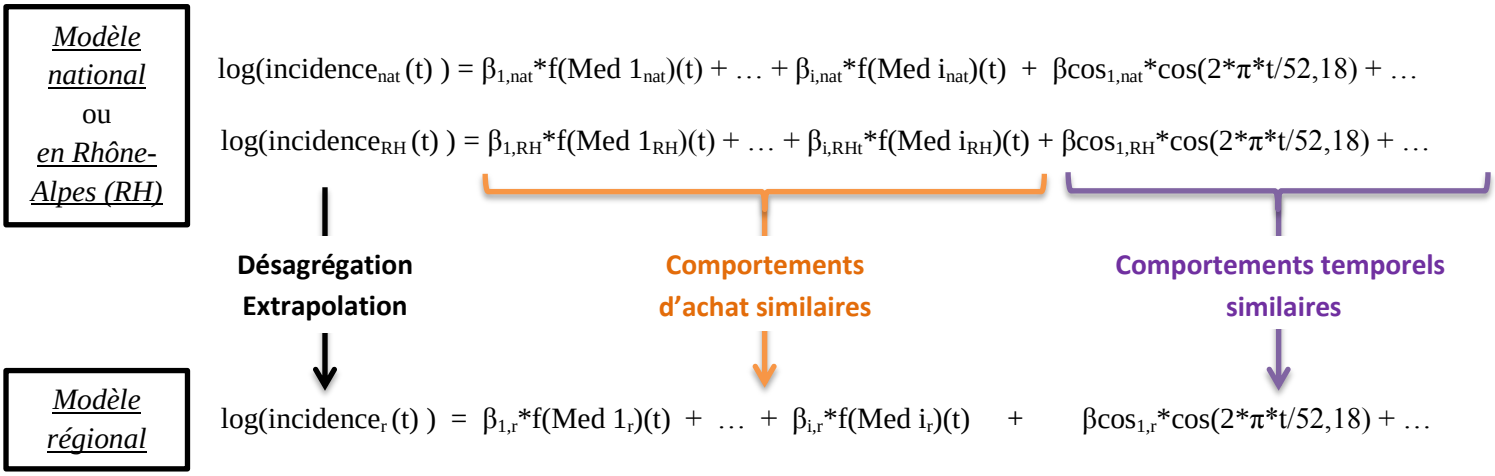


Figure 10. Présentation des hypothèses de la désagrégation et de l'extrapolation spatiales

2.3. « Désagrégation » et « extrapolation » spatiales

2.3.1. Le principe de désagrégation spatiale

La désagrégation de modèle est une problématique déjà abordée dans la littérature, notamment concernant la désagrégation temporelle de séries chronologiques (Sax & Steiner 2013), ou encore la désagrégation spatiale en hydrologie (Mehrotra & Singh 1998). Toutefois, la recherche bibliographique effectuée n'a pas su trouver de méthodes adaptées à notre cas.

Dans cette étude, la désagrégation spatiale du modèle national, pour l'estimation des incidences régionales, repose sur le postulat que les incidences nationales sont estimées correctement. A partir de cela, afin d'établir le lien entre le modèle national et les modèles régionaux (Figure 10), il est nécessaire d'établir quelques hypothèses. Si nous émettons les suppositions suivantes :

- *(H1) : Le comportement d'achat des médicaments face aux SG au niveau national correspond à ceux dans chaque région.*
- *(H2) : Le comportement temporel des SG au niveau national correspond à ceux dans chaque région.*

nous pouvons établir que *: chaque coefficient $\beta_{i,r}$ du modèle régional égal le $\beta_{i,nat}$ national.* Nous pouvons alors désagréger en appliquant directement le modèle national dans les différentes régions. Cela nécessitera toutefois une vérification des hypothèses *(H1)* et *(H2)*. Pour cela, les délivrances nationales des classes ayant participé à la désagrégation spatiale seront comparées aux délivrances de ces mêmes classes dans les différentes régions, en termes de corrélation. En effet, une bonne corrélation indiquerait des comportements d'achat similaires. Enfin, la vérification de l'hypothèse concernant le comportement temporel des SG s'effectuera indirectement, en comparant les estimations des incidences du RS dans les régions de confiance à celles obtenues par la désagrégation spatiale.

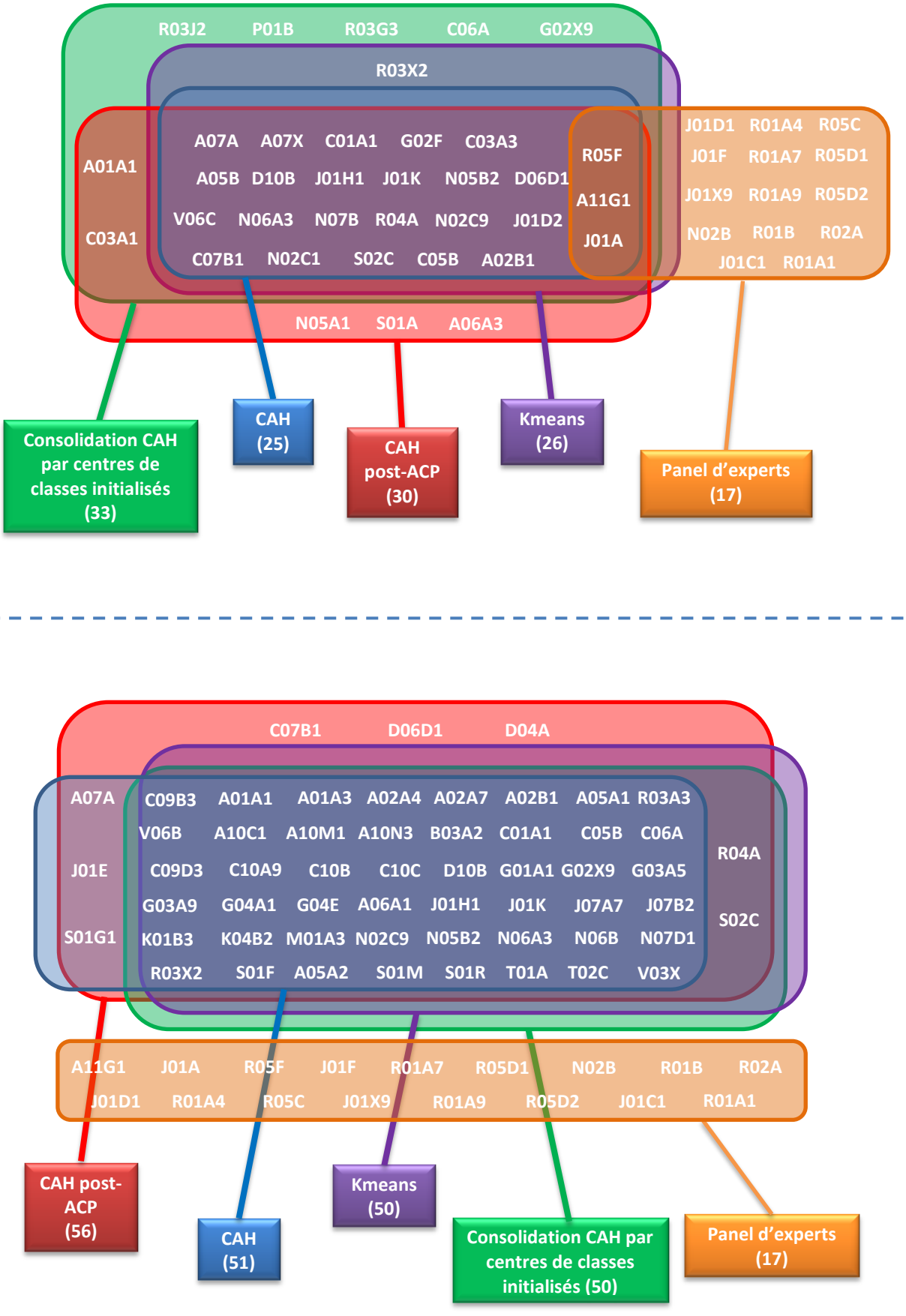
2.3.2. Le principe d'extrapolation spatiale

L'extrapolation spatiale à partir d'une région de confiance repose également sur le postulat que les incidences obtenues dans cette région sont bien estimées. Ce postulat semble à première vue plus raisonnable pour les régions de confiance que pour les incidences nationales, ces dernières étant issues de l'ensemble des incidences régionales pour lesquelles on n'a parfois que peu de confiance.

Parmi les régions de confiance, il a été fait le choix de réaliser l'extrapolation spatiale à partir de la région Rhône-Alpes, celle-ci surclassant toutes les autres régions en termes de participations hebdomadaires moyennes par les MS (Tableau 1). Pour réaliser l'extrapolation spatiale, on peut supposer que :

- *(H1') : Le comportement d'achat des médicaments face aux SG en Rhône-Alpes correspond à ceux dans chacune des autres régions.*
- *(H2') : Le comportement temporel des SG en Rhône-Alpes correspond à ceux dans chacune des autres régions.*

Ainsi pouvons-nous appliquer le modèle établi en Rhône-Alpes dans les autres régions, et même au niveau national. Cela nécessitera toutefois de vérifier les hypothèses *(H1')* et *(H2')*. Ces vérifications seront similaires à celles réalisées pour la désagrégation spatiale.



<u>Figure 11</u>. Résultats de la présélection des classes sur la période longue (en haut) et la période courte (en bas)

II. Résultats

1. La présélection de classes

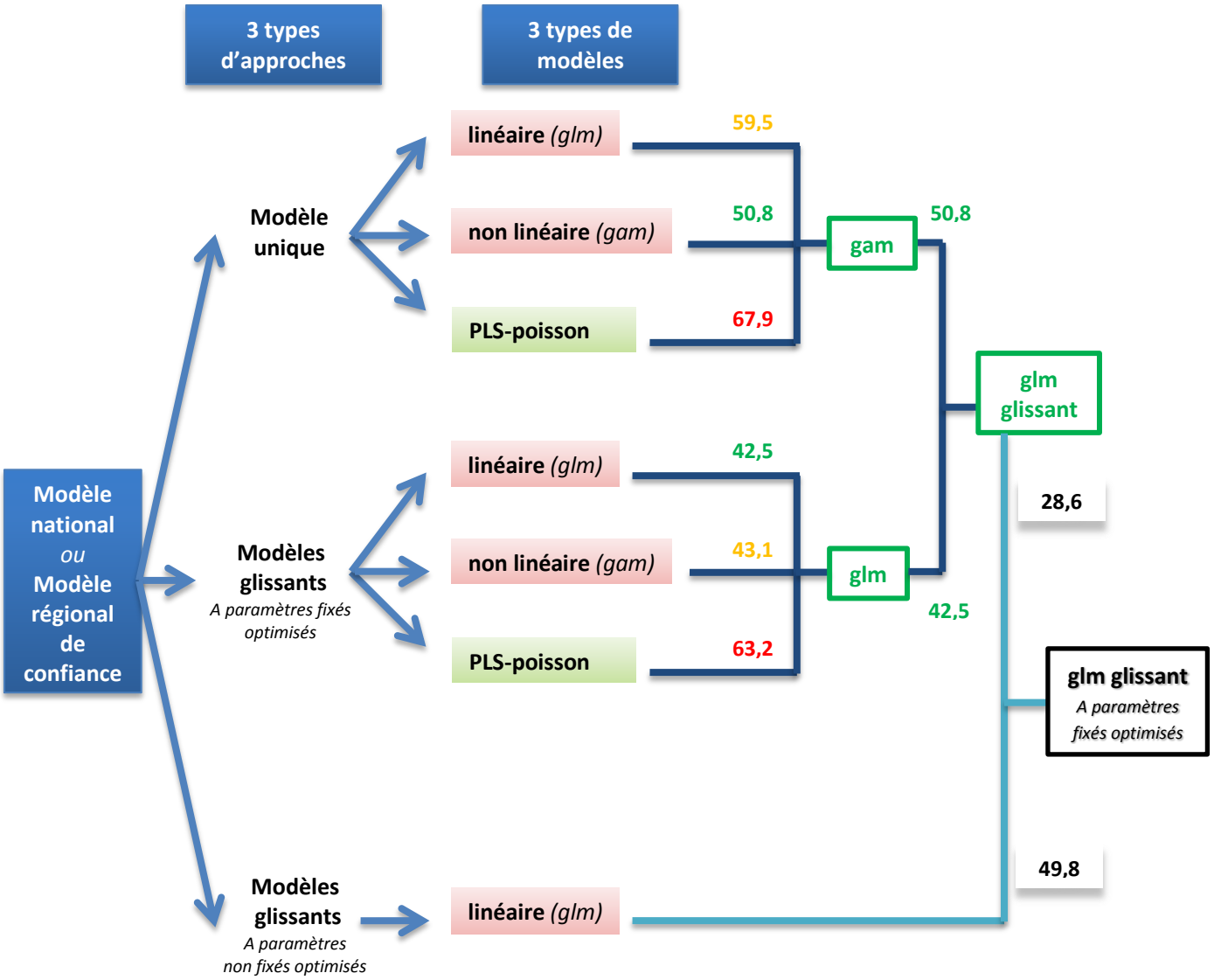
1.1. Les classes présélectionnées sur la période 2004-2014

Les différentes classifications réalisées ont permis la présélection de 36 classes médicamenteuses sur la période longue 2004-2014 (descriptions en Annexe II). En effet, une CAH a dans un premier temps permis la présélection de 25 classes. Puis, une CAH post-ACP a permis de présélectionner 5 classes supplémentaires (les 25 de la CAH + 5). Pour cela, seules les 3 premières dimensions ont été conservées où 99,63% de l'inertie totale est représentée, le reste pouvant être considéré comme du « bruit » (Annexe II). En consolidant la CAH par une K-means avec initialisation des centres de classes trouvées par la CAH, 5 nouvelles classes ont été présélectionnées en lien avec les incidences des SG. Enfin, une segmentation par la méthode des K-means en indiquant le nombre de classe (K = 14) a permis la sélection d'une classe supplémentaire, ce qui fait bien au total 36 classes présélectionnées par classifications.

L'ajout des 14 classes du panel d'experts (3 étant déjà en commun avec les classes présélectionnées par classifications) nous a conduit à **50 classes présélectionnées** pour la période 2004-2014 (Figure 11 en haut).

1.2. Les classes présélectionnées sur la période 2010-2014

Le même travail a été réalisé sur la période courte 2010-2014. Cette fois, la CAH a permis la présélection de 51 classes. La consolidation de la CAH par centres de classes initialisés, ainsi que la segmentation par méthode de K-means, ont permis la présélection des mêmes 50 classes, dont 47 en commun avec celles de la CAH seule. Au final, avec les classes issues de la CAH post-ACP (2 dimensions conservées), 56 classes ont ainsi été présélectionnées par les différentes méthodes de classifications (descriptions en Annexe III), ce qui, en ajoutant les 17 classes du panel d'experts, fait un total de **73 classes présélectionnées** sur cette période (Figure 11, en bas).



<u>Figure 12</u>. Résultats de la comparaison d'approches et de modèles basées sur la qualité de prédictions des incidences nationales

Les valeurs au-dessus des branches correspondent aux RMSEP minimaux associés aux paramètres optimaux trouvés

<u>Branches bleues claires</u> : prévisions des semaines 201139-201338, avec les classes présélectionnées sur la période courte, et avec une recherche non exhaustive de la fenêtre d'ajustement 'f' (un pas de 10 semaines)

<u>Branches bleues foncées</u> : prévisions des semaines 201039-201338 ; avec les classes présélectionnées sur la période longue, et avec une recherche exhaustive de la fenêtre d'ajustement 'f'

2. La comparaison et sélection des meilleurs modèles de prédiction

2.1. Le meilleur modèle de prédiction

Les résultats de la recherche du meilleur modèle sont résumés sur l'organigramme en Figure 12. Les nombres au-dessus des branches correspondent aux RMSEP associées aux paramètres optimaux ('f', 'n' ou 'nt', 'df'), ou, dit autrement, aux erreurs de prévisions les plus faibles obtenues pour un type d'approche et un type de modèle donnés. Les résultats spécifiques à chaque approche et chaque modèle sous formes graphiques sont tous en Annexe IV, Annexe V et Annexe VI. Sur l'organigramme, les branches de couleur bleu foncé spécifient que les comparaisons réalisées se basent sur la prévision de 3 épidémies (156 semaines) et que la recherche des paramètres optimaux s'est effectuée sur la période longue (2004-2014), avec l'utilisation des 50 classes présélectionnées sur cette période. En revanche, les branches bleues claires se basent uniquement sur la prévision de 2 épidémies, sur la période courte (2010-2014), où ont été utilisées les 73 classes présélectionnées sur cette période.

Un premier résultat est que l'approche dite du « modèle unique » prédit mieux les incidences nationales si le modèle est log-non linéaire (rmsep = 50,8), alors que ce sont les modèles log-linéaires les mieux prédictifs avec l'approche dite des « modèles glissants à paramètres fixés optimisés » (rmsep = 42,5). Par ailleurs, si on compare ces deux approches, on remarque que les modèles glissants de type log-linéaire sont ceux qui prédisent le mieux les incidences. Si on la compare désormais avec la 3ème approche dite « des modèles glissants à paramètres non fixés optimisés », on observe que cette dernière est moins bien prédictive des incidences (rmsep de 49,8 comparée à 28,6). Ainsi, ne pas fixer le couple (f,n) et le laisser se ré-optimiser chaque semaine ne semble pas être une meilleure approche.

Par conséquent, la comparaison d'approches et de modèles nous amène à conclure que le meilleur modèle en termes de prédiction des incidences nationales est le modèle log-linéaire glissant à paramètres fixés optimisés.

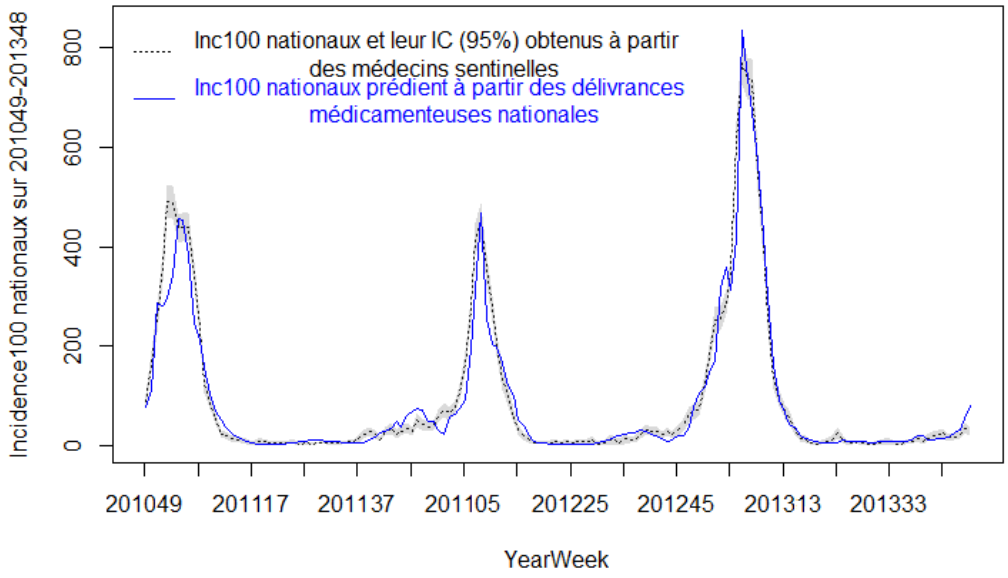
2.2. Affinement du meilleur modèle sélectionné

2.2.1. Le meilleur modèle prédictif des incidences nationales

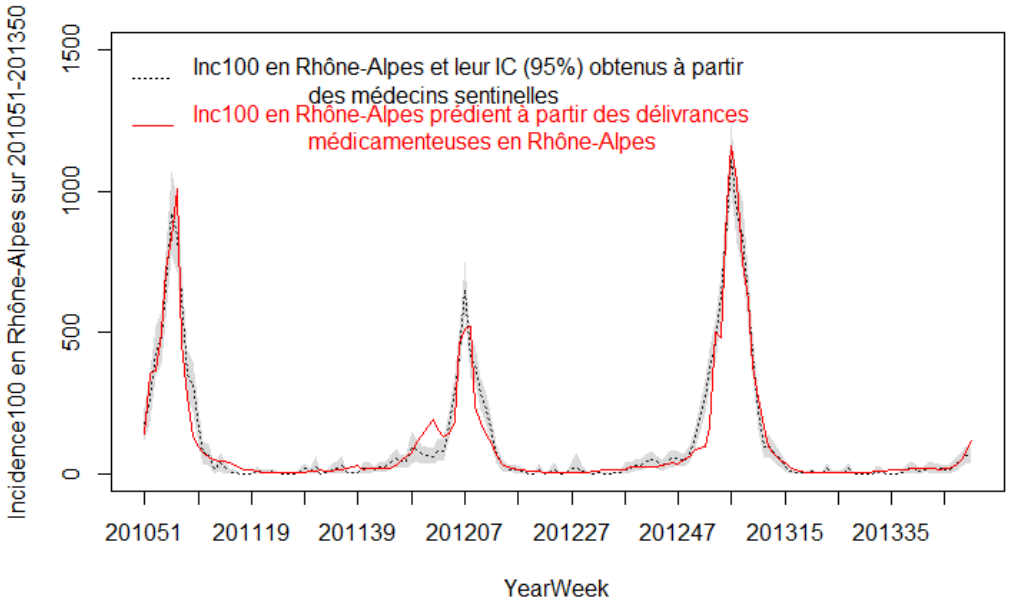
L'étape de sélection de modèle nous a permis de trouver le meilleur modèle prédictif des incidences. Toutefois, pour des soucis techniques, cette sélection s'est soit effectuée en se basant sur les prévisions de 3 épidémies mais avec une recherche non exhaustive du couple (f,n) optimal (variation de la taille de la fenêtre d'ajustement 'f' avec un pas de 10 semaines), soit avec une recherche exhaustive du couple (f,n) mais en se basant uniquement sur la prévision de 2 épidémies. Nous allons donc affiner la recherche du couple (f,n) optimal en prédisant cette fois 3 épidémies de SG de 201049 à 201348.

Un premier résultat est qu'à partir de 6 classes de médicaments inclues dans les modèles, le sur-ajustement entraine des prévisions parfois moins bonnes que si les incidences n'étaient fonctions que des variables temporelles, sans inclusion de classe (surtout vrai à partir de 8 classes inclues, Annexe VII).

De plus, la minimisation des erreurs de prédictions des 3 épidémies est obtenue en incluant 'n' = 3 classes dans les modèles et en ajustant sur les 'f' = 45 dernières semaines (RMSEP = 37,37). Les prédictions associées (trait plein bleu) à ce couple optimal sont données en Figure 13.



<u>Figure 13</u>. Prédictions des taux d'incidence nationaux sur 201049-201348, à partir du meilleur modèle ('f' = 45 et 'n' = 3)



<u>Figure 14</u>. Prédictions des taux d'incidence en Rhône-Alpes sur 201051-201350, à partir du meilleur modèle ('f' = 50 et 'n' = 2)

2.2.2. Le meilleur modèle prédictif des incidences en Rhône-Alpes

L'approche et le type de modèle sélectionnés au niveau national sont ceux utilisés pour l'ajustement en Rhône-Alpes, afin de réaliser l'extrapolation spatiale. La minimisation des erreurs de prédictions est cette fois-ci associée à un ajustement sur les 'f' = 50 dernières semaines et un nombre 'n' de classes inclus dans les modèles égale à 2 (rmsep = 45,7, Annexe VII). Les prédictions associées (trait plein rouge) sont observables en Figure 14.

2.3. Analyse du meilleur modèle sélectionné

L'objectif de cette partie est d'améliorer la compréhension de notre meilleur modèle, que ce soit au niveau national qu'en Rhône-Alpes. Ne pouvant pas analyser le modèle de chaque incidence prédite (156 modèles), il a été fait le choix de prendre deux semaines en exemple : les modèles prédisant les incidences en 201124 (période hors épidémie) et en 201308 (période épidémique). Tous les résultats figurent en Annexe VIII et Annexe IX.

2.3.1. Analyse du modèle prédictif des incidences nationales

- **Analyse en 201124 : période hors épidémie**

L'analyse des résidus studentisés en fonction du temps indique une petite structuration en forme sinusoïdale laissant présager l'existence d'une autocorrélation des résidus non suffisamment captée dans le modèle. De plus, bien qu'il soit tout à fait normal d'avoir α% de valeurs aberrantes (2-3 valeurs sur les f=45 semaines ajustées, α=0.05), on remarque que trois semaines associées au démarrage de l'épidémie sont particulièrement mal ajustées par le modèle : 201032, 201045, 201102. Cette dernière participe de surcroit très fortement avec la semaine 201101 dans l'ajustement du modèle. Enfin, bien que 3 classes aient été initialement inclues dans le modèle, la sélection stepwise sous critère 'BIC' n'a conservé que la classe 'R05D1'. L'analyse des résidus partiels en fonction des taux de délivrance laisse entrevoir une courbe en cloche, ce qui sous-entend qu'une transformation serait ici envisageable.

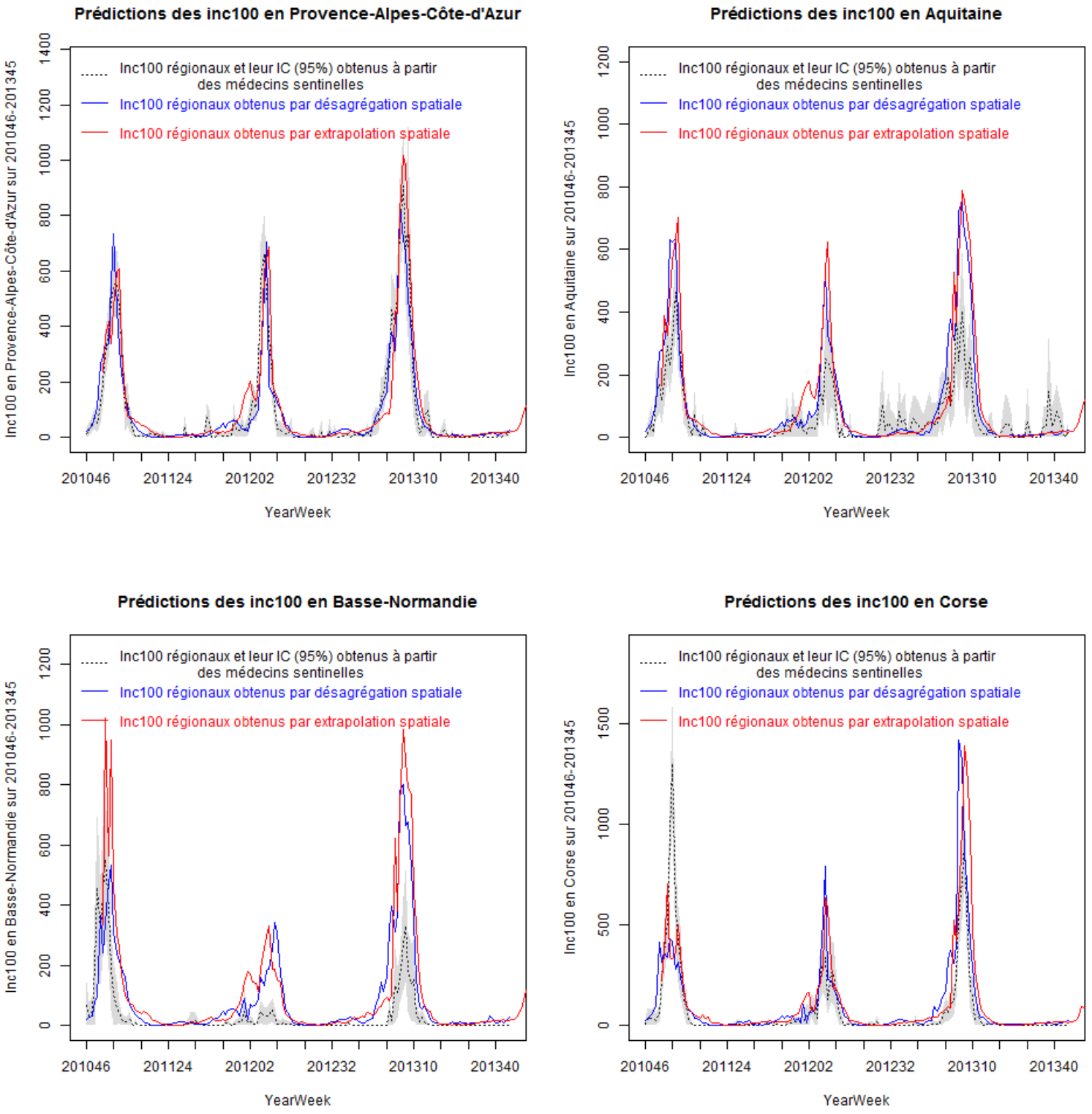
- **Analyse en 201308 : période épidémique**

L'analyse des résidus studentisés en fonction du temps ne montre pas cette fois de structuration particulière : l'hypothèse d'homoscédasticité est donc bien vérifiée. En revanche, on observe 4 semaines aberrantes qui semblent confirmer un assez mauvais ajustement de notre meilleur modèle au démarrage des épidémies. Par ailleurs, on observe également que ces 4 semaines sont celles qui influencent le plus le modèle. Les semaines épidémiques sont donc celles qui ont le plus de poids dans l'ajustement de notre meilleur modèle. Enfin, dans cet exemple, seules les classes 'A06A1' et 'R05D1' sont présentes dans le modèle après sélection stepwise. L'analyse des résidus partiels de ces 2 classes semblent ne pas indiquer le besoin d'une transformation particulière de ces classes.

2.3.2. Analyse du modèle prédictif des incidences en Rhône-Alpes

- **Analyse en 201124 : période hors épidémie**

L'analyse des résidus studentisés en Rhône-Alpes indique que de très nombreuses semaines sont mal expliquées par le modèle, avec là encore une structuration sinusoïdale. Par ailleurs, on observe également que plusieurs semaines ont un poids très important dans l'ajustement du modèle, la majorité étant des semaines épidémiques. Enfin, l'analyse des résidus partiels des classes présentes dans ce modèle ('R05D1' et 'A11G1') semblent suggérer que des transformations seraient nécessaires, principalement pour les faibles valeurs de délivrances.



<u>Figure 15</u>. Prédictions des taux d'incidence des SG dans les différentes régions par désagrégation (bleu) et extrapolation (rouge) spatiales sur 201046-201345

De gauche à droite et de haut en bas : En PACA, Aquitaine, Basse-Normandie et Corse

- **Analyse en 201138 : période épidémique**

De très nombreuses semaines sont ici aussi mal expliquées par le modèle. Concernant la normalité des résidus, la probabilité critique au test de Shapiro-Wilk est de 0,04, et donc à la limite de la non-significativité. En outre, l'analyse des résidus partiels de la seule classe présente dans ce modèle ('A11G1') indique la nécessité d'une transformation de cette variable dans l'espoir d'améliorer les prédictions des incidences des SG.

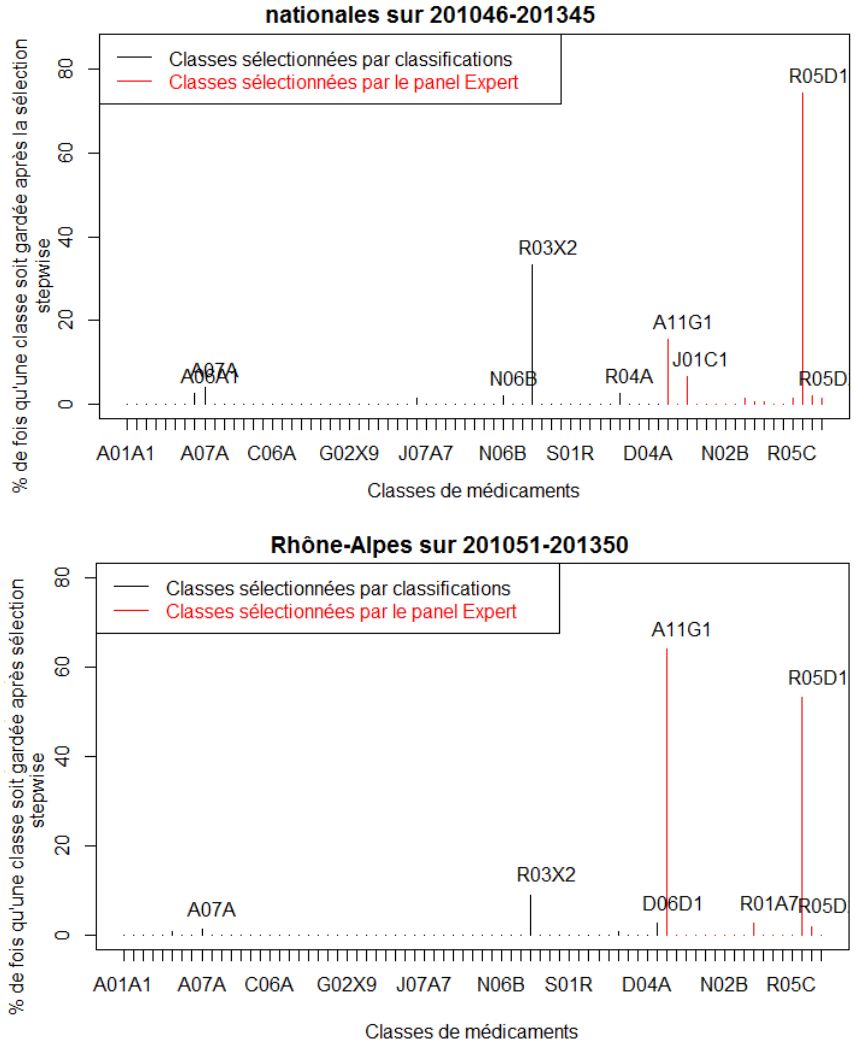
3. La désagrégation et l'extrapolation spatiales

3.1. Les nouvelles estimations des incidences régionales

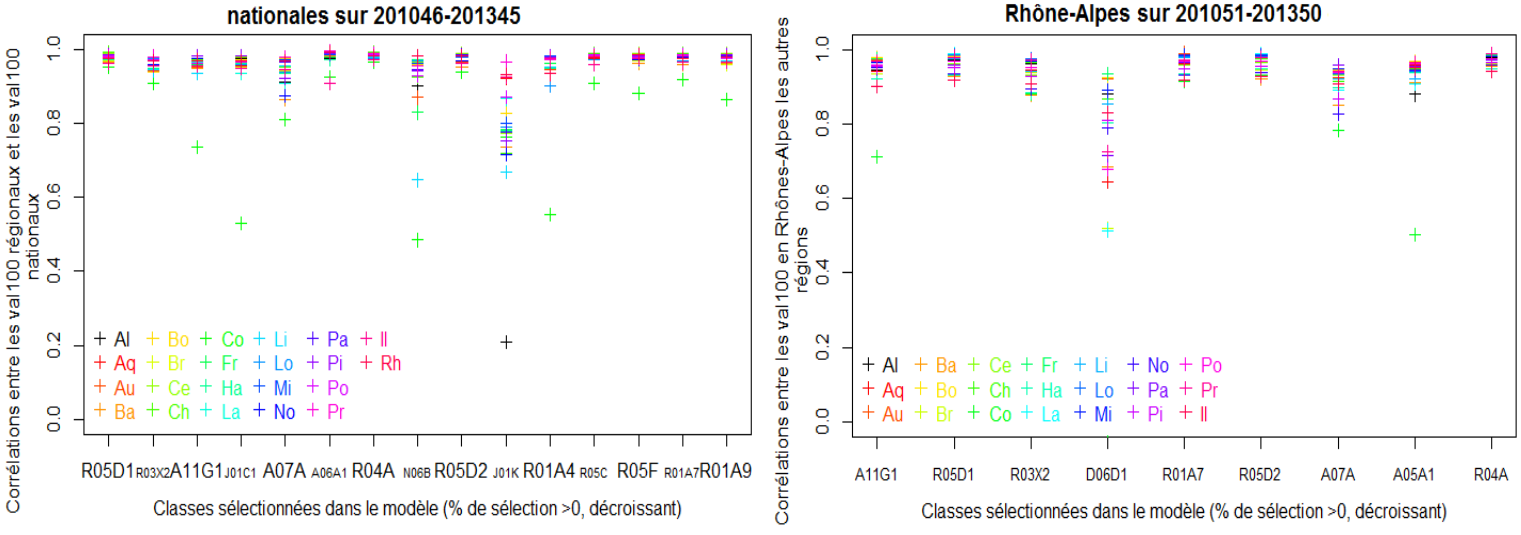
Les nouvelles incidences régionales obtenues par désagrégation spatiale (trait bleu plein) et par extrapolation spatiale (trait rouge plein) sont présentées en Annexe X, ainsi que ci-contre pour les régions PACA, Aquitaine, Basse-Normandie et Corse (Figure 15).

Un premier résultat est que, pour chaque région, les incidences prédites par désagrégation et extrapolation spatiales sont relativement similaires, avec un écart absolu moyen de 31.7 cas pour 100 000 (toutes régions et semaines confondues). Ainsi, bien que les incidences nationales servant à la désagrégation soient partiellement issues d'incidences régionales peu stables, il semble y avoir un relatif consensus entre ces deux nouvelles méthodes d'estimation des incidences régionales. Seule l'épidémie de 2010-2011 en Basse-Normandie fait exception, avec une prévision deux fois supérieure des incidences par l'extrapolation spatiale. Par ailleurs, on peut observer que la méthode d'extrapolation prédit souvent des incidences légèrement plus importantes que celles issues de la désagrégation.

Ces deux méthodes d'estimations apportent des résultats différents selon les régions. Ainsi avons-nous des régions où les nouvelles estimations sont très semblables à celles du RS (trait pointillé noir, avec IC à 95% en zones grisées) : les régions PACA, Auvergne, Centre, Franche-Comté, Midi-Pyrénées, Alsace, Haute-Normandie, Île-de-France et, pour la désagrégation spatiale, Rhône-Alpes. On peut également regrouper les régions où ces deux nouvelles méthodes semblent améliorer les estimations, au moins concernant leur stabilité : Bretagne, Languedoc-Roussillon, Limousin, Poitou-Charentes, Aquitaine, Champagne-Ardenne et Lorraine. Les estimations actuelles par les MS de ces régions avaient pour point commun de posséder de nombreuses fluctuations au cours du temps que l'on peut associer à des artefacts de mesures. Celles-ci étant bien moins présents avec les nouvelles estimations, il semble donc y avoir une amélioration. Aussi, un autre groupe peut être réalisé avec les régions dont les nouvelles estimations sont parfois plus élevées que celles du RS : Pays-de-la-Loire, Picardie, Basse-Normandie et Bourgogne. Enfin, deux régions ont des nouvelles incidences estimées très différentes de ce que l'on a actuellement à partir des MS : les régions Nord-Pas-de-Calais et Corse. Pour cette première, le faible nombre de MS et la faible confiance en les estimations actuelles, avec des pics au-delà de 1500 cas/100 000 habitants, font qu'il n'est pas inenvisageable de considérer ces nouvelles estimations. En revanche, la Corse faisant partie des régions de confiance, les estimations par ces nouvelles méthodes ne peuvent être prises en comptent pour cette région.



<u>Figure 16</u>. Pourcentage de fois que chaque classe soit présente dans les modèles nationaux (haut) ou dans les modèles en Rhône-Alpes (bas)



<u>Figure 17</u>. Corrélations entre les délivrances nationales (gauche) ou en Rhône-Alpes (droite) avec les délivrances dans les différentes régions, pour les classes présentes dans les modèles

Les régions associées aux '+' sont désignées par leurs 2 premières lettres (ex : 'Co' = Corse)

3.2. Validations des nouvelles estimations des incidences régionales

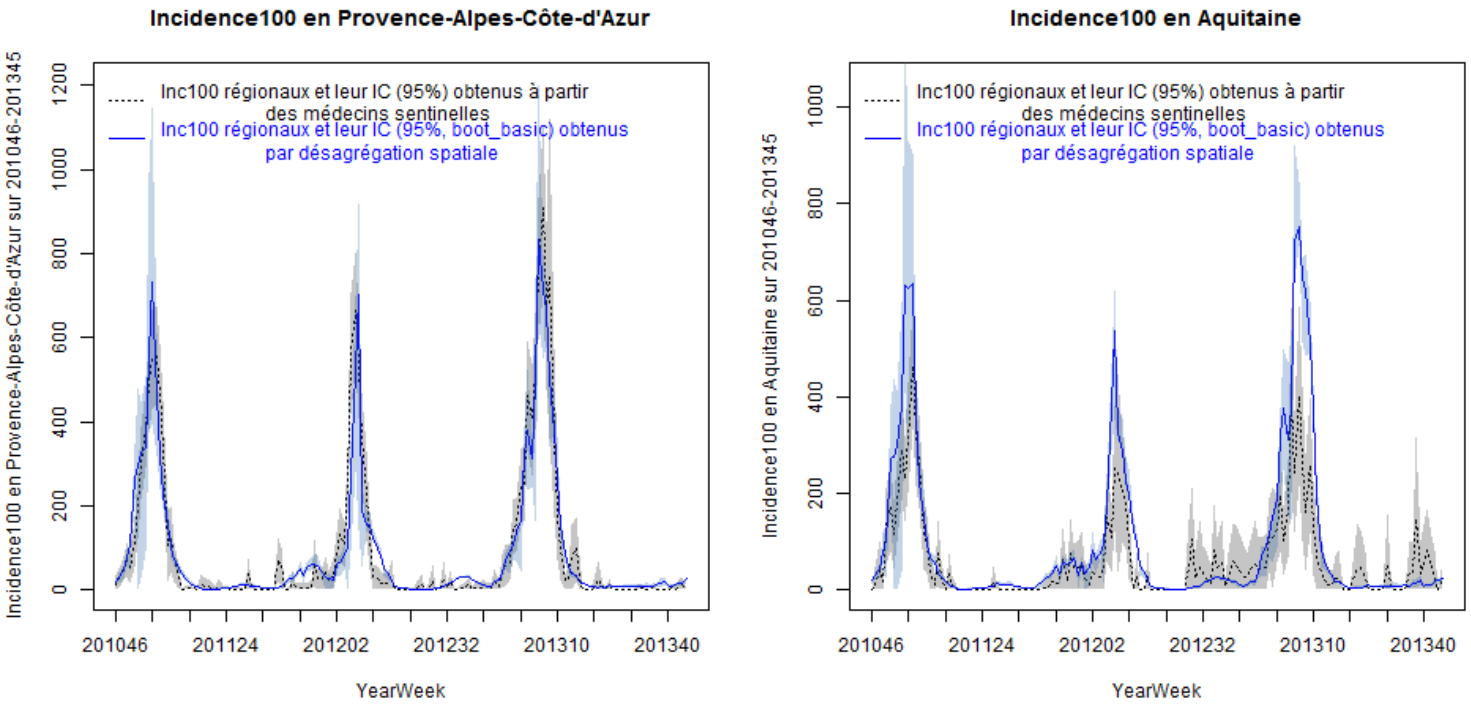
3.2.1. Vérifications des hypothèses inhérentes à la désagrégation et à l'extrapolation

La désagrégation et l'extrapolation spatiales reposent sur des hypothèses de similitudes concernant le comportement d'achat de médicaments face aux SG ((*H1*) et (*H1'*)) et le comportement temporel des SG ((*H2*) et (*H2'*)), sur l'ensemble du territoire français (voir I.2.3.).

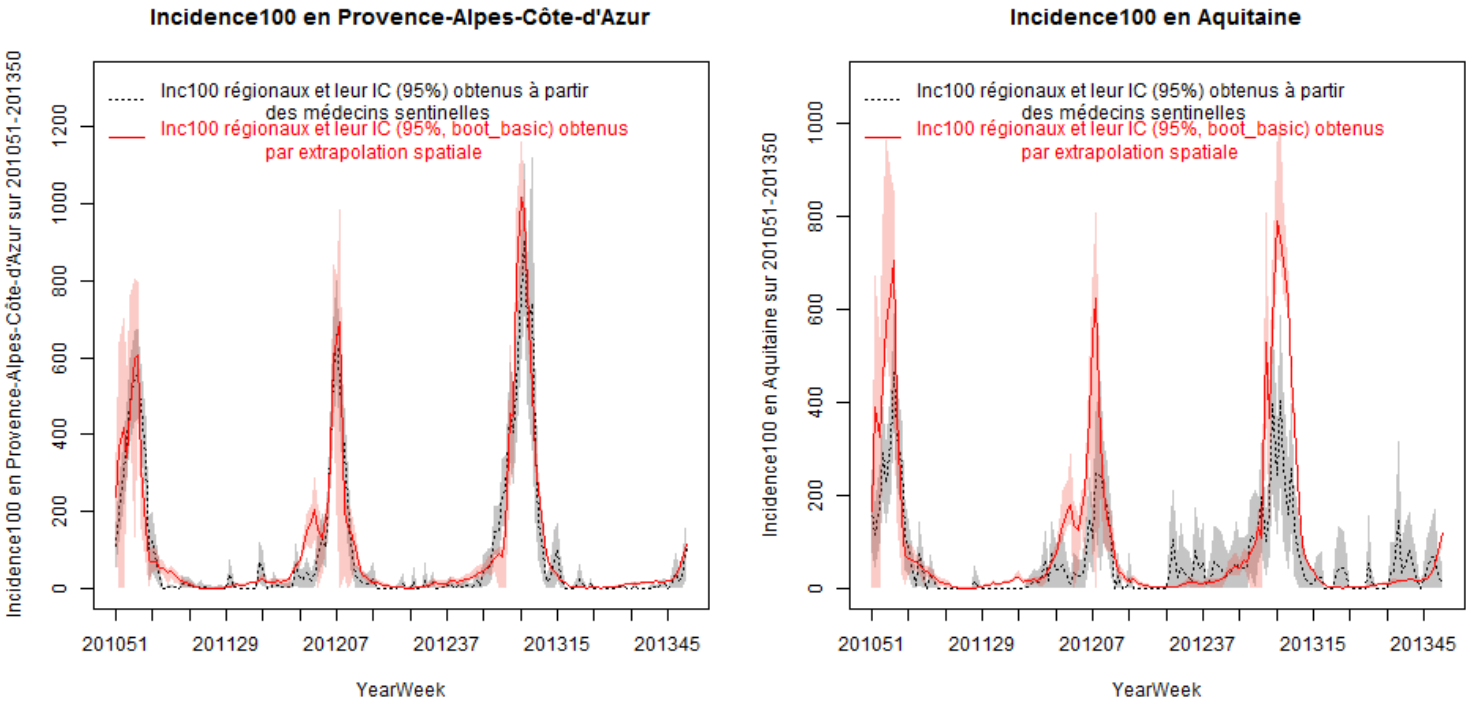
Une manière de valider les hypothèses relatives aux comportements d'achat, est de déterminer si les taux de délivrance nationaux des classes ayant servi à la désagrégation (respectivement les taux de délivrance en Rhône-Alpes des classes ayant servi à l'extrapolation) sont corrélés au cours du temps aux taux de délivrance régionaux de ces mêmes classes. Une bonne corrélation indiquerait alors des comportements d'achat similaires. Pour ce faire, il a fallu déterminer dans un premier temps les classes figurant dans les modèles nationaux (ou les modèles en Rhône-Alpes) servant ainsi à l'estimation des incidences régionales. On remarque sur la Figure 16 (haut) que la classe 'R05D1' est celle qui est la plus de fois présente dans les modèles nationaux prédisant la période 201046-201345 (74%), suivie par les classes 'R03X2', 'A11G1' et 'J01C1'. Les corrélations au cours du temps entre les taux de délivrance de ces classes au niveau national avec ceux au niveau régional sont indiquées en Figure 17 (gauche). On observe une très bonne corrélation entre les délivrances nationales et régionales, indiquant des comportements d'achat similaires. Toutefois, on observe que la région Corse (croix verte) se distingue du national pour de nombreuses classes de médicaments ('R03X2', 'A11G1', 'J01C1',…). Par conséquent, si l'hypothèse *(H1)* semble bien être vérifiée pour la quasi-totalité des régions, elle ne peut être validée pour la région Corse qu'il faut donc considérer séparément.

Ce travail a également été réalisé en région Rhône-Alpes pour vérifier l'hypothèse *(H1')* de l'extrapolation. Cette fois, ce sont les classes 'A11G1' et 'R05D1' qui sont les plus de fois présentes dans les modèles, avec 64% et 53% respectivement sur la période 201051-201350 (Figure 16, bas). On remarque sur la Figure 17 (droite) que la région Corse se distingue là également pour la classe 'A11G1', indiquant un comportement d'achat différent de celui de la région Rhône-Alpes. L'hypothèse *(H1')* semble donc être vérifiée pour l'ensemble des régions, excepté la Corse.

Nos données ne nous permettant pas de connaitre le comportement temporel des SG dans chaque région en France, la vérification des hypothèses *(H2)* et *(H2')* ne peut se faire qu'indirectement. Ainsi, la vérification de ces hypothèses s'appuie ici sur les régions de confiance (hors Corse) - en comparant les nouvelles estimations à celles du RS -, ainsi que sur l'ensemble des régions afin de déterminer la cohérence des pics épidémiques, uniquement en termes de démarrage et de fin des épidémies (et non des valeurs associées). On remarque dans un premier temps que les nouvelles estimations prédisent globalement les mêmes pics épidémiques que celles du RS (Annexe X). Aussi, pour les régions PACA, Rhône-Alpes, et dans une moindre mesure Auvergne, les nouvelles estimations sont très proches des incidences du RS. S'agissant de régions de confiance, on peut considérer de manière indirecte que les hypothèses *(H2)* et *(H2')* sont acceptables.



<u>Figure 18</u>. *Bootstrap* des taux d'incidence obtenus par désagrégation spatiale sur 201046-201345.
Gauche : en PACA. Droite : en Aquitaine



<u>Figure 19</u>. *Bootstrap* des taux d'incidence obtenus par extrapolation spatiale sur 201051-201350.
Gauche : en PACA. Droite : en Aquitaine

3.2.2. *Bootstrap* des nouvelles estimations des incidences régionales

Afin d'obtenir des intervalles de confiance des nouvelles prédictions, un *bootstrap* à l'aide du package *boot* (Canty & Ripley 2014) a été réalisé pour la désagrégation et l'extrapolation spatiales. Pour ce faire, pour chaque nouvelle estimation, 999 échantillons de taille 'f' optimale issues de tirages aléatoires avec remise ont été générées à partir des données nationales ou en Rhône-Alpes. De plus, le choix a été fait d'obtenir des intervalles de *bootstrap basic*, les hypothèses étant moins restrictives concernant la distribution normale (comparée au *bootstrap normal*), conférant ainsi de meilleures propriétés aux petits échantillons (Canty, A. J 2002).

Les résultats sont en Annexe XI, ainsi que ci-contre pour les régions PACA et Aquitaine (Figure 18 et Figure 19).

III. Discussion

Afin d'améliorer la stabilité des estimations des incidences régionales des SG, de nombreux choix méthodologiques ont été réalisés. La discussion abordera donc dans un premier temps la pertinence des choix entrepris, avant de discuter plus longuement des résultats obtenus.

1. Retour sur les choix méthodologiques entrepris

Pour répondre à la problématique, le choix a été fait d'utiliser les données médicamenteuses comme source d'information externe, puisqu'elles avaient déjà fait l'objet d'une étude probante en lien avec les SG (Vergu et al. 2006). Cependant, d'autres bases de données potentiellement informatives sont disponibles, à l'instar des données du système de surveillance syndromique « OSCOUR » (Organisation de la surveillance coordonnée des urgences), qui recueille en temps réel les informations médicales associées à près de 63% des passages aux urgences en France (source InVS). De même, on peut à nouveau citer la base de données des requêtes sur le moteur de recherche Google, présentée en introduction. Enfin, les données d'absentéisme scolaire semblent également prometteuses. Toutefois, la base de données médicamenteuses avait pour avantage d'être à la fois la plus représentative de la population générale – *a contrario* d'OSCOUR et des données d'absentéisme –, et robuste dans le temps, à la différence des données « google », où les requêtes sur le moteur de recherche sont nettement liées aux médias et aux crises potentielles (Butler 2013). Il serait en revanche envisageable d'améliorer ces recherches en modélisant les incidences par plusieurs sources de données à la fois.

Aussi, dans la démarche méthodologique, il a été fait le choix de réaliser une présélection de classes médicamenteuses en lien avec les SG. Cette présélection de classes est une étape très lourde de conséquences pour la suite, puisqu'elle est le socle sur lequel se basent tous les autres résultats. Ainsi, il est important de souligner que d'autres protocoles de sélection sont envisageables, tels que l'utilisation de forêts aléatoires. De même, prendre en compte dans la sélection de classes un critère assurant la bonne corrélation des délivrances nationales (ou en Rhône-Alpes) avec celles de l'ensemble des autres régions, aurait pu être plus rigoureux pour mieux satisfaire les hypothèses inhérentes à l'extrapolation et la désagrégation. Par ailleurs, il aurait pu être intéressant de réaliser la présélection des classes sur les résidus d'une régression modélisant les incidences par des fonctions temporelles uniquement. Les classes ainsi présélectionnées seraient informatives des SG sans la saisonnalité afin d'éviter la redondance avec les fonctions du temps dans les modèles. En outre, le choix de ne pas utiliser seulement les classes du panel d'experts, qui sont spécifiques à la grippe, peut être remis en question. Enfin, on peut noter que cette étape de présélection de classes peut être évitée par régression d'un modèle PLS-poisson sur l'ensemble des classes médicamenteuses disponibles. Les résultats figurant en Annexe XII indiquent cependant une moins bonne prévision des incidences, que ce soit avec l'approche des « modèles uniques », qu'avec des « modèles glissants ».

Concernant la recherche du meilleur modèle prédictif des incidences, certains choix peuvent être soumis à discussion. Premièrement, il a été fait le choix de ne pas réaliser de validations croisées, et cela pour un souci principalement technique. En effet, la validation croisée nécessite des temps de calculs plus long qui, sur l'ensemble de notre étude, auraient été bien trop importants.

Aussi, l'ensemble des comparaisons de modèles et d'approches ont été uniquement

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
200506	3.996758	936.0413	313.0447	2930.065	1188.8168	6.421599e-05
200401	3.936751	753.6934	261.5606	2328.538	953.4138	8.259224e-05
200505	3.896190	874.2476	300.7264	2758.713	1122.6552	9.771761e-05
200507	3.835777	855.0983	299.0918	2718.810	1105.5123	1.251679e-04
200652	3.784905	697.1231	238.1735	2287.767	924.7983	1.537671e-04
200504	3.782166	787.4990	281.0316	2499.169	1021.2876	1.554698e-04
200405	3.754328	690.5512	256.0932	2132.258	882.5770	1.738072e-04
200453	3.749498	735.1726	265.1367	2334.705	956.0816	1.771886e-04
200508	3.747233	778.3714	276.5779	2506.833	1021.2955	1.787959e-04
200552	3.736907	651.2241	230.9345	2107.231	857.7751	1.862977e-04
200901	3.671190	715.6401	237.3115	2477.363	993.7029	2.414234e-04
200801	3.648815	705.8333	240.5661	2415.616	972.4955	2.634523e-04
200701	3.641947	672.0116	235.1691	2262.637	914.8047	2.705841e-04
200752	3.624613	729.8623	249.7601	2509.069	1010.2040	2.893945e-04
200452	3.603727	774.3325	289.2102	2496.019	1026.6830	3.136861e-04
200605	3.588637	793.4214	287.4964	2641.853	1075.2109	3.324118e-04
200406	3.586294	699.4172	268.5002	2212.294	916.3995	3.354108e-04
200606	3.578654	837.9235	302.2262	2810.212	1141.6601	3.453687e-04

Figure 20. Extrait de la caractérisation des 50 classes présélectionnées sur la période longue, par la fonction 'catdes'

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
201306	1.964989	511.6668	260.9957	2484.579	1212.375	0.04941557

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
201306	5.795621	1931.0767	260.9957	4922.099	1212.3746	6.806878e-09
201307	5.753703	1876.6365	256.2529	4799.098	1184.8672	8.730936e-09
201301	5.743178	1349.5392	187.4230	3381.553	851.3273	9.291584e-09
201052	5.542031	1774.5940	249.4845	4232.868	1157.7942	2.989829e-08
201250	5.507564	1694.5696	247.4622	4369.769	1105.4536	3.638340e-08
201249	5.397226	1603.8625	237.9652	4210.900	1064.7477	6.767899e-08
201302	5.367517	1743.0246	259.8294	4683.612	1162.5836	7.982816e-08
201242	5.340573	1527.6959	230.1872	4055.082	1022.1667	9.265341e-08
201101	5.338892	1907.3471	278.9597	4754.519	1283.2341	9.351646e-08
201240	5.338594	1541.2623	231.2984	4095.333	1032.3614	9.367035e-08
201252	5.336033	1388.6336	197.0874	3405.912	939.4889	9.500211e-08
201305	5.335017	1899.3255	264.9135	4804.317	1288.9173	9.553581e-08
201105	5.326318	1722.9008	247.4153	4253.265	1165.4863	1.002238e-07
201051	5.299484	1732.4558	254.2542	4182.206	1173.5440	1.161305e-07
201241	5.283394	1545.4570	234.6346	4125.472	1043.8309	1.268124e-07
201104	5.250183	1659.9388	242.7533	4126.319	1135.6684	1.519484e-07
201152	5.233626	1435.6659	209.9272	3517.048	985.3589	1.662163e-07

Figure 21. Extraits de la caractérisation des 73 classes présélectionnées sur la période courte (en haut), et uniquement des 17 classes du panel d'experts (en bas), par la fonction 'catdes'

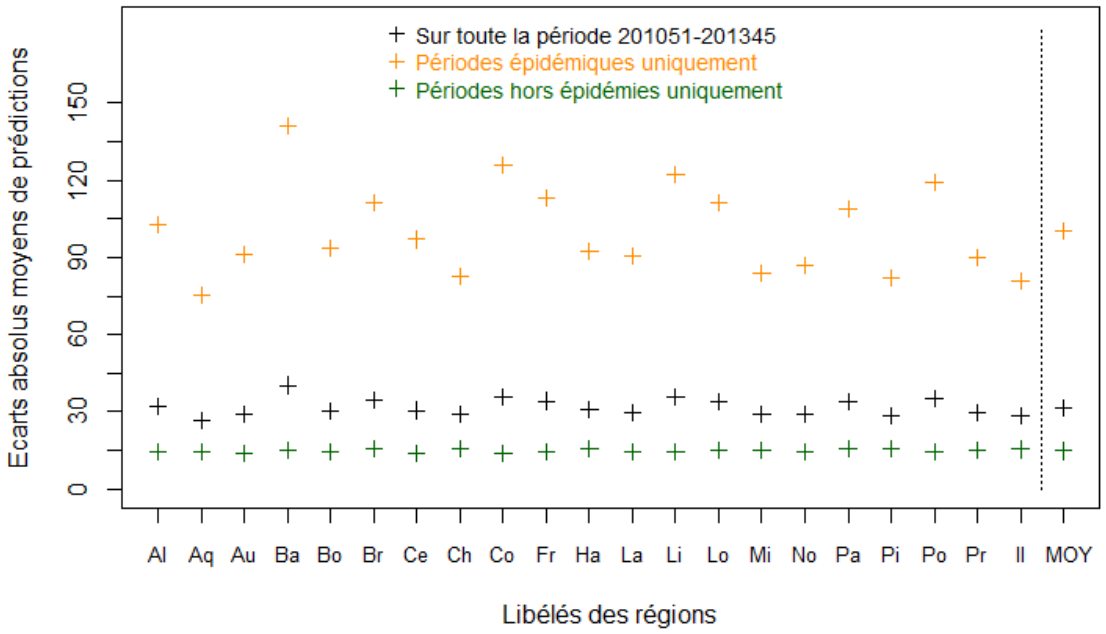
réalisées au niveau national pour la désagrégation spatiale. Il aurait pu être plus rigoureux d'effectuer également toutes ces comparaisons en Rhône-Alpes pour l'extrapolation, bien qu'il s'agisse d'un même type de données. Enfin, la règle d'inclusion des classes dans les modèles, qui est fonction des corrélations, est une contrainte forte et restrictive. En effet, ce critère empêche l'inclusion de classes moins corrélées avec les incidences mais pouvant être plus informatives, car non redondantes avec les autres classes déjà inclues.

Les principes de désagrégation spatiale et d'extrapolation spatiale consistent en la prédiction des incidences dans les différentes régions par l'application directe des modèles ajustés au niveau national ou en Rhône-Alpes, sans aucune transformation. La non-prise en compte d'une variabilité régionale, notamment en ce qui concerne le comportement d'achat de médicament face aux SG, est une hypothèse forte. En effet, il est probable qu'il existe des disparités de comportement, avec des régions où l'accessibilité à la santé est plus importante (selon l'offre de soins par exemple) ou encore des régions plus empreintes à l'achat de médicaments et à l'automédication (selon les philosophies de vie). Cela nécessiterait donc la prise en compte de co-variables au niveau de chaque région. Une autre approche serait de considérer un modèle mixte, où chaque coefficient $\beta_{i,r}$ des modèles régionaux serait constitué d'un effet fixe égal à $\beta_{i,nat}$ ou $\beta_{i,RH}$ (selon qu'il s'agisse de désagrégation ou d'extrapolation) et d'un effet aléatoire autour de $\beta_{i,r}$.

2. Retour sur les résultats

Afin de juger de la pertinence des classes présélectionnées sur la période longue (2004-2014) et sur la période courte (2010-2014), une caractérisation de ces classes a été réalisée. Celle-ci consiste, à partir de la forme du jeu de données décrite au I.2.1., en la détermination des variables (semaines) pour lesquelles les classes présélectionnées prennent des valeurs significativement différentes des classes qui ne le sont pas, par comparaisons de moyenne à l'aide de la fonction 'catdes' sous « R ». Sur la période longue, cette étude a permis de montrer que les classes présélectionnées sont caractéristiques des périodes hivernales (Figure 20), puisqu'elles ont des taux de délivrance significativement supérieurs aux autres classes pour les semaines hivernales. Elles semblent donc bien caractéristiques des épidémies de SG. En revanche, sur la période courte, la caractérisation de ces 73 classes n'a révélé qu'une seule semaine pour laquelle nos classes présélectionnées sont significativement plus délivrées que les autres : la 6ème semaine de l'année 2013 (Figure 21, haut), semaine épidémique. Toutefois, dans le détail, si nous caractérisons uniquement les 56 classes présélectionnées par les méthodes de classification, on remarque que celles-ci ne sont caractéristiques d'aucune semaine en particulière. En revanche, la caractérisation des 17 classes du panel d'experts montre bien que ces dernières sont très fortement caractéristiques des semaines épidémiques (Figure 21, bas). Ceci apporte donc certains doutes concernant la pertinence des classes présélectionnées. De même, que ce soit pour la période longue ou courte, on observe une grande diversité d'action et d'utilité de ces classes (Annexe II et Annexe III): en effet, si certaines classes sont des analgésiques antimigraineux (« N02C1 » et « N02C9 »), d'autres semblent plus éloignées des SG, à l'instar des dentifrices (« A01A1 »).

Concernant la recherche du meilleur modèle, il peut sembler curieux que l'approche des modèles glissants avec les paramètres (f,n) qui se ré-optimisent au cours du temps, ne soit pas mieux prédictive des incidences que celle où les paramètres sont fixés. Toutefois, une explication de cela réside dans ce qui est inhérent aux méthodes de prévision : le compromis biais/variance. En effet, on assiste à un meilleur ajustement des modèles glissants avec



<u>Figure 22</u>. Ecarts absolus moyens de prédictions des inc100 régionaux entre les méthodes de désagrégation et d'extrapolation spatiales, sur 201051-201345
MOY : Moyenne des écarts absolus moyens de prédictions pour l'ensemble des régions

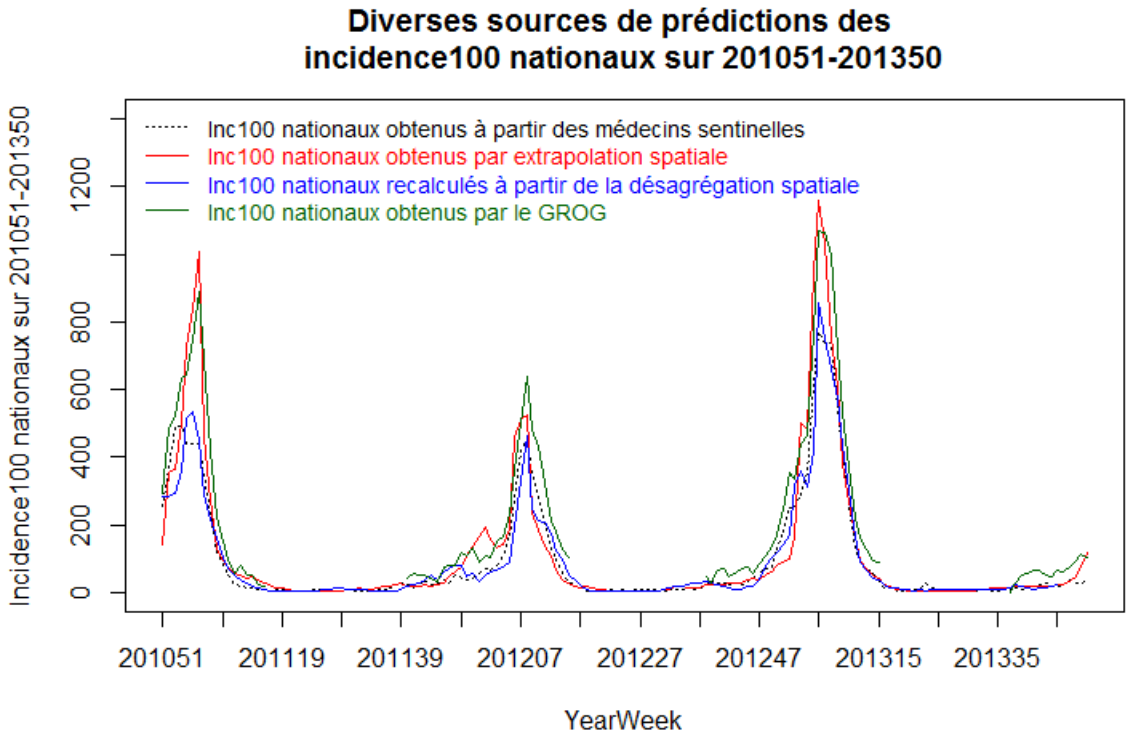
les paramètres non fixés (rmse d'ajustement moyen de 7,88 cas pour 100 000, sur 201139-201338, Annexe XIII) qu'avec ceux où les paramètres sont fixés (rmse moyen de 8,67 cas pour 100 000), ce qui, en contrepartie, diminue la capacité prédictive des modèles.

Par ailleurs, à partir du meilleur modèle sélectionné, la modélisation des incidences nationales ou en Rhône-Alpes conduit à des estimations très proches des taux d'incidence du RS (Figure 13 et Figure 14), avec une bonne prévision des « pics » et du démarrage des épidémies. On peut toutefois noter quelques anomalies de prévision, comme dans le modèle national au niveau des premières semaines de l'année 2011, mais surtout dans le modèle en Rhône-Alpes avec une surestimation des incidences juste avant l'épidémie de 2012, dont la cause pourrait être imputée à une capture de l'épidémie de VRS (Virus Respiratoire Syncytial) qui a eu lieu à la même période (source GROG).

Comme annoncé au II.3., les nouvelles estimations des incidences régionales semblent également plausibles (hors Corse), avec des prévisions relativement similaires entre désagrégation et extrapolation spatiales (écart absolu moyen de 31,7 cas pour 100000, toutes régions et semaines confondues, Figure 22). Toutefois, ce résultat masque de fortes disparités de prédiction entre les périodes épidémiques – où l'écart absolu moyen est de 100 cas pour 100000 (avec un écart-type moyen de 89) – et les périodes hors épidémies, où l'écart absolu moyen est de 14,8 cas pour 100000 (écart-type de 24). Ainsi, en période épidémique, bien qu'il y ait un consensus dans la forme des pics – en termes de démarrage et de fin des épidémies –, les valeurs estimées des taux d'incidence par les méthodes d'extrapolation et de désagrégation peuvent être très différentes. Néanmoins, les fortes valeurs des écart-types associées indiquent que ces différences sont dues à quelques écarts extrêmes. De même, si l'on analyse les écarts absolus moyens de prédiction dans chaque région, on observe un consensus plus prononcé entre les méthodes de désagrégation et d'extrapolation pour les régions Aquitaine, Île-de-France et Charente-Maritime ; tandis qu'il l'est moins pour les régions Basse-Normandie et Limousin. Enfin, la région Corse semble être un cas particulier, de par sa spécificité de consommation médicamenteuse. La Corse nécessite donc un modèle revu, ajusté par exemple par des classes spécifiquement choisies. Néanmoins, en tant que région de confiance, les incidences du RS en Corse sont actuellement déjà bien stables.

Concernant la « validation » des nouvelles estimations des incidences régionales, celle-ci s'apparente plus à des vérifications, le vrai nombre de cas de SG n'étant pas connu. Il est intéressant de noter que les délivrances médicamenteuses sont plus corrélées entre le national et les régions, qu'entre la région Rhône-Alpes et les autres régions (Figure 17). Ceci présage d'un comportement d'achat plus similaire entre le national et les régions, qu'entre Rhône-Alpes et les autres régions. Les hypothèses effectuées semblent donc être plus raisonnables pour la désagrégation que pour l'extrapolation. De même, l'analyse des modèles au II.2.3 indique une meilleure modélisation des taux d'incidence nationaux qu'en région Rhône-Alpes, pour laquelle de nombreuses semaines sont mal expliquées par le modèle. En revanche, le postulat de l'extrapolation - qui stipule la véracité des incidences en Rhône-Alpes - est probablement plus juste que celui de la désagrégation, les incidences nationales étant issues d'incidences régionales parfois peu stables.

Aussi, concernant les intervalles de confiance obtenus par *bootstrap*, il est difficile de mesurer le sens à donner à ces intervalles, les échantillons étant issus de tirages aléatoires au niveau national ou en Rhône-Alpes, tandis que les modèles prédisent au niveau régional. Par ailleurs, l'extrapolation spatiale n'a été réalisée qu'à partir de modèles basés sur la seule région Rhône-Alpes, empêchant ainsi de prendre une quelconque variabilité régionale. Un axe d'étude serait de réaliser l'extrapolation spatiale à partir de modèles agrégés de modèles



<u>Figure 23</u>. Estimations des inc100 nationaux sur 201051-201350, par le RS (trait pointillé noir), par extrapolation spatiale (trait rouge), par le GROG (trait vert), et par recalcul à partir des inc100 régionaux issues de la désagrégation spatiale (trait bleu)

régionaux de confiance, en s'appuyant notamment sur ce qui a été réalisé concernant les stratégies d'agrégation de modèles (Biau et al. 2013).

Enfin, il est intéressant de constater sur la Figure 23 que si l'on réalise l'extrapolation spatiale au niveau national (trait rouge plein), les nouvelles incidences nationales estimées sont assez éloignées de celles obtenues par le RS (trait noir pointillé), avec des incidences plus proches de celles du GROG (Groupe Régionaux d'Observation de la Grippe, trait vert plein), qui est un autre réseau de surveillance. Toutefois, les estimations des incidences par le GROG reposent sur une patientèle différente, et sur une définition historiquement moins spécifique. Enfin, si on recalcule les incidences nationales à partir des estimations régionales issues de la désagrégation (trait bleu), ces incidences nationales sont très proches de celles actuelles par les MS.

Conclusions

Les incidences des syndromes grippaux sont estimées à partir des déclarations d'un réseau basé en médecine générale. Le nombre de médecins participant dans une région donnée étant fluctuant au cours du temps, la stabilité des estimations peut être très variable selon les régions. Cette étude visait donc à proposer une nouvelle méthode d'estimation des incidences régionales, par l'utilisation de la base des délivrances médicamenteuses comme source de données externe. Pour cela, une présélection de classes en lien avec les syndromes grippaux a été réalisée par des méthodes de classification. Aussi, une recherche du meilleur modèle de prédiction des incidences nationales a été effectuée, en comparant différents types de modèles de régression périodique (log-linéaire, log-non linéaire et PLS-Poisson), ainsi que différentes approches (modèles glissants ou non, à paramètres fixés ou non). Cette recherche a permis de montrer que les modèles glissants log-linéaires à paramètres fixés étaient les plus adaptées à notre étude, avec une prédiction des incidences nationales (pour la désagrégation) et en Rhône-Alpes (pour l'extrapolation) très proche de celles des données de validation. Finalement, la « désagrégation » et « l'extrapolation » spatiales ont permis de proposer de nouvelles estimations des incidences régionales (hors Corse, cas particulier), de meilleures stabilités. Un intervalle de confiance pour chaque nouvelle estimation a alors été déterminé par *bootstrap*.

Bibliographie

Bertrand, F., Meyer, N. & Maumy-Bertrand, M., 2014. Partial least squares Regression for generalized linear models, R package version 1.1.0. Available at: http://cran.r-project.org/web/packages/plsRglm/index.html.

Biau, G. et al., 2013. COBRA: A Nonlinear Aggregation Strategy. *arXiv preprint arXiv:1303.2236*.

Butler, D., 2013. When Google got flu wrong. *Nature*, 494(7436), p.155.

Canty, A. & Ripley, B., 2014. boot: Bootstrap R (S-Plus) Functions. R package version 1.3-11. Available at: http://cran.r-project.org/web/packages/boot/index.html.

Canty, A. J, 2002. Resampling Methods in R: The boot Package. *R News*, p.2–7.

Costagliola, D. 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), p.97‑99.

Husson, F., Josse, J. & Pages, J., 2010. Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data. *Applied Mathematics Department*.

Mehrotra, R. & Singh, R.D., 1998. Spatial disaggregation of rainfall data. *Hydrological Sciences Journal-Journal des Sciences Hydrologiques*, 43(1), p.91‑102.

Pelat, C. et al., 2010. A method for selecting and monitoring medication sales for surveillance of gastroenteritis. *Pharmacoepidemiology and drug safety*, 19(10), p.1009‑1018.

Pelat, C. et al., 2009. More diseases tracked by using Google Trends. *Emerging infectious diseases*, 15(8), p.1327.

Sax, C. & Steiner, P., 2013. Temporal Disaggregation of Time Series.

Scarpino, S.V., Dimitrov, N.B. & Meyers, L.A., 2012. Optimizing provider recruitment for influenza surveillance networks. *PLoS computational biology*, 8(4), p.e1002472.

Serfling, R.E., 1963. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public health reports*, 78(6), p.494.

source GROG, Groupes Régionaux d'Observation de la Grippe. Available at: http://www.grog.org/# [Consulté le juillet 31, 2014].

source IMS Health, Home - IMS Health. Available at: http://www.imshealth.com/portal/site/imshealth [Consulté le juillet 31, 2014].

source IMS Pharmastat, ::IMS Pharmastat.fr:: Available at: http://www.ims-pharmastat.fr/quisommes-nous/Decouvrir-Pharmastat [Consulté le juillet 31, 2014].

source InVS, Available at: http://www.invs.sante.fr/fr../Dossiers-thematiques [Consulté le juillet 31, 2014].

source RS, Réseau Sentinelles. Available at: http://websenti.u707.jussieu.fr/sentiweb/?site=fr [Consulté le juillet 31, 2014].

Souty, C. et al., 2014. Improving disease incidence estimates in primary care surveillance systems. *Population Health Metrics*, 12(1), p.19.

Thacker, S.B., Parrish, R.G. & Trowbridge, F.L., 1988. A method for evaluating systems of epidemiological surveillance. *World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales*, 41(1), p.11–18.

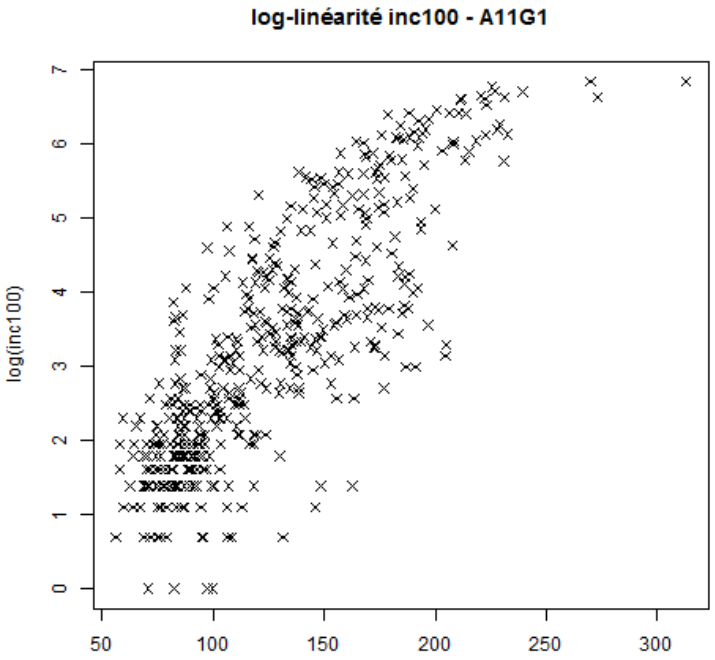
Valleron, A.-J., 2012. *L'épidémiologie humaine: Conditions de son développement en France, et rôle des mathématiques*, EDP sciences.

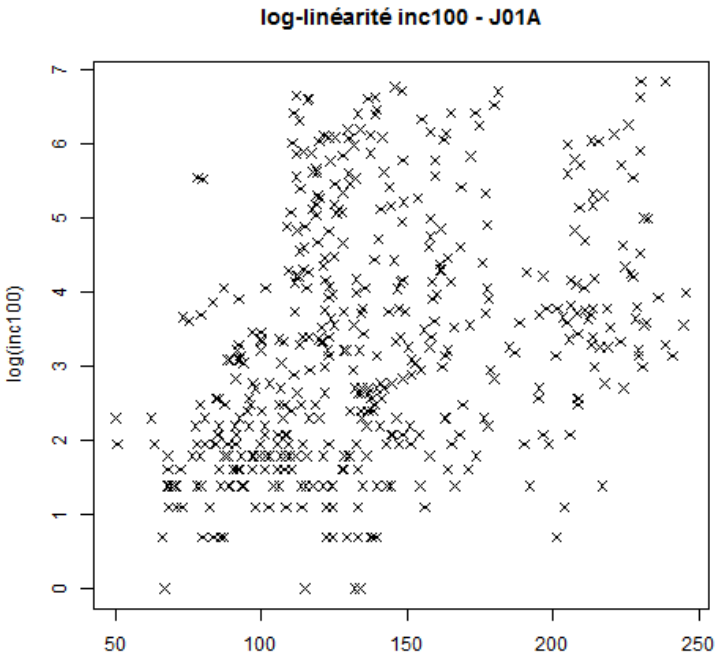
Vergu, E. et al., 2006. Medication sales and syndromic surveillance, France. *Emerging infectious diseases*, 12(3), p.416.

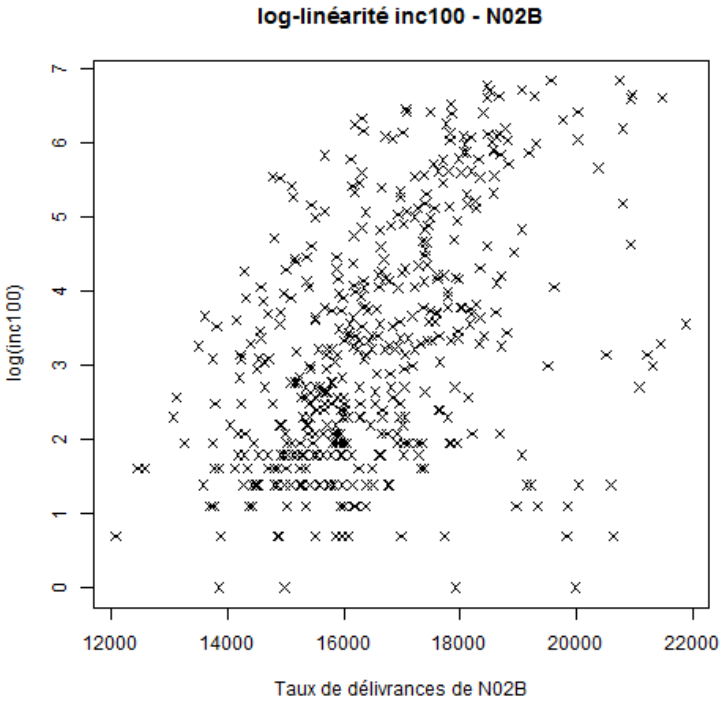
Annexes

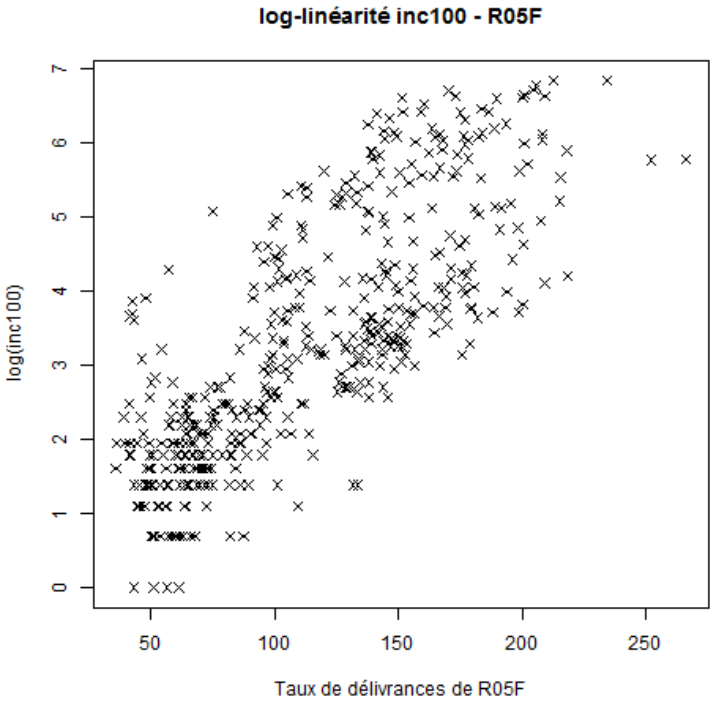
Annexe I. Tracés de la log-linéarité entre les taux d'incidence des SG et les taux de délivrance des classes.

Exemples de 4 classes du panel d'experts : A11G1, J01A, N02B et R05F.









Annexe II. Descriptions des classes présélectionnées sur 2004-2014 et valeurs propres de la CAH post-ACP (sortie "R").

Les 17 classes du panel d'experts (WHOCC)

Classes	Descriptions générales	Descriptions plus détaillées
A11G1	VIT.C +ASSOC.AV MINERAUX	VIT.C SEULE
J01A	ANTIBACTERIENS SYSTEMIQ.	TETRACYCLINES ET ASSOCIAT
R05F	ANTITUSSIFS PREP BRONCH	AUTRES ANTITUS+P BR PULM
J01C1	PENICILLINE LARGE SPECTRE	PENICIL.LARG.SPECT.US.OR.
J01D1	CEPHALOSPORINES	CEPHALOSPOR.USAGE ORAL
J01F	ANTIBACTERIENS SYSTEMIQ.	MACROLIDES ET APPARENTES
J01X9	AUTRES ANTIBACTERIENS	TS AUTRES ANTIBACTERIENS
N02B	ANALGESIQUES	ANALGES NON NARC ANTIPYR
R01A1	PREP RHINOLOGIQUES LOC.	CORTIC.RHINOL.SS ANTIINF.
R01A4	PREP RHINOLOGIQUES LOC.	ANTIINF.RHINOL SS CORTIC.
R01A7	PREP RHINOLOGIQUES LOC.	DECONGESTIONNANTS NASAUX
R01A9	PREP RHINOLOGIQUES LOC.	AUT.PREP.RHINOL.TOPIQUES
R01B	ANTIINF DECONGEST RHINO	PREP RHINOLOGIQUES V.GEN.
R02A	ANTIINF DECONGEST PHARYNX	ANTIINF DECONGEST PHARYNX
R05C	ANTITUSSIFS PREP BRONCH	EXPECTORANTS
R05D1	SEDATIFS DE LA TOUX	ANTITUSSIFS SEULS
R05D2	SEDATIFS DE LA TOUX	AUTR.ANTITUSSIFS + ASSOC.

Classes	Descriptions générales	Descriptions plus détaillées
A01A1	STOMATOLOGIE	DENTIFRICES
A02B1	ANTIULCEREUX	ANTAGONIST.RECEPTEURS H2
A05B	CHOLAGOGUES.HEPATOPROTEC	HEPATOPROTEC.LIPOTROPES
A07A	A-DIAR.AP.ELECT.A-INF INT	ANTIINFECT INTESTINAUX
A07X	A-DIAR.AP.ELECT.A-INF INT	AUTRES ANTIDIARRHEIQUES
C01A1	GLUCOSIDES CARDIAQUES	GLUCOSIDES SEULS
C03A1	DIURETIQUES	EPARGNEURS POTASSIUM SEUL
C03A3	DIURETIQUES	THIAZIDES +APPARENT.SEULS
C05B	ANTIVARIQUEUX/ANTIHEMORR.	ANTIVARIQUEUX TOPIQUE
C06A	AUTRES CARDIOVASCULAIRES	AUTRES CARDIOVASCULAIRES
C07B1	BETA BLOQUANTS EN ASSOC.	ASS.HYPOTENS ET/OU DIURET
D06D1	PROD.ANTIVIRAUX LOCAUX	ANTIVIRAUX TOPIQUES
D10B	PRDT ANTI ACNE	PRDT ANTI ACNE V. ORALE
G02F	AUTRES PRODUITS GYNECOLOG	HORMON.SEXUELLES V.TOP
G02X9	AUT.PRDTS GYNECOLOGIQUES	AUT.PROD.GYNECOLOGIQUES
J01D2	CEPHALOSPORINES	CEPHALOSPOR.INJECTABLE
J01H1	PENICIL:SPECTR.MOY.ETROIT	PENI.SPECT.MOY+ETROIT SL.
J01K	ANTIBACTERIENS SYSTEMIQ.	AMINOGLYCOSIDES
N02C1	ANALGES ANTIMIGRAINEUX	TRIPTANS ANTIMIGRAINEUX
N02C9	ANALGES ANTIMIGRAINEUX	AUTRES ANTIMIGRAINEUX
N05B2	HYPNOTIQUES ET SEDATIFS	HYPNOT ASS DE NON BARBIT
N06A3	ANTIDEPRESS.& REG.HUMEUR	REGULATEURS DE L'HUMEUR
N07B	A.PROD ACT SUR SNC	PRODUITS ANTI TABAC
P01B	ANTIPROTOZ.&ANTHELMINTIQ	ANTHELMINT. SF SCHISTOS.
R03G3	ASSOC.ANTICHOL/B2 STIMUL.	ANTICHOLINERG INH SEULS
R03J2	ANTIASTH.ANTILEUKOTRIENE	ANTIASTH.ANTILEUKO.SYST.
R03X2	AUT.A-ASTHM & COPD. PRDTS	AUT.A-ASTHM & COPD. SYS
R04A	REVULSIFS PERCUTAN.INHAL	REVULSIFS ET PRDT INHAL
S02C	PRDT OTOLOGIQUES	PRDT OTOLOG CORT +ANTIINF
V06C	DIETETIQUE GENERALE	DIETETIQUE DIVERS
A06A3	LAXATIFS	LAXATIFS AUGM.BOL FOECAL
N05A1	ANTIPSYCHOTIQUES	ANTIPSYCHOTIQUES ATYPIQ.
S01A	PRDT OPHTALMOLOGIQUES	PRDT OPHT ANTIINFECTIEUX

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	514.14	98.68	98.68
comp 2	3.52	0.68	99.36
comp 3	1.43	0.28	99.63
comp 4	0.67	0.13	99.76
comp 5	0.58	0.11	99.87
comp 6	0.17	0.03	99.91
comp 7	0.12	0.02	99.93
comp 8	0.09	0.02	99.95

Annexe III. Descriptions des classes présélectionnées sur 2010-2014 (hors panel d'experts) et valeurs propres de la CAH post-ACP (sortie "R").

Classes	Descriptions générales	Descriptions plus détaillées
A01A1	STOMATOLOGIE	DENTIFRICES
A01A3	STOMATOLOGIE	ANTIINFLAM&ANALG.PURS BUC
A02A4	A-ACID.A-FLATUL.CARMINAT.	A-ACID.+A-FLAT.OU CARMIN.
A02A7	A-ACID.A-FLATUL.CARMINAT.	A-FLAT.ET/OU CARM+A-PRDTS
A02B1	ANTIULCEREUX	ANTAGONIST.RECEPTEURS H2
A05A1	CHOLAGOGUES.CHOLERETIQUE	CHOLERETIQ.CHOLECYSTOKIN.
A05A2	CHOLAGOGUES.CHOLERETIQUE	ANTILITHIASIQUES
A06A1	LAXATIFS	LAXATIFS EMOLLIENTS
A07A	A-DIAR.AP.ELECT.A-INF INT	ANTIINFECT INTESTINAUX
A10C1	INSULINES HUMAINES+ANAL.	INSUL.HUM+ANAL.ACT.RAPID
A10M1	ANTIDIABET.ORAUX	ANTIDIABET.SULFONYLURES
A10N3	ANTIDIABET.ORAUX	ANTIDIABET.SULFONYLURES
B03A2	ANTIANEMIQ.FER ET ASSOC.	ANTIANEMIQ.FER EN ASSOC.
C01A1	GLUCOSIDES CARDIAQUES	GLUCOSIDES SEULS
C05B	ANTIVARIQUEUX/ANTIHEMORR.	ANTIVARIQUEUX TOPIQUE
C06A	AUTRES CARDIOVASCULAIRES	AUTRES CARDIOVASCULAIRES
C09B3	IEC EN ASSOCIATION	IEC ASS ANTAG CALCIQUES
C09D3	ANTAG ANGIOTENS II ASSOC	ARA 2 ASS ANTAG CALCIQUE
C10A9	PR.REGUL.CHOLEST&TRIGLYC	TS A.REG.CHOLEST&TRIGLYC
C10B	PR.LIPID-REGUL.A-ATHEROM	ANTI-ATHEROM. ORIG.NATUR
C10C	PR.LIPID-REGUL.A-ATHEROM	REG LIP ASS AV A.REG LIP
D10B	PRDT ANTI ACNE	PRDT ANTI ACNE V. ORALE
G01A1	TRICHOMONACIDES	TRICHOMONACIDES V.GENER
G02X9	AUT.PRDTS GYNECOLOGIQUES	AUT.PROD.GYNECOLOGIQUES
G03A5	CONTRACEPTIFS HORM SYST.	PROGESTATIFS ORAUX
G03A9	CONTRACEPTIFS HORM SYST.	AUT.CONTRACEPT.HORM.SYST.
G04A1	ANTISEPT.ANTIINFECT URIN.	ANTIBIOTIQ.+SULFAMID.URIN
G04E	UROLOGIE	PR DYSFONCTION ERECTILE
J01E	ANTIBACTERIENS SYSTEMIQ.	ASS AV TRIMETHOP.APPARENT
J01H1	PENICIL:SPECTR.MOY.ETROIT	PENI.SPECT.MOY+ETROIT SL.
J01K	ANTIBACTERIENS SYSTEMIQ.	AMINOGLYCOSIDES
J07A7	VACCINS SEULS	VACCINS PNEUMOCOCCIQUES
J07B2	VACCINS EN ASSOCIATION	ASS AV ROUG ET/OU OREILL
K01B3	SOLUTION STANDARD	SOL.D'HYD.DE CARBONE<=10%
K04B2	SOL STANDARD <100ML	SOL STAND >20ML ET <100ML
M01A3	ANTIRHUMAT NON STEROIDIEN	COXIBS SEULS
N02C9	ANALGES ANTIMIGRAINEUX	AUTRES ANTIMIGRAINEUX
N05B2	HYPNOTIQUES ET SEDATIFS	HYPNOT ASS DE NON BARBIT
N06A3	ANTIDEPRESS.& REG.HUMEUR	REGULATEURS DE L'HUMEUR
N06B	PSYCHO ANALEPTIQUES	PSYCHOSTIMULANTS -AMPHET
N07D1	PRODUITS ANTI-ALZHEIMER	PROD.ANTI-ALZ.INHIB.CHOL
R03A3	STIMULANTS RECEPTEURS B2	B2-STIMUL. INH. A.LONG.
R03X2	AUT.A-ASTHM & COPD. PRDTS	AUT.A-ASTHM & COPD. SYS
S01F	PRDT OPHTALMOLOGIQUES	MYDRIATIQUES ET CYCLOPLEG
S01G1	ANTIS. DECONG. A/ALL OPH.	A/ALLERG. A/HISTAMIN OPH
S01M	PRDT OPHTALMOLOGIQUES	TONIQ. VITAMINES VIS.OPH
S01R	PRDT OPHTALMOLOGIQUES	ANTIINFL NON STEROID.OPHT
T01A	DIAGNOSTIC PAR IMAGERIE	URO-ANGIOGRAP.FAIB.OSMOL.
T02C	TESTS DIAGNOSTICS	AUTRES TESTS DIAGNOSTICS
V03X	MEDICAMENTS DIVERS	TS AUT PROD THERAPEUTIQ.
V06B	DIETETIQUE GENERALE	DIETETIQUE DIVERS
R04A	REVULSIFS PERCUTAN.INHAL	REVULSIFS ET PRDT INHAL
S02C	PRDT OTOLOGIQUES	PRDT OTOLOG CORT +ANTIINF
C07B1	BETA BLOQUANTS EN ASSOC.	ASS.HYPOTENS ET/OU DIURET
D04A	ANTIPRURIGINEUX	ANTIPRURIGINEUX
D06D1	PROD.ANTIVIRAUX LOCAUX	ANTIVIRAUX TOPIQUES

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	205.66	98.87	98.87
comp 2	1.58	0.76	99.63
comp 3	0.34	0.16	99.80
comp 4	0.21	0.10	99.90
comp 5	0.10	0.05	99.95
comp 6	0.04	0.02	99.97
comp 7	0.02	0.01	99.98
comp 8	0.01	0.00	99.98

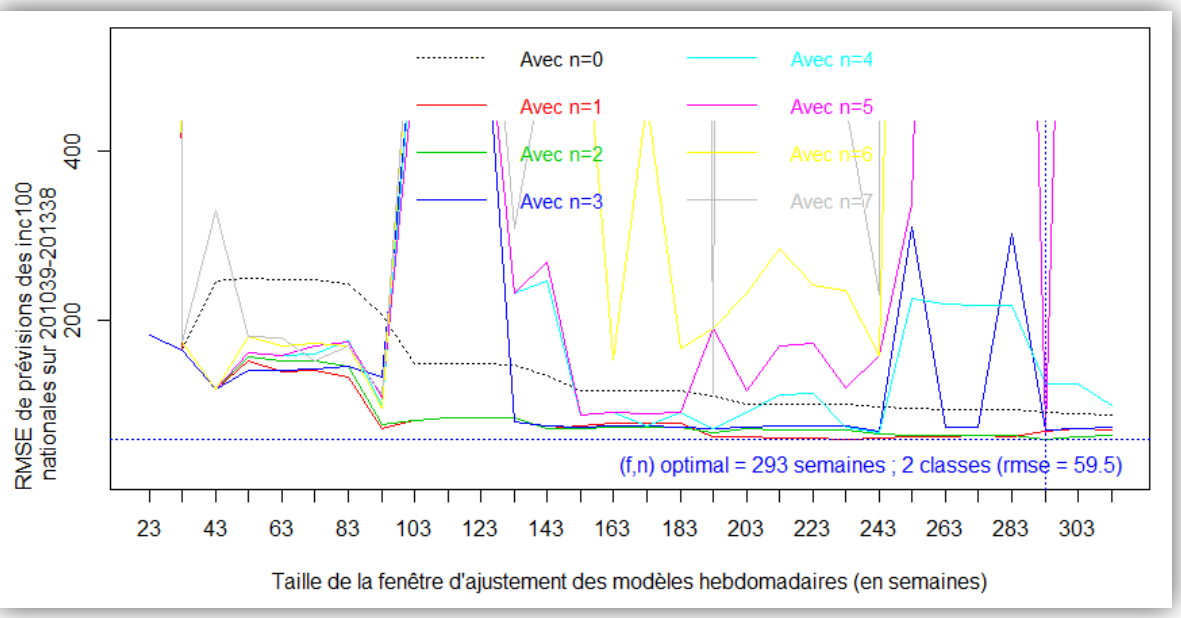
Annexe IV. Les paramètres optimaux minimisant les erreurs de prédictions des inc100 nationaux sur 201039-201338, pour l'approche du type "modèle unique"

Modèle log-linéaire

Paramètres optimaux :

'f' = 293 semaines
'n' = 2 classes

RMSEP = 59,5

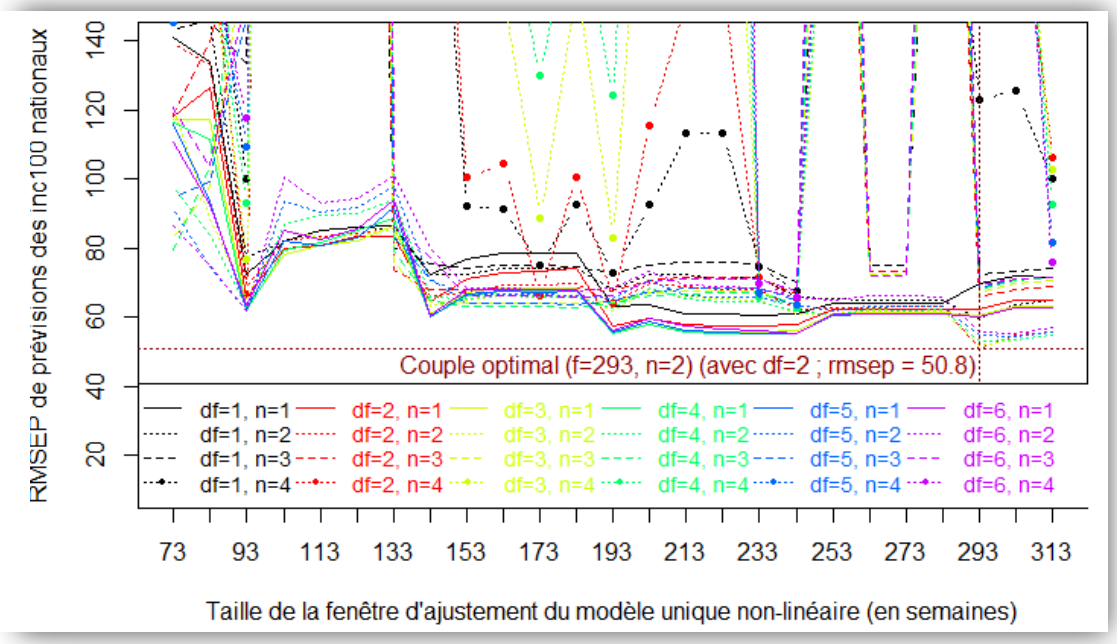


Modèle log-non linéaire

Paramètres optimaux :

'f' = 293 semaines
'n' = 2 classes
'df' = 2

RMSEP = 50,8

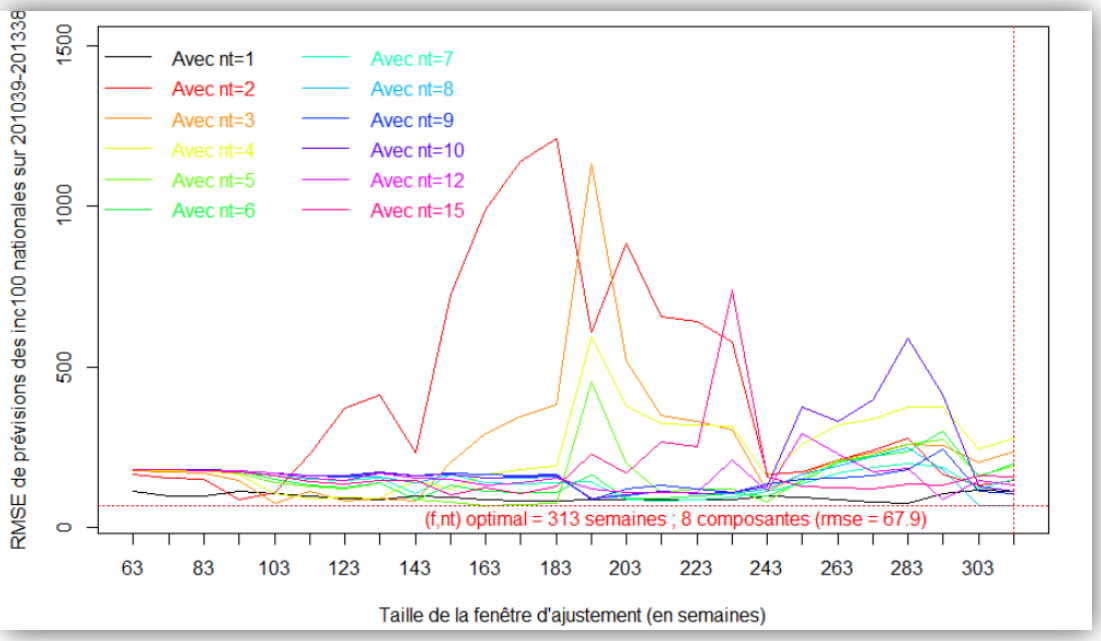


Modèle PLS-Poisson

Paramètres optimaux :

'f' = 313 semaines
'nt' = 8 valeurs latentes

RMSEP = 67,9



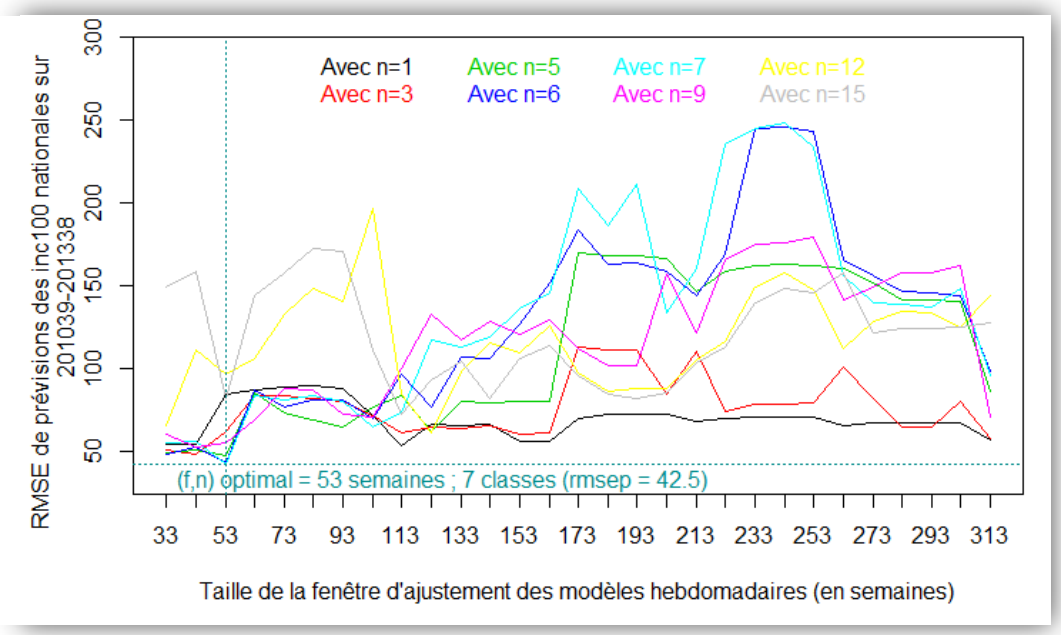
Annexe V. Les paramètres optimaux minimisant les erreurs de prédictions des inc100 nationaux sur 201039-201338, pour l'approche du type "modèles glissants à paramètres fixés optimisés"

Modèle log-linéaire

Paramètres optimaux :

'f' = 53 semaines
'n' = 7 classes

RMSEP = 42,5

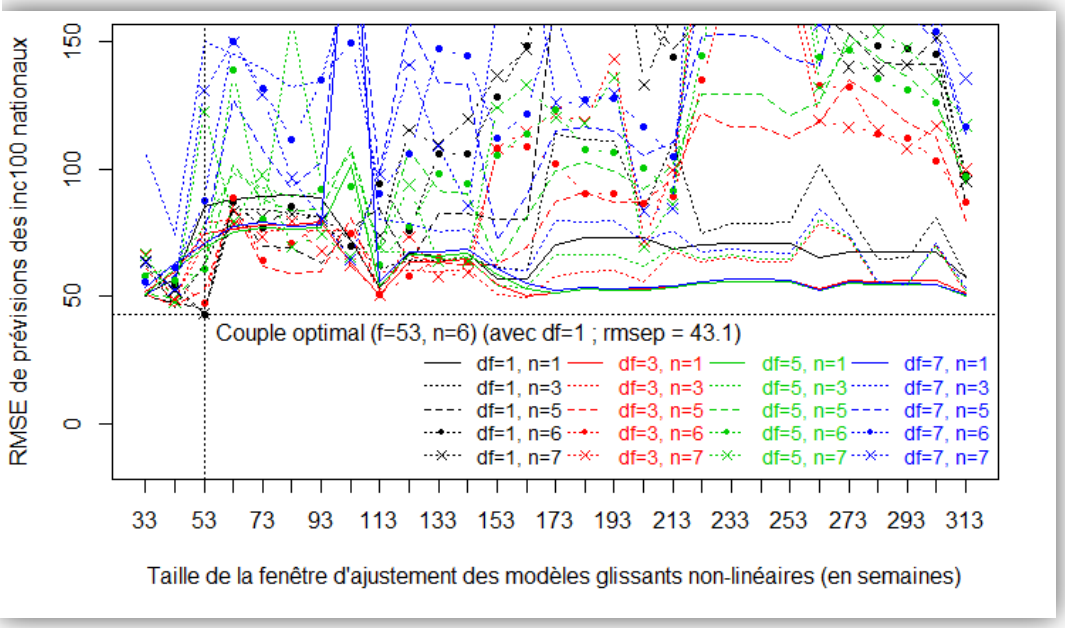


Modèle log-non linéaire

Paramètres optimaux :

'f' = 53 semaines
'n' = 6 classes
'df' = 1

RMSEP = 43,1

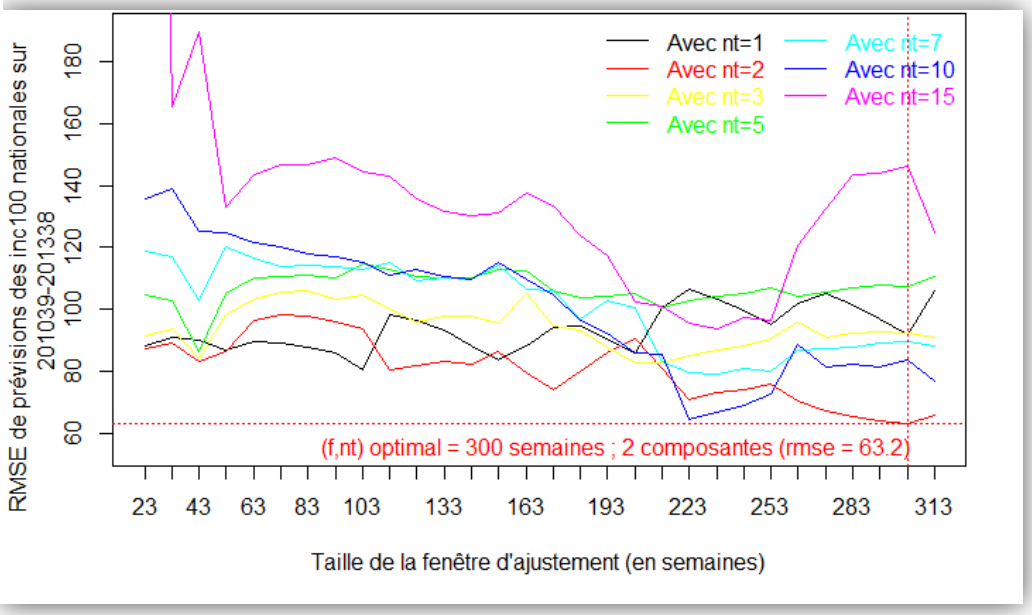


Modèle PLS-Poisson

Paramètres optimaux :

'f' = 300 semaines
'nt' = 2 valeurs latentes

RMSEP = 63,2

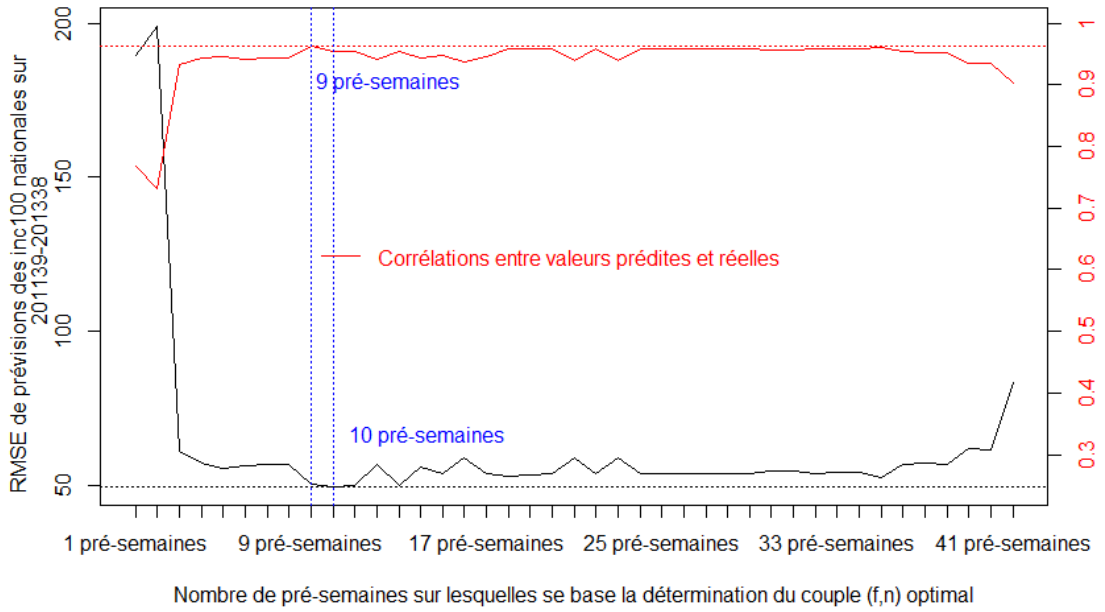


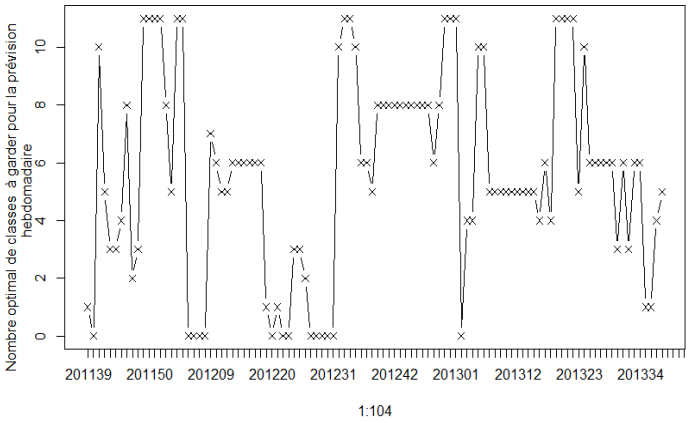
Annexe VI. Comparaisons des erreurs de prédictions par les approches du type « modèles glissants », entre ceux "à paramètres fixés optimisés" et ceux "à paramètres non fixés optimisés ", sur 201139-201338

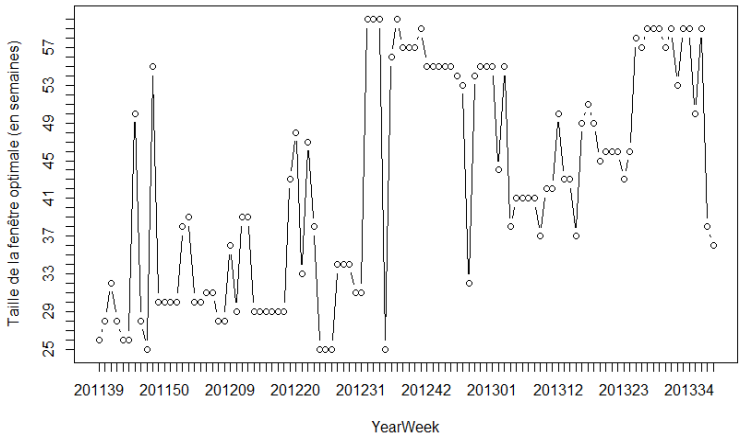
Nombre de pré-semaines ‘j’ sur lequel doit se baser la ré-optimisation hebdomadaire des paramètres ‘f’ et ‘n’ pour la prédiction des inc100 nationaux sur 201139-201338

‘j’ optimal
=
10 pré-semaines

RMSEP = 49,8







Les valeurs optimales prises par les paramètres ‘n’ (à gauche) et ‘f’ (à droite) pour chaque prédiction (de 201139 à 201338), du fait de la ré-optimisation hebdomadaire basée sur les ‘j’ = 10 pré-semaines

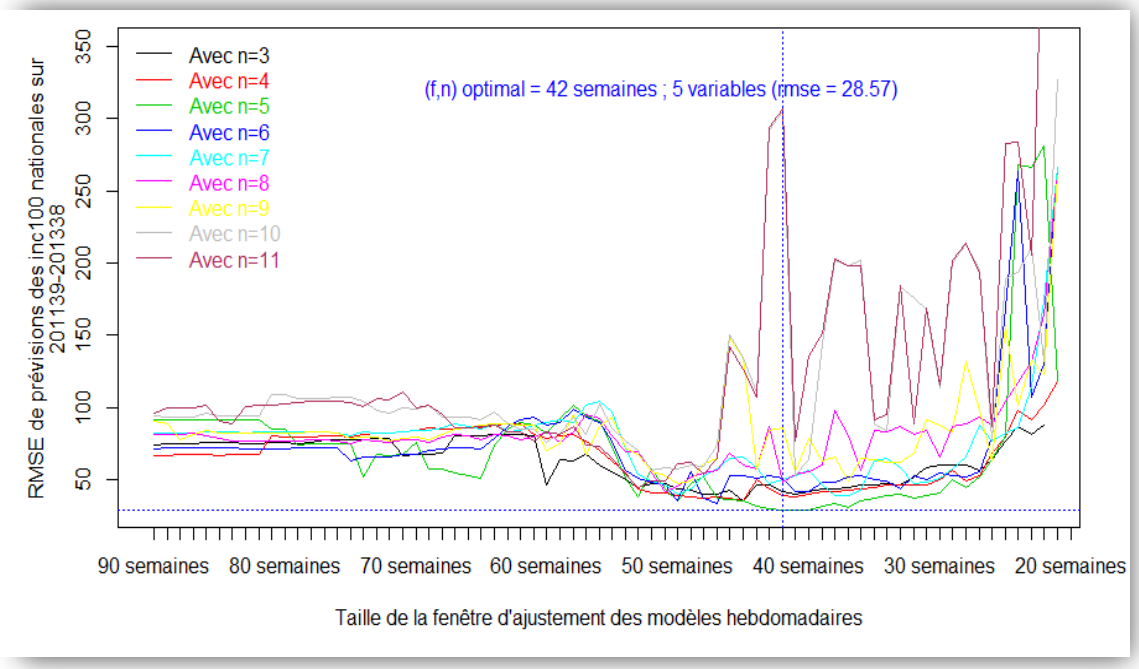
---

Modèles glissants log-linéaires à paramètres fixés optimisés sur 201139-201338

Paramètres optimaux :

‘f’ = 42 semaines
‘n’ = 5 classes

RMSEP = 28,57



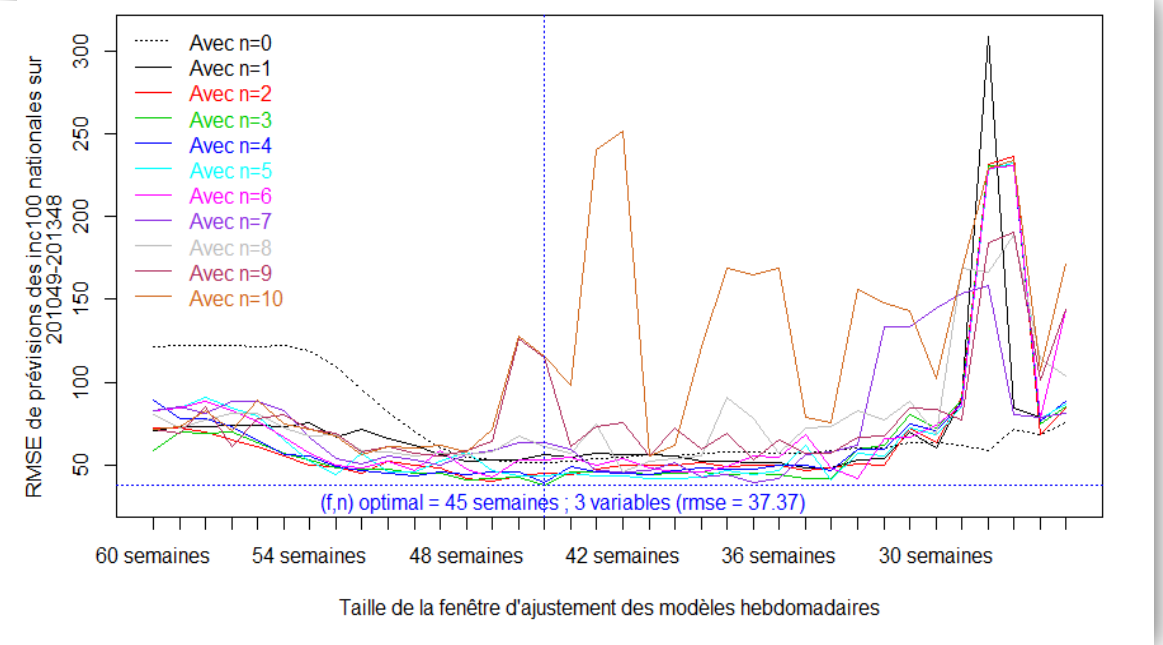
Annexe VII. Affinements au niveau national et en Rhône-Alpes du meilleur modèle : Obtention des paramètres 'f' et 'n' minimisant les erreurs de prédictions des inc100

Affinement au niveau national sur 201049-201348 :

Paramètres optimaux :

'f' = 45 semaines
'n' = 3 classes

RMSEP = 37,37

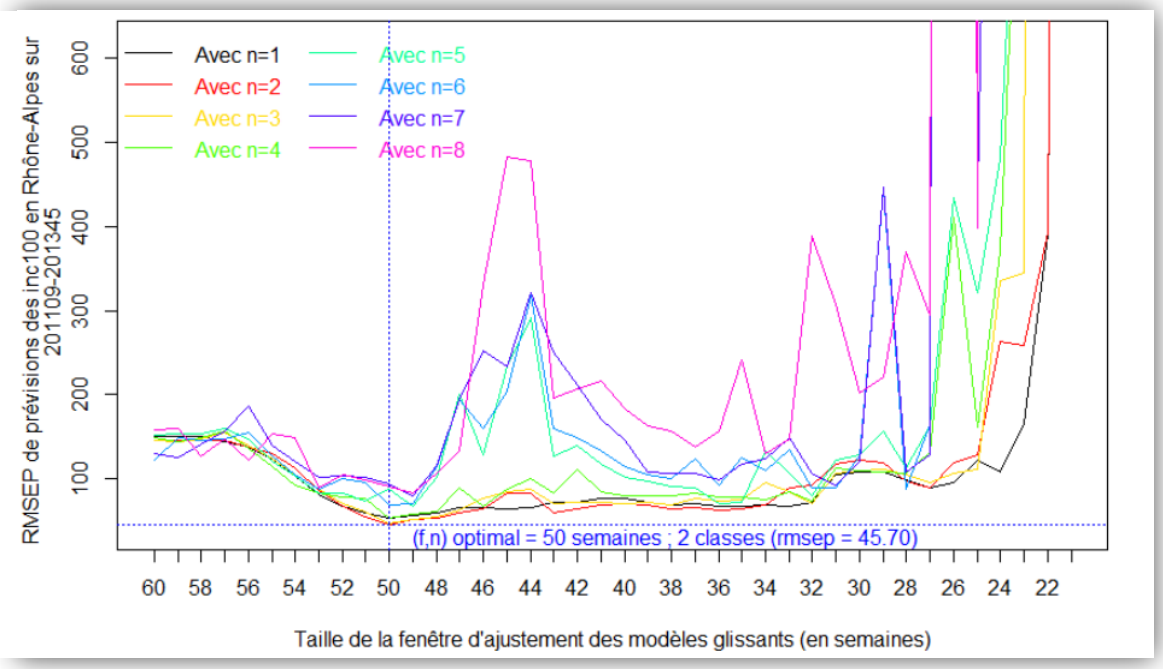


Affinement en Rhône-Alpes sur 201109-201345 :

Paramètres optimaux :

'f' = 50 semaines
'n' = 2 classes

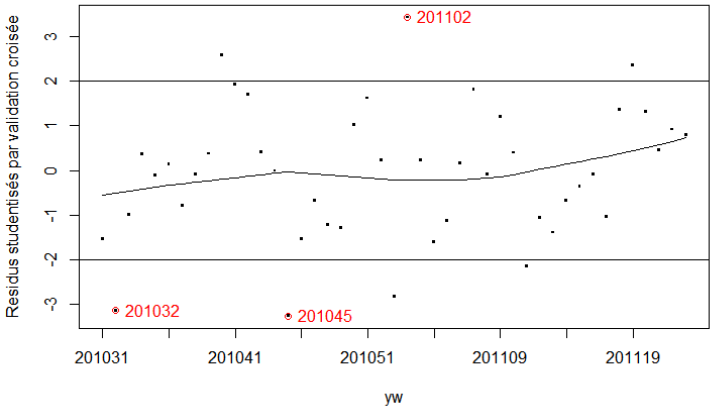
RMSEP = 45,7



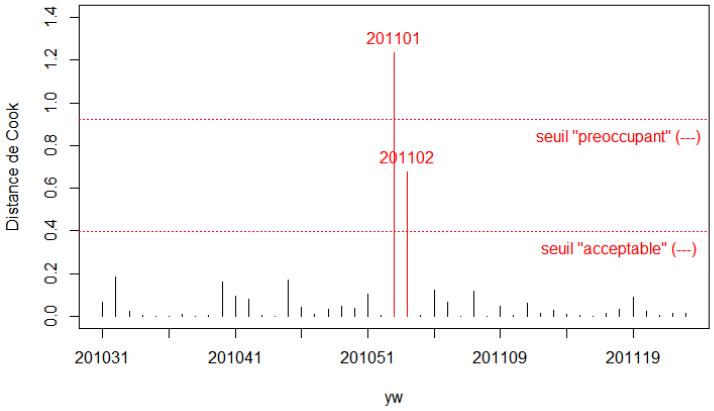
Annexe VIII. Analyses du meilleur modèle de prédiction des incidences nationales : Résidus studentisés, distance de Cook et résidus partiels

Pour la prédiction de la semaine 201124

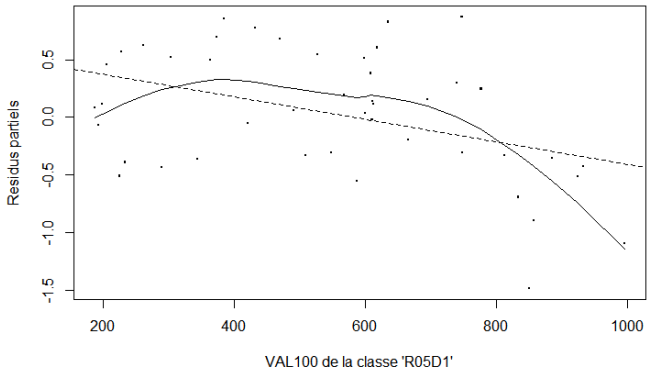
Résidus studentisés par validation croisée du meilleur modèle prédictif pour la semaine 201124



Analyse de l'influence des observations dans le modèle de régression - 201124

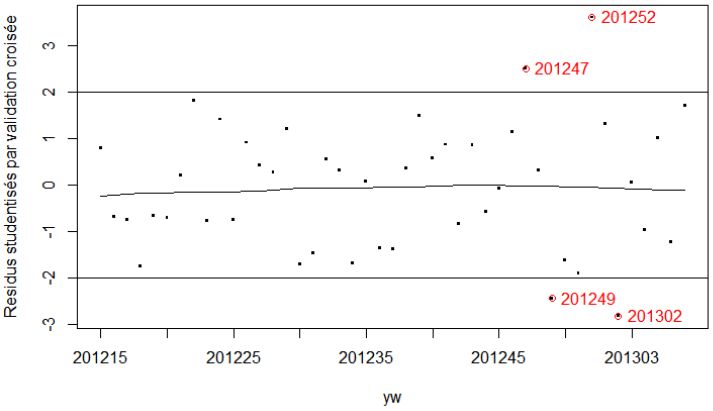


Residus partiels de la classe 'R05D1' pour la prédiction de 201124 par le meilleur modèle

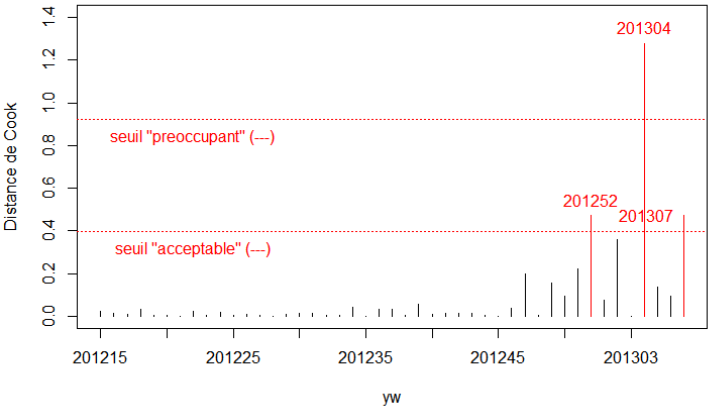


Pour la prédiction de la semaine 201308

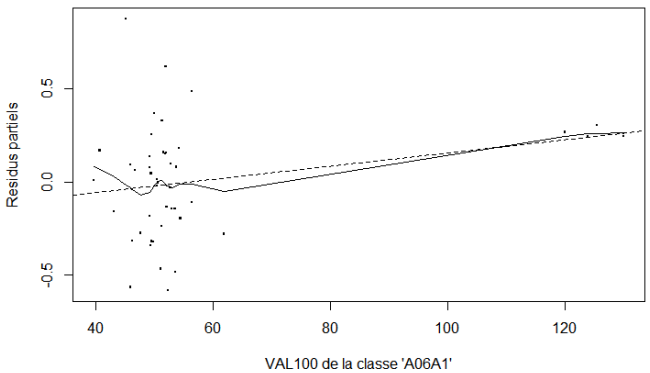
Résidus studentisés par validation croisée du meilleur modèle prédictif pour la semaine 201308



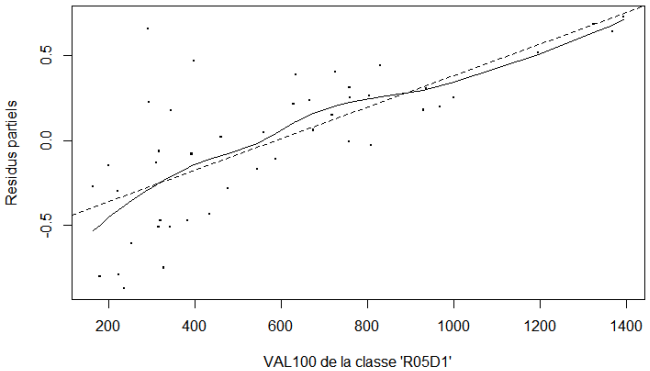
Analyse de l'influence des observations dans le modèle de régression - 201308



Residus partiels de la classe 'A06A1' pour la prédiction de 201308 par le meilleur modèle

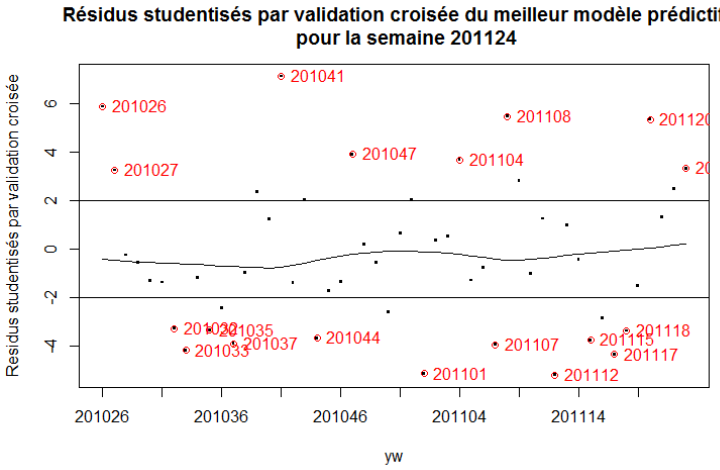


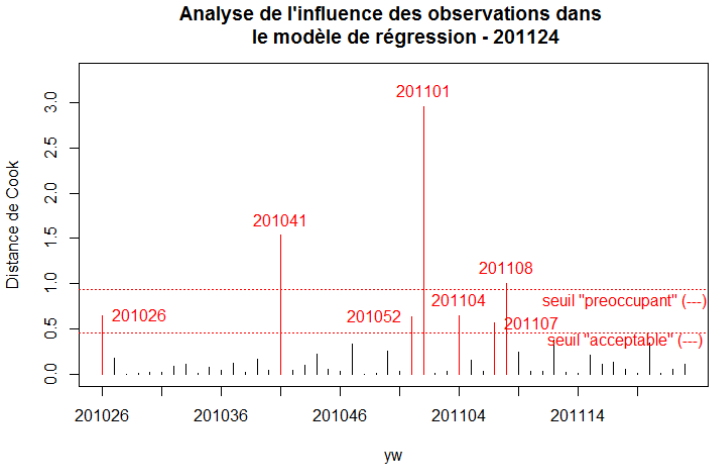
Residus partiels de la classe 'R05D1' pour la prédiction de 201308 par le meilleur modèle

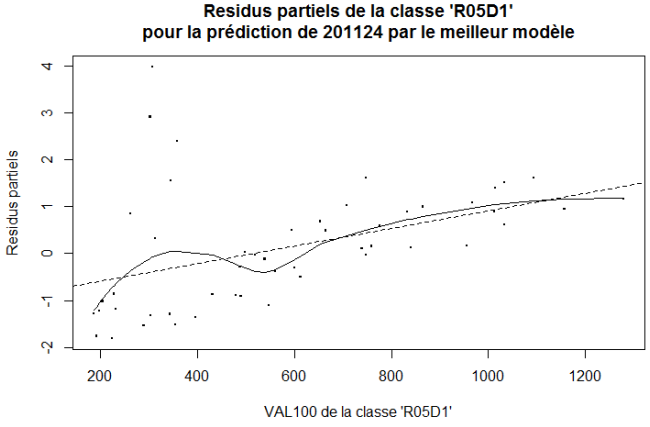


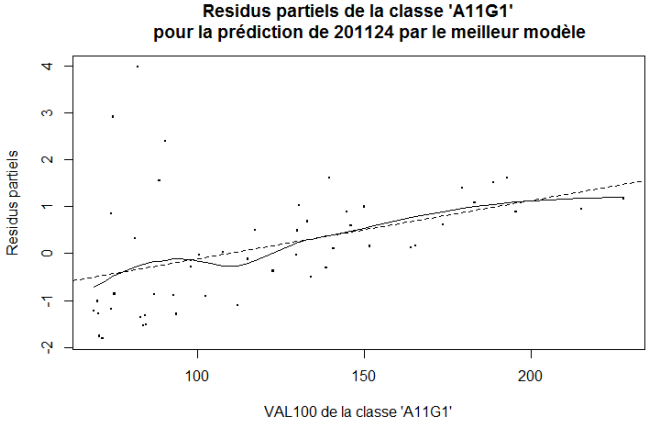
Annexe IX. Analyses du meilleur modèle de prédiction des incidences en Rhône-Alpes : Résidus studentisés, distance de Cook et résidus partiels

Pour la prédiction de la semaine 201124

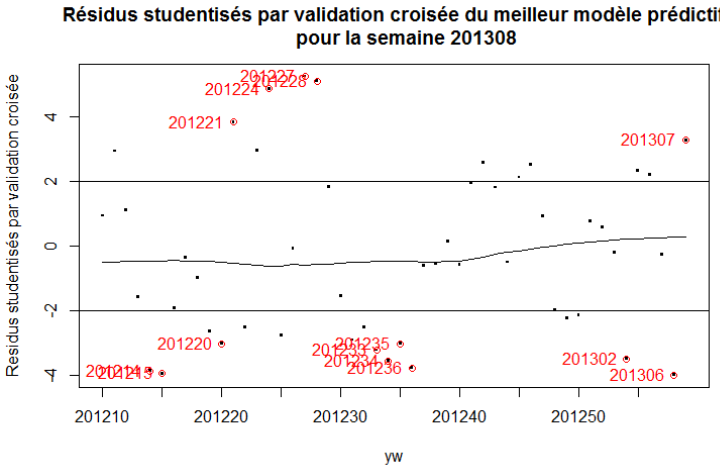


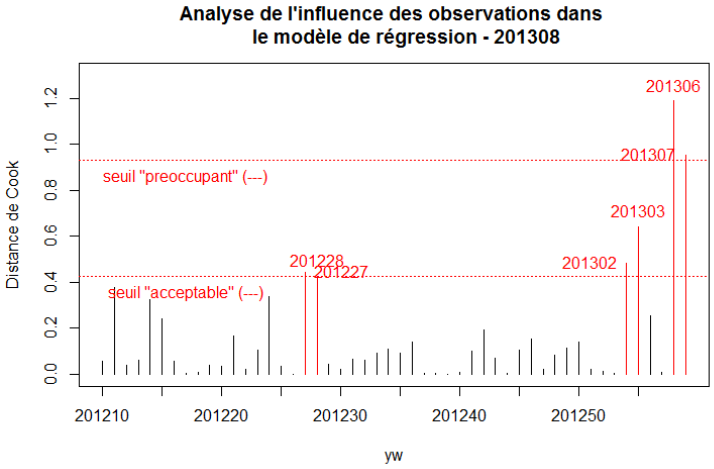


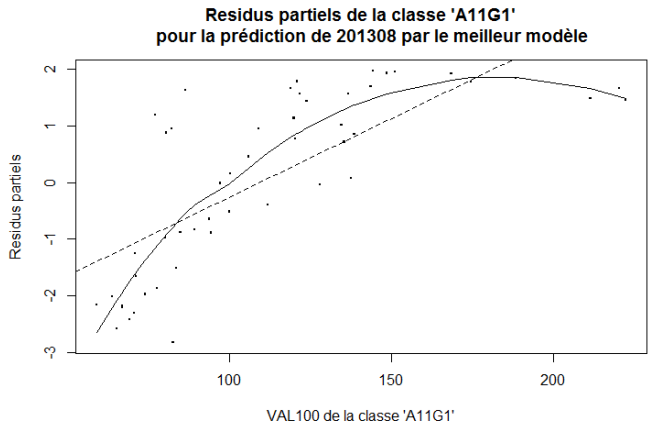




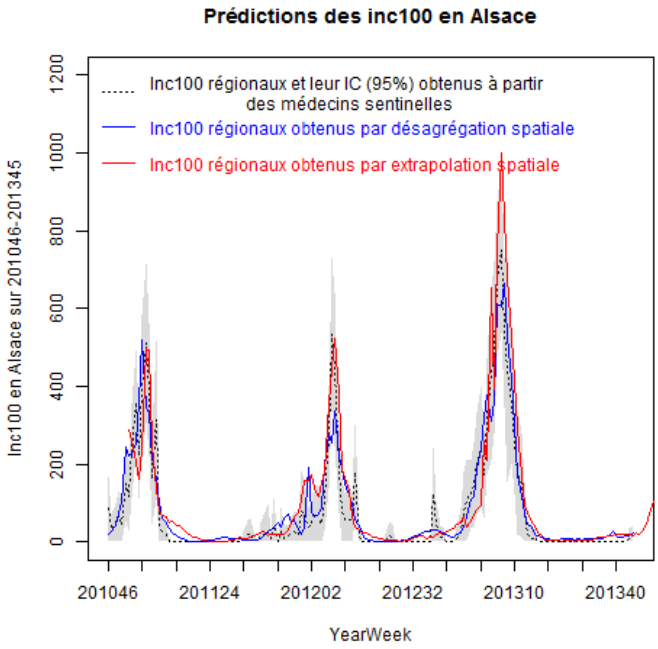
Pour la prédiction de la semaine 201308

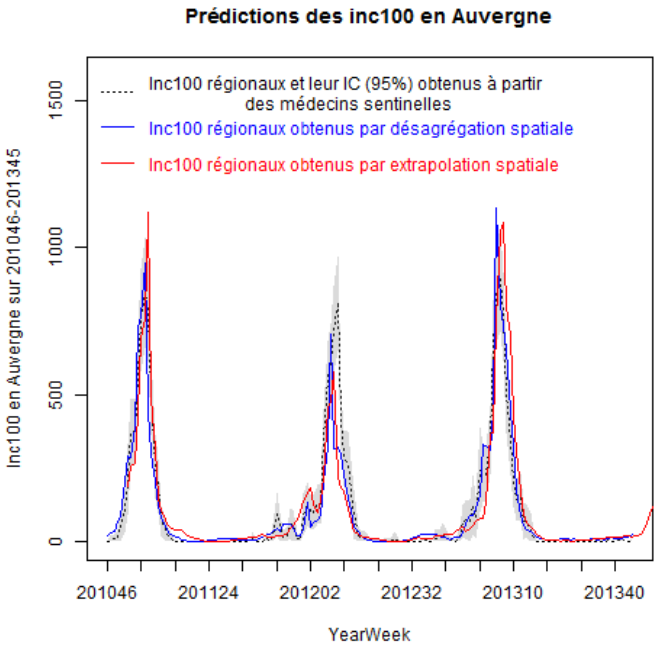


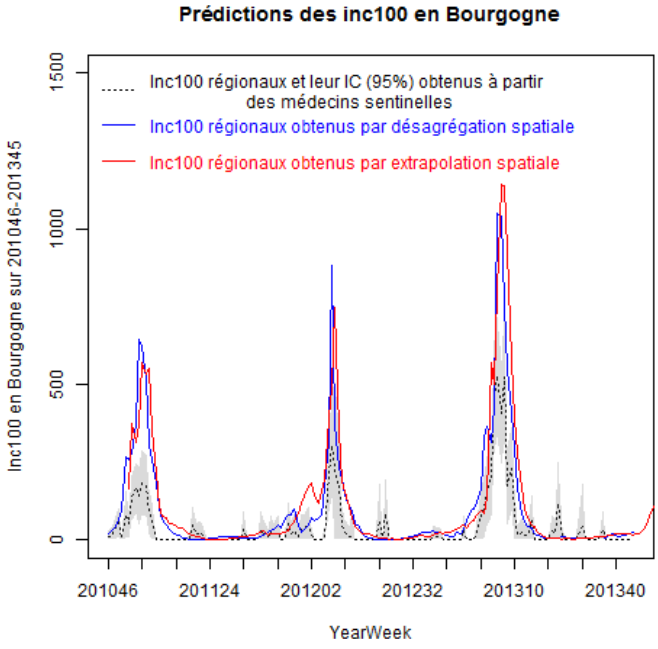


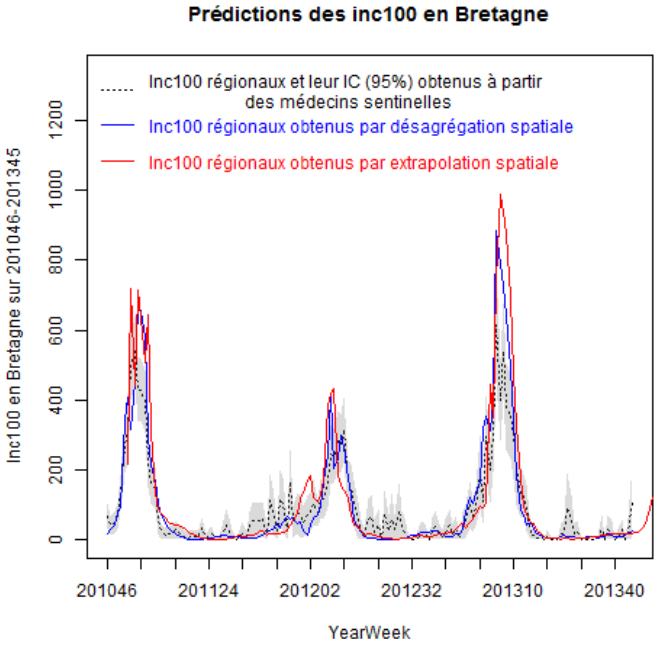


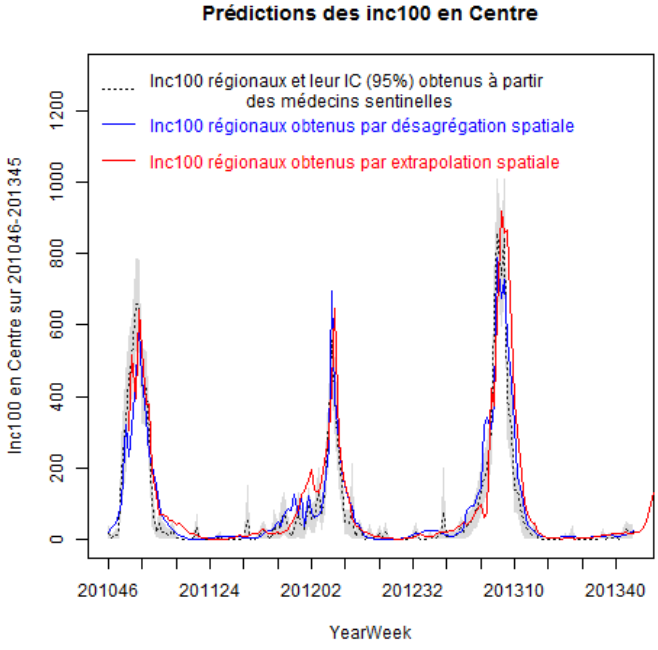
Annexe X. Prédictions des nouvelles incidences régionales par désagrégation et extrapolation spatiales

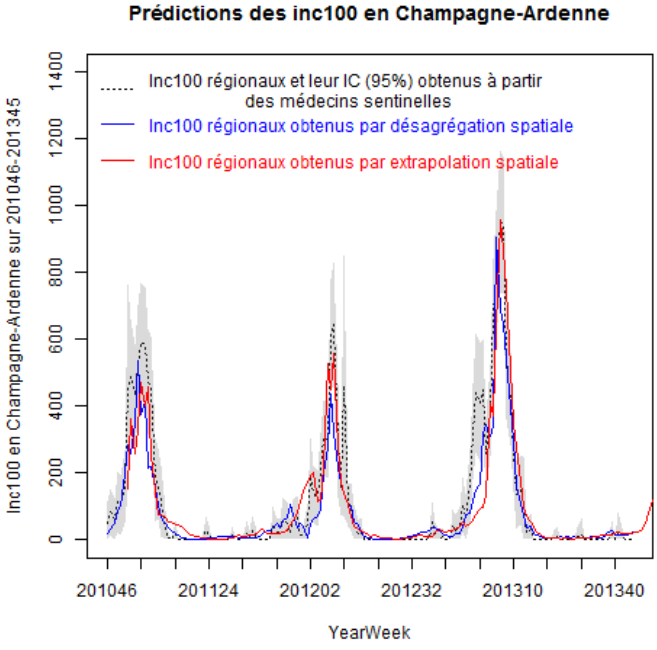




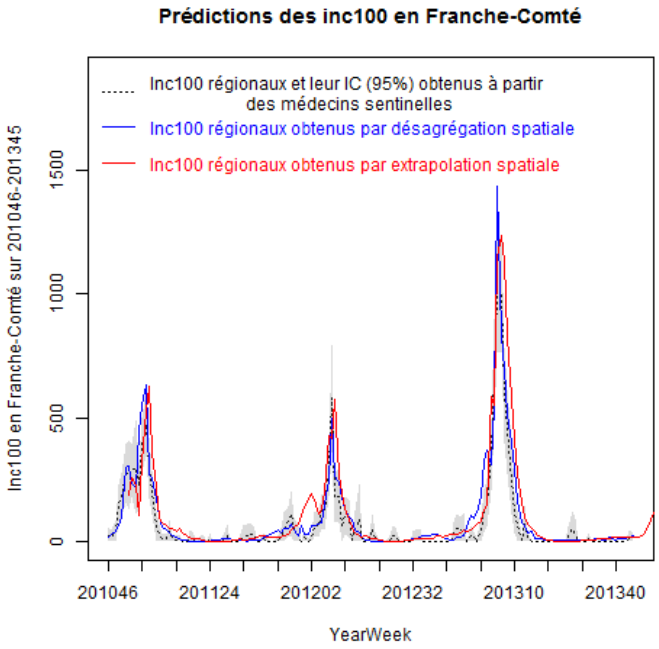


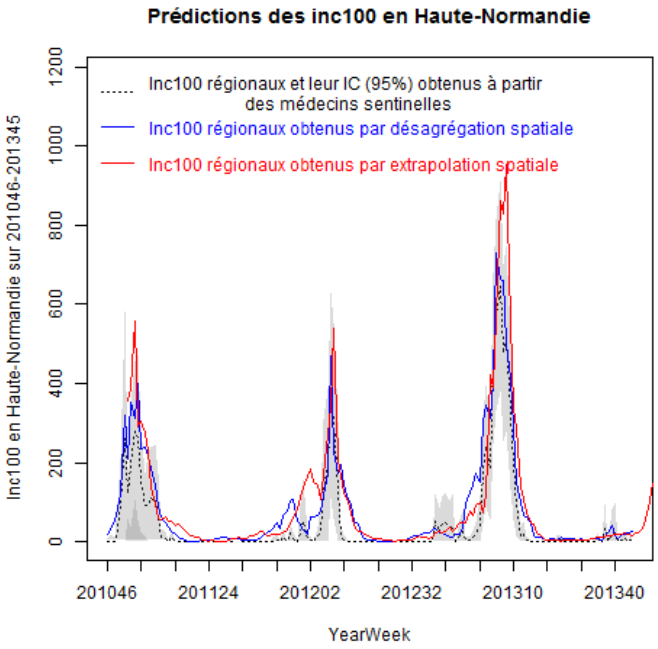


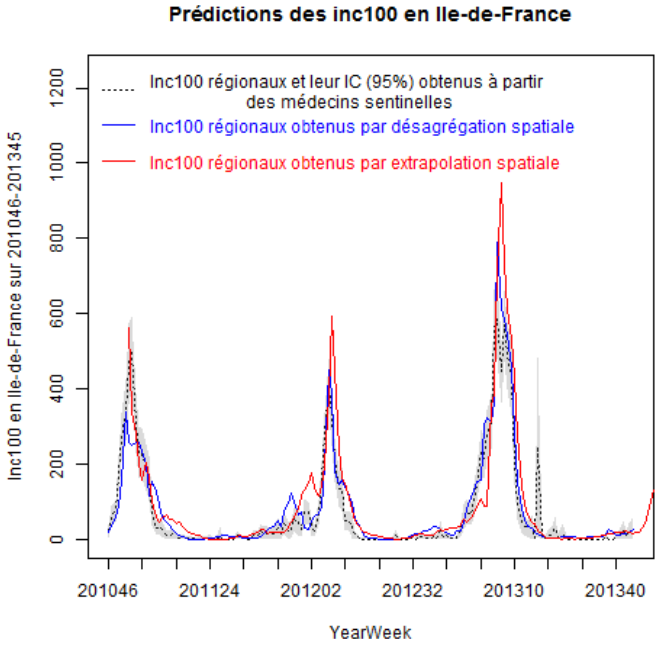


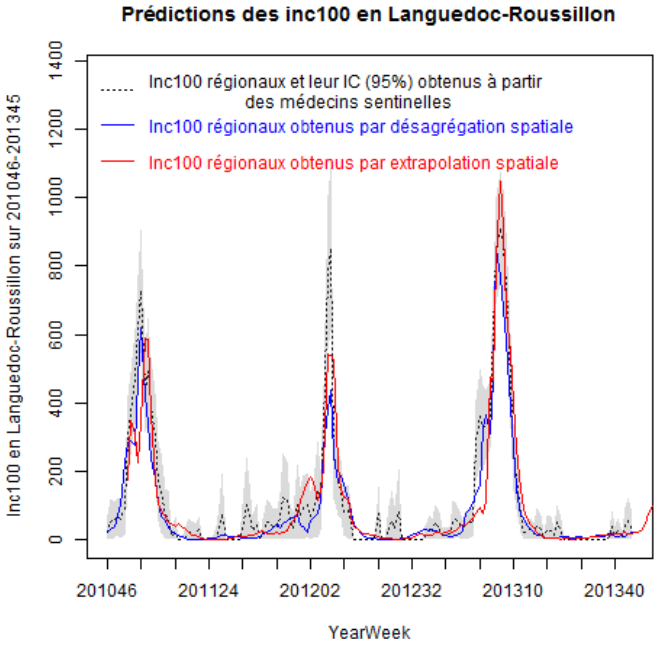


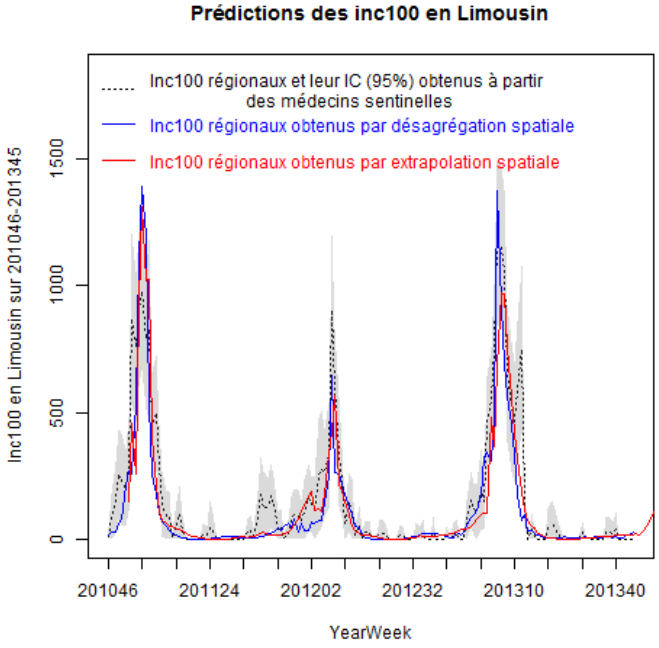
Annexe X - Suite (2/3)

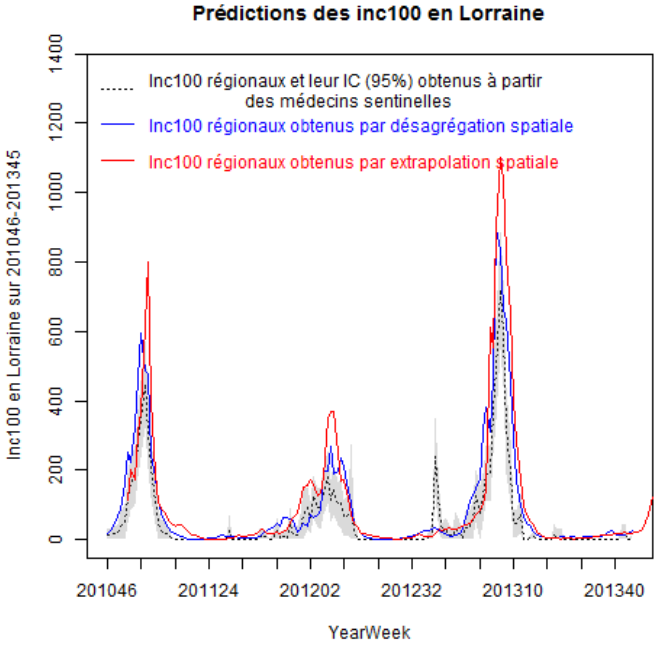




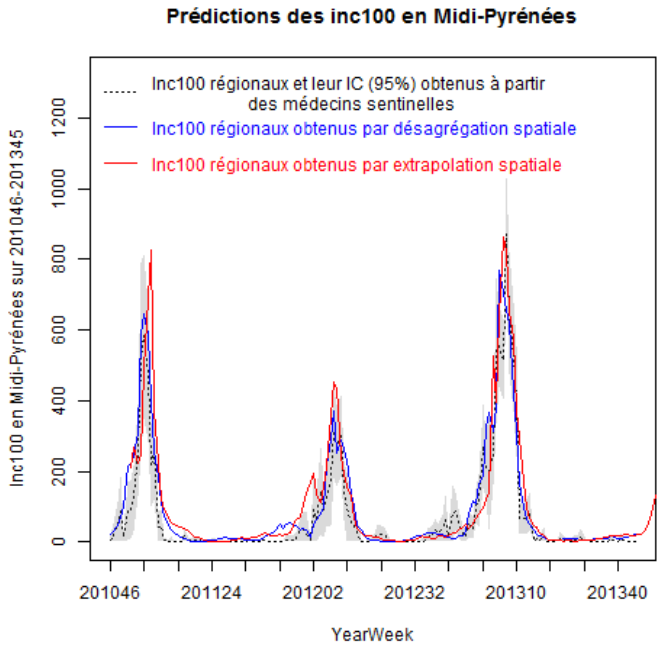


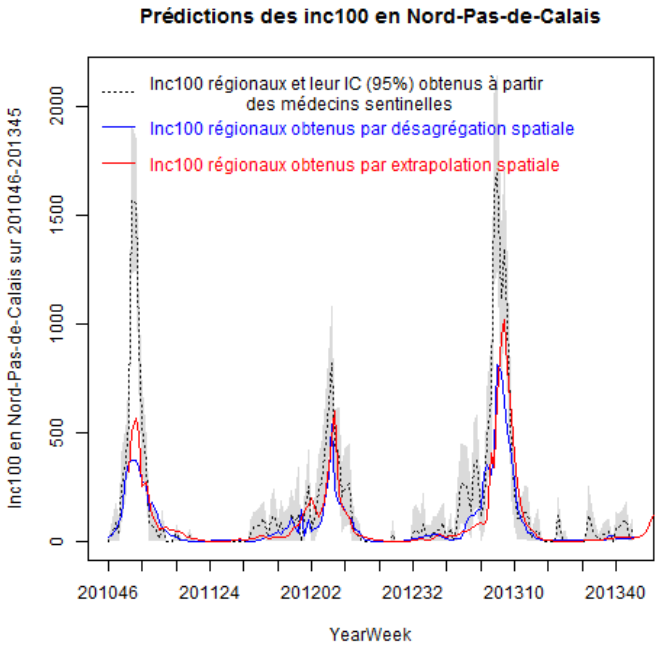


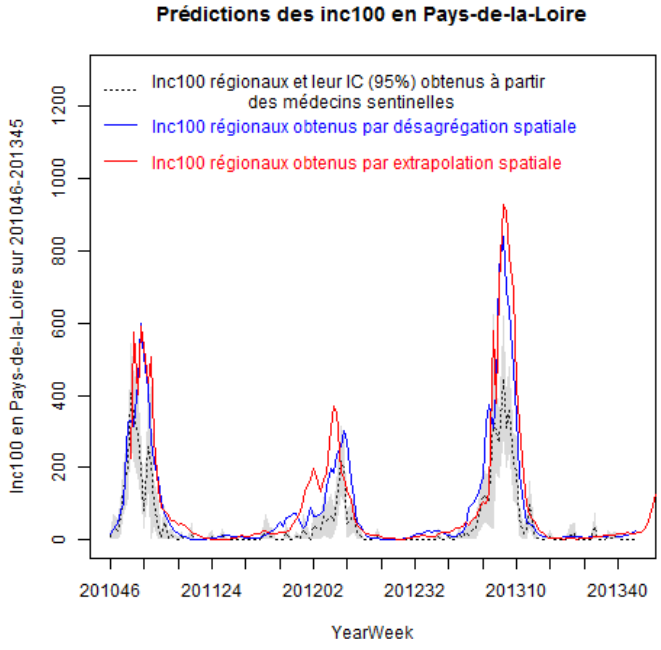


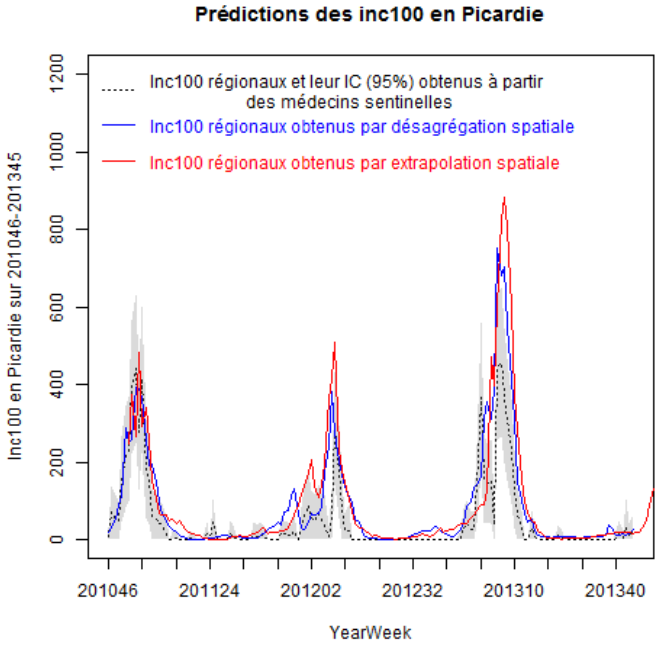


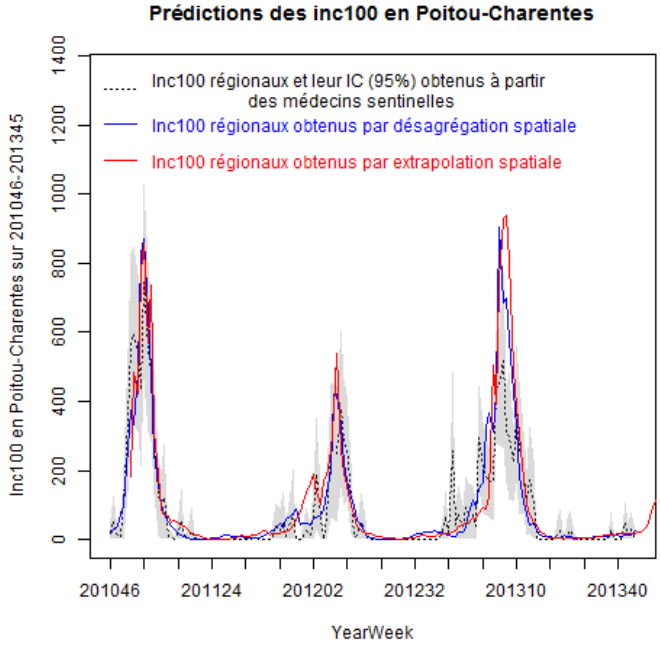
Annexe X - Suite (3/3)

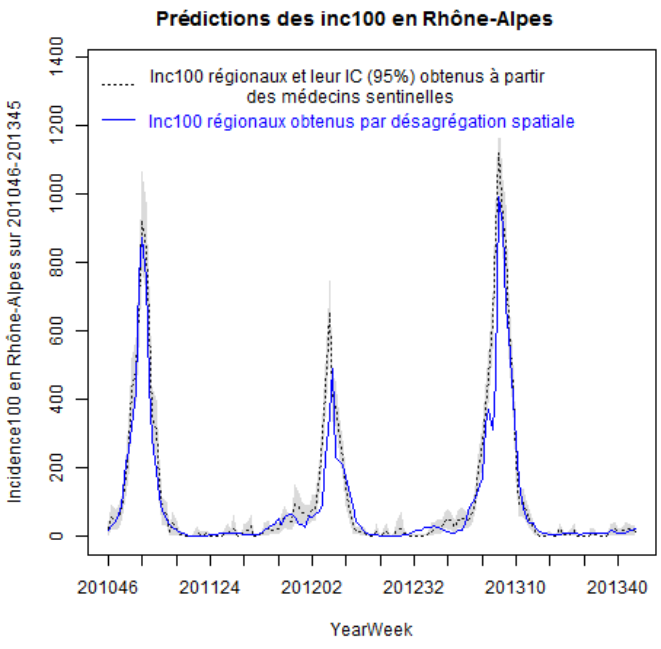




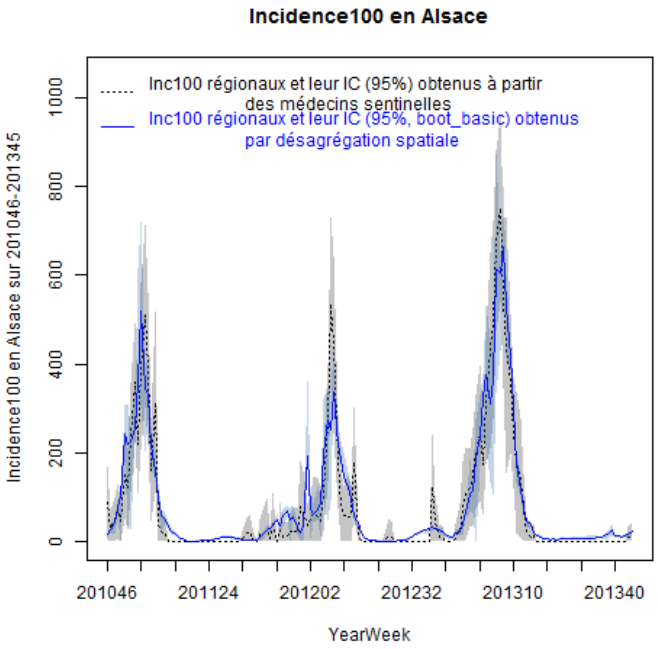


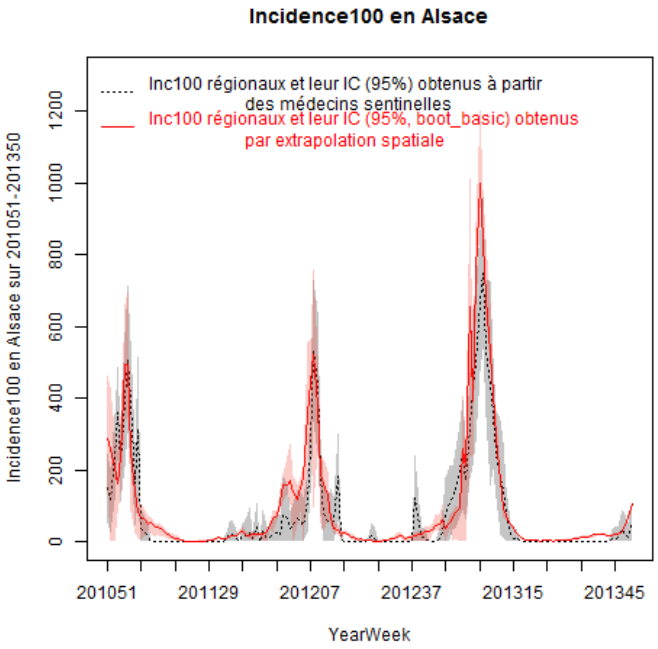


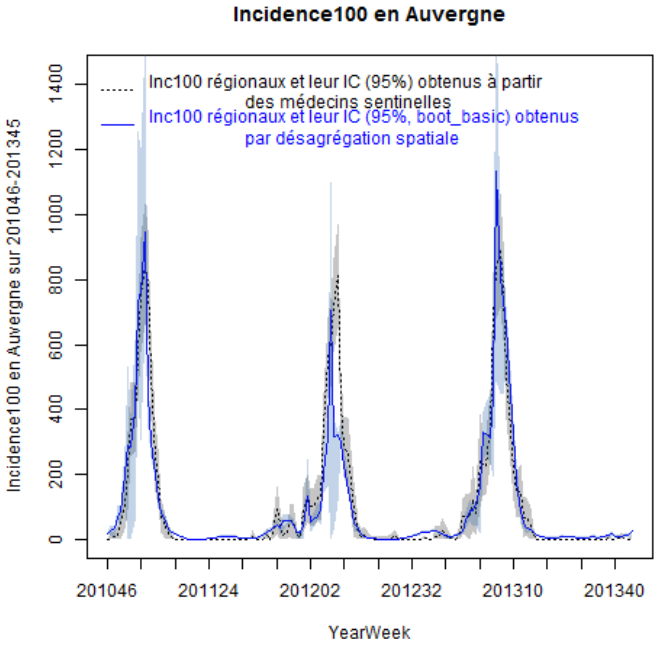


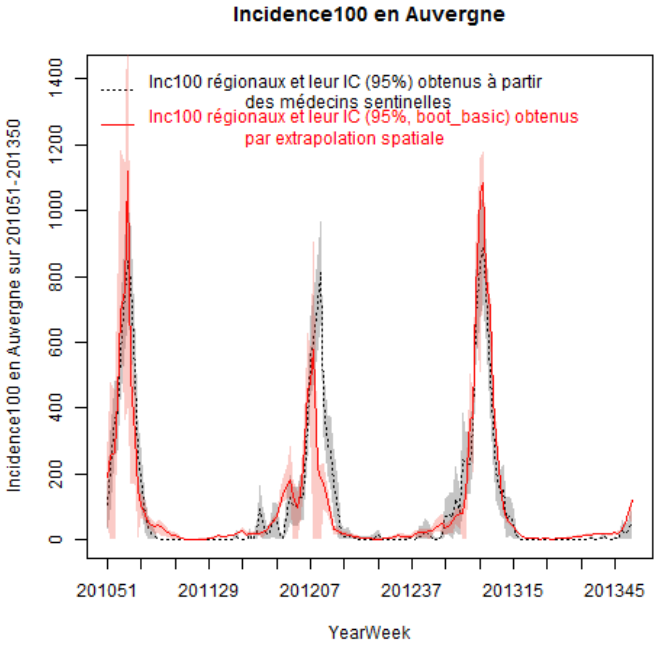


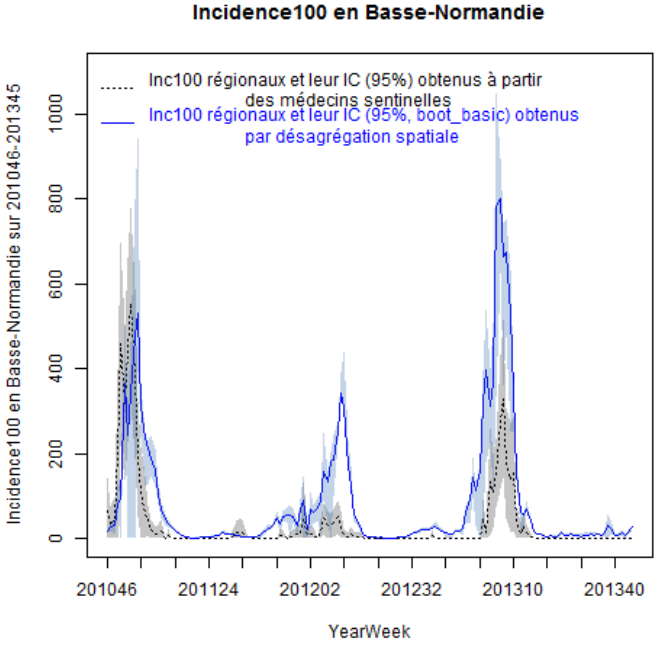
Annexe XI. *Basic bootstrap* des prédictions régionales des inc100 issues de la désagrégation (gauche) et de l'extrapolation (droite) spatiales

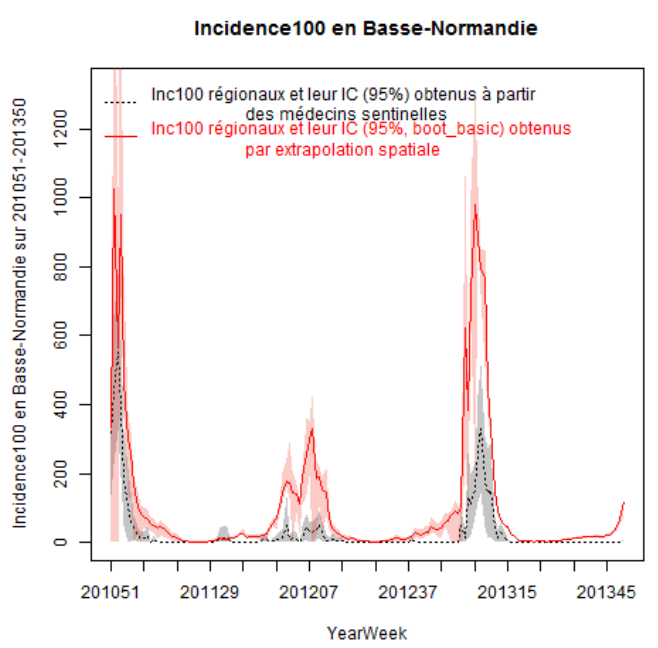




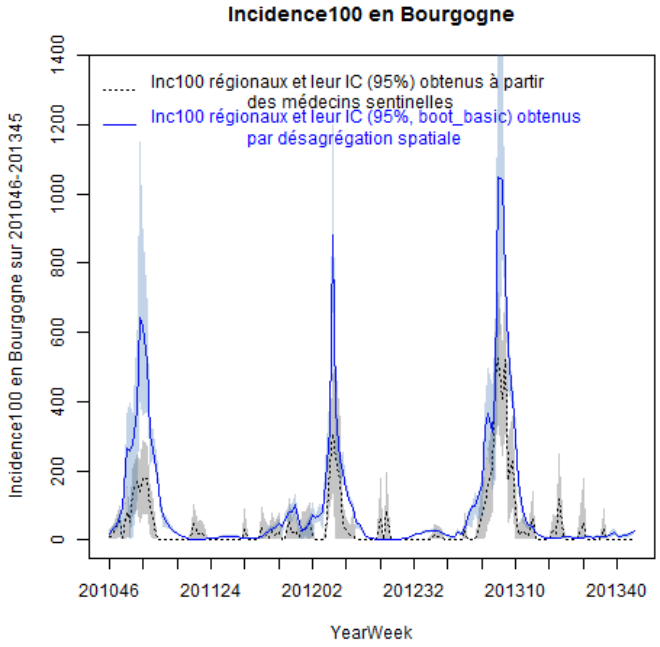


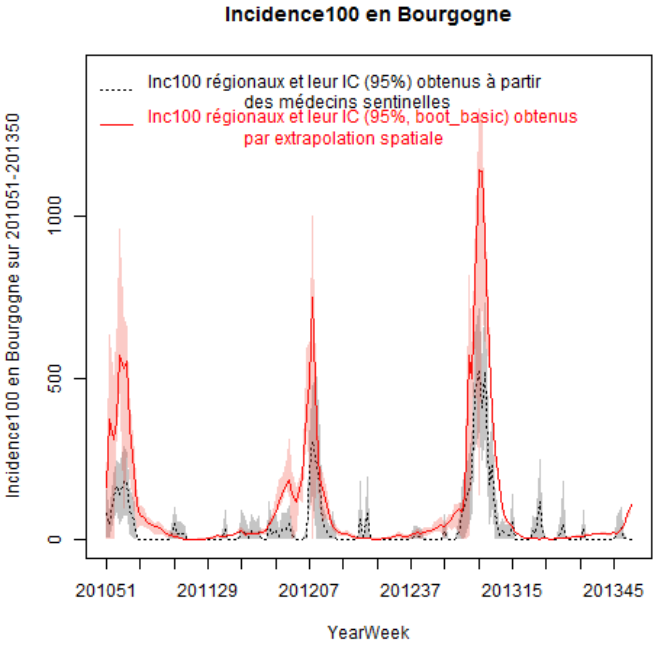


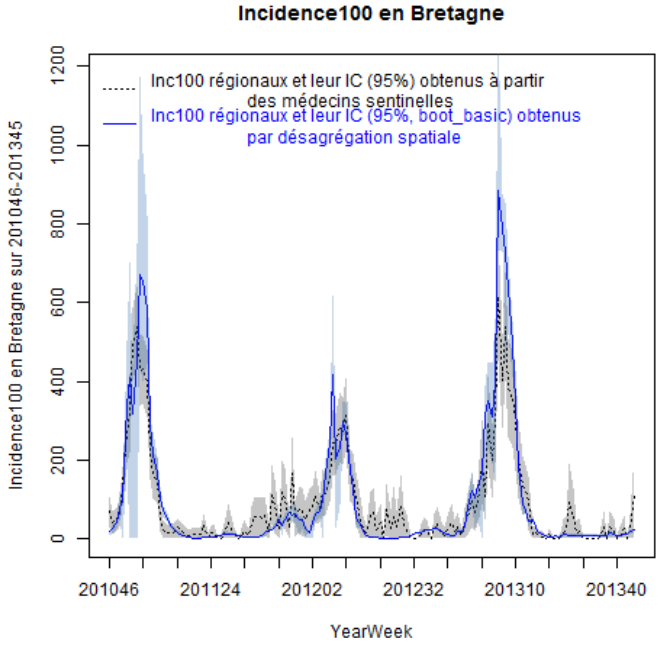


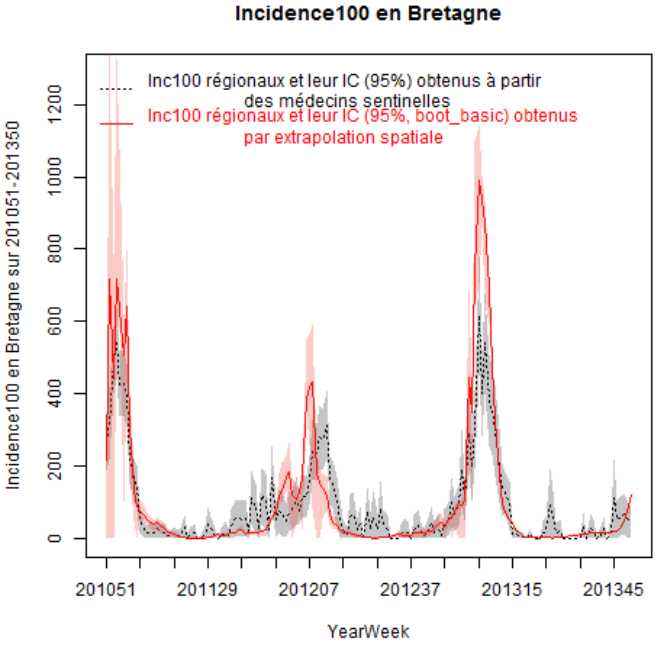


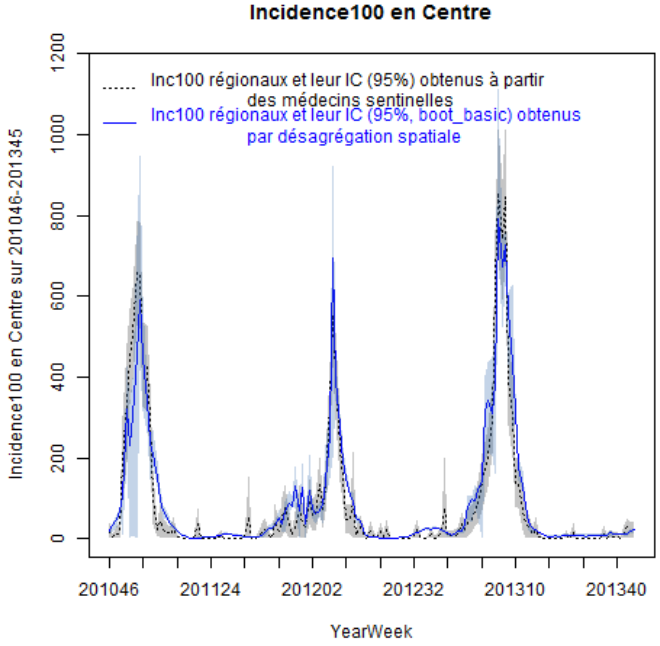
Annexe XI - Suite (2/7)

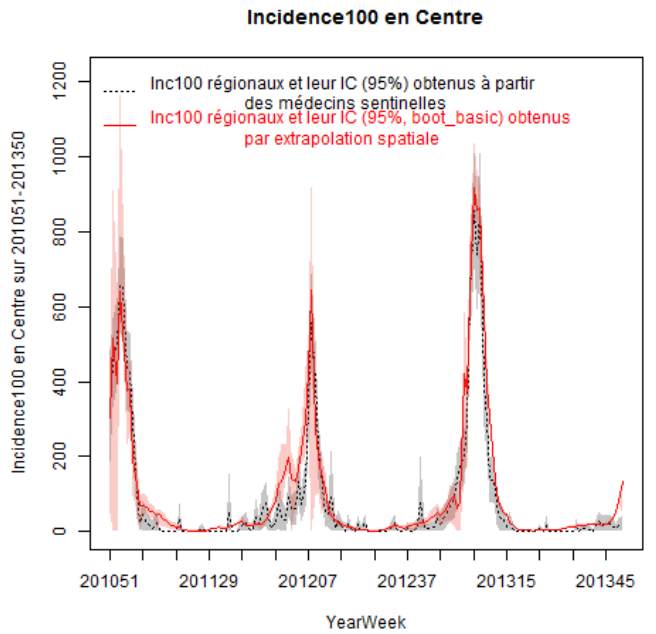




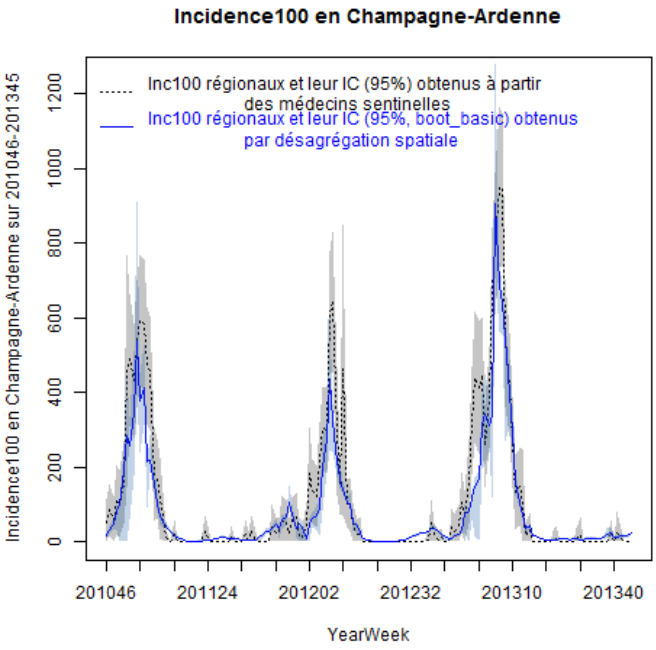


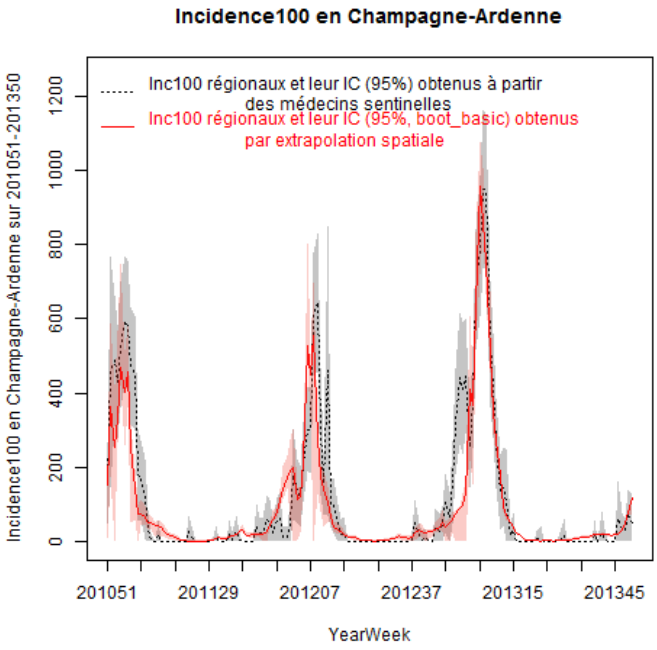


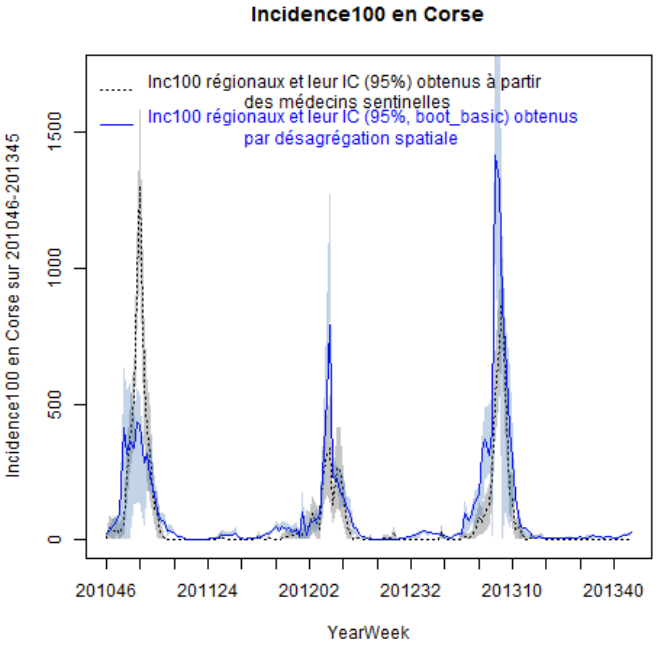


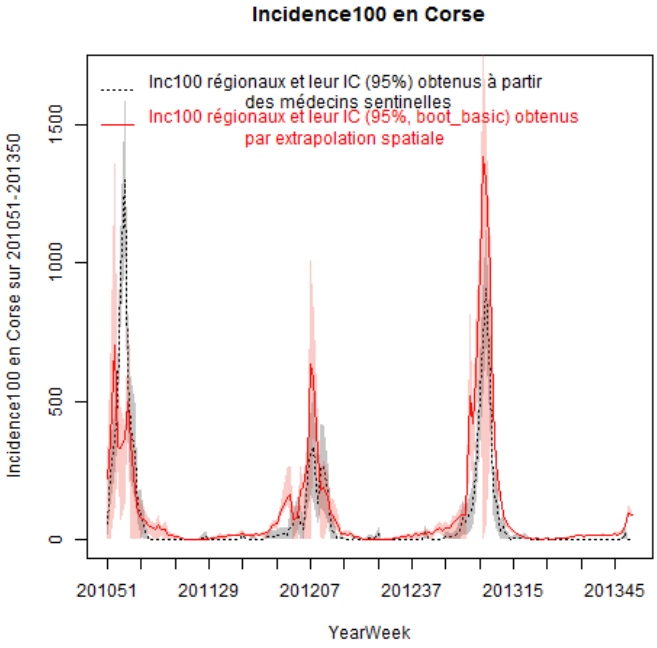


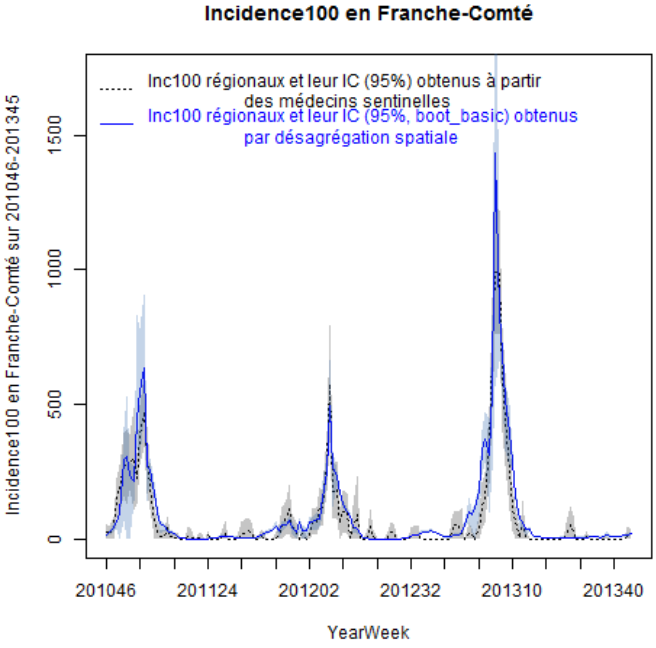
Annexe XI - Suite (3/7)

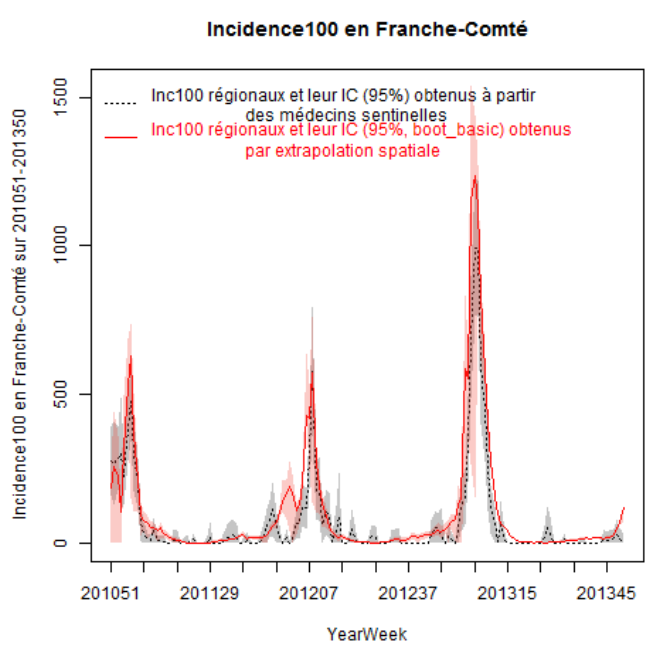




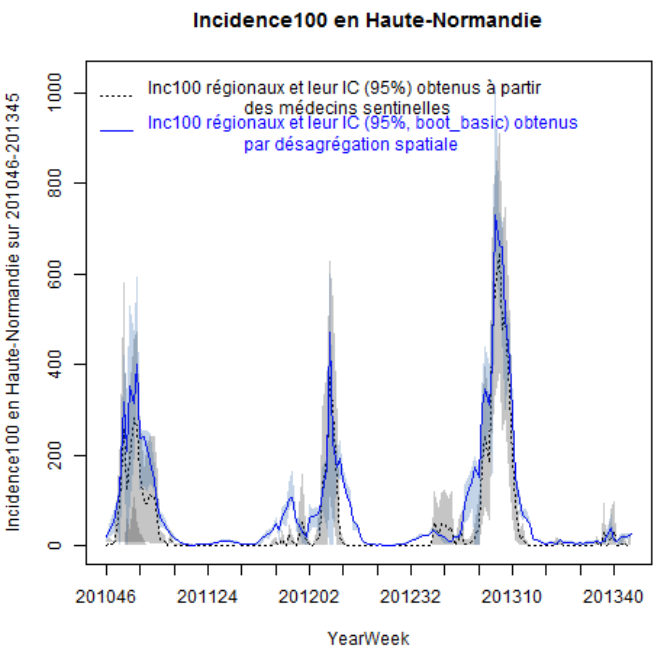


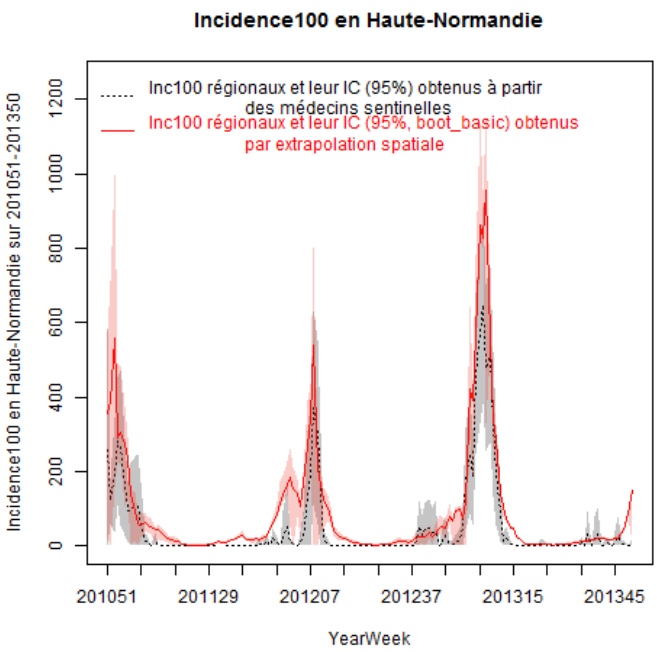


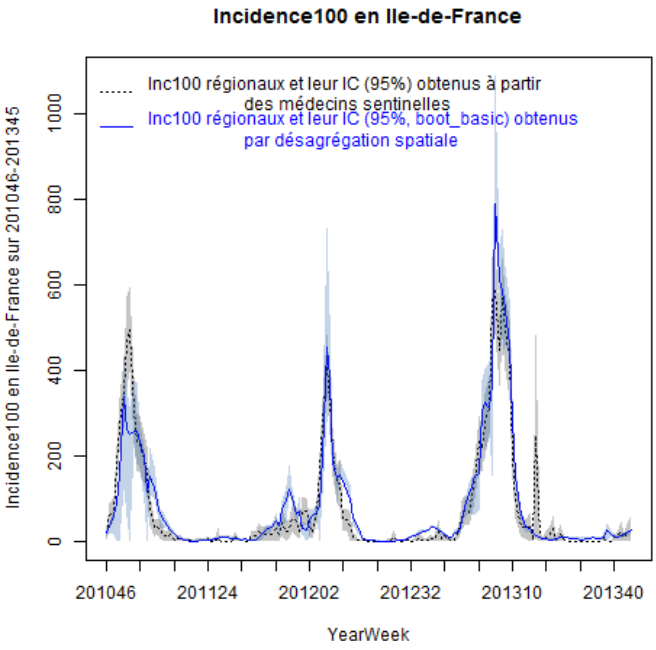


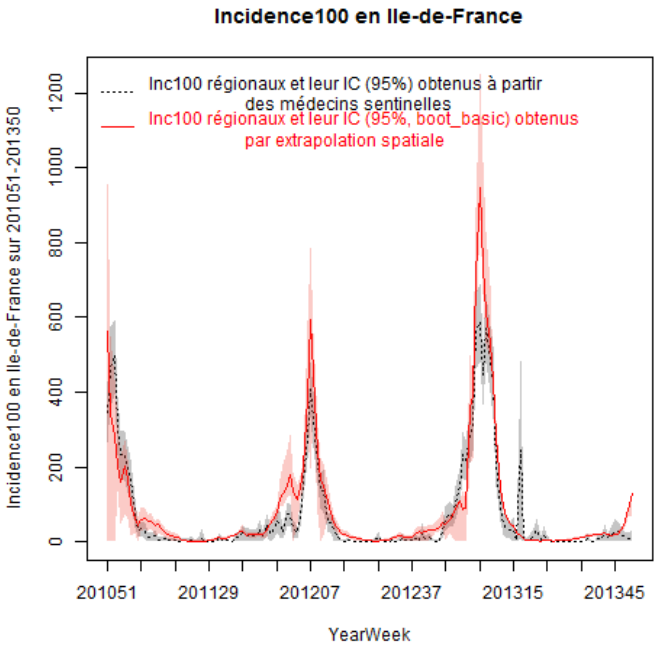


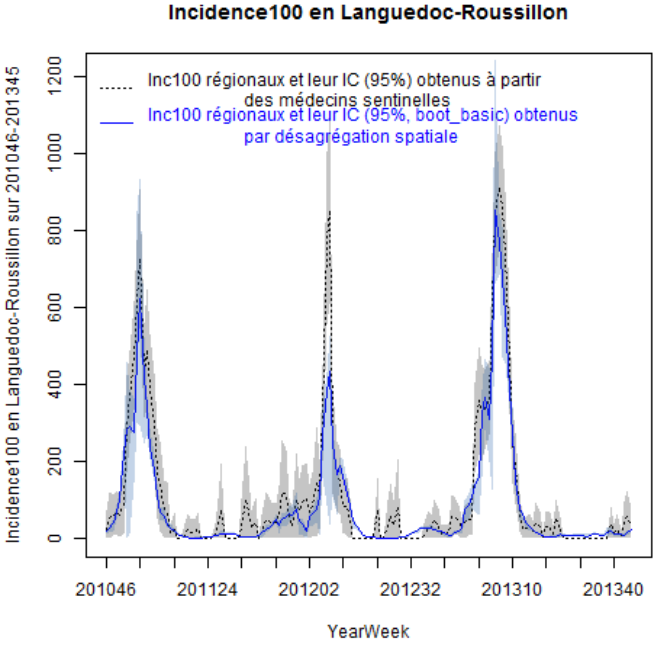
Annexe XI - Suite (4/7)

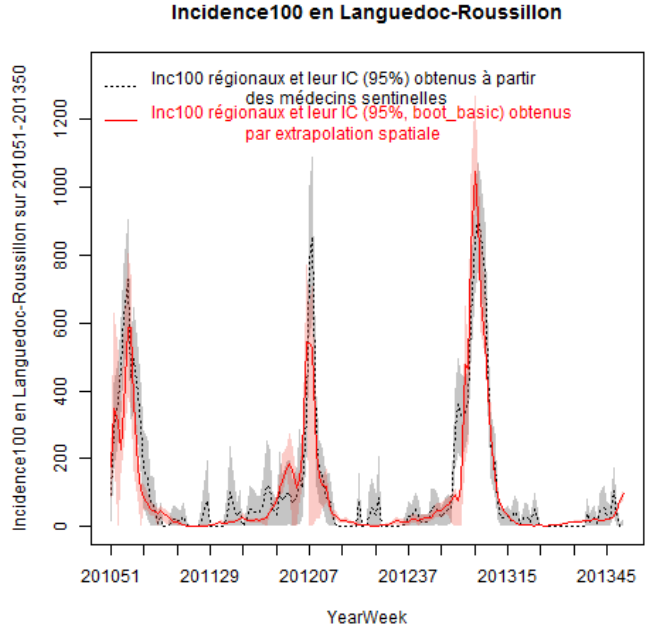




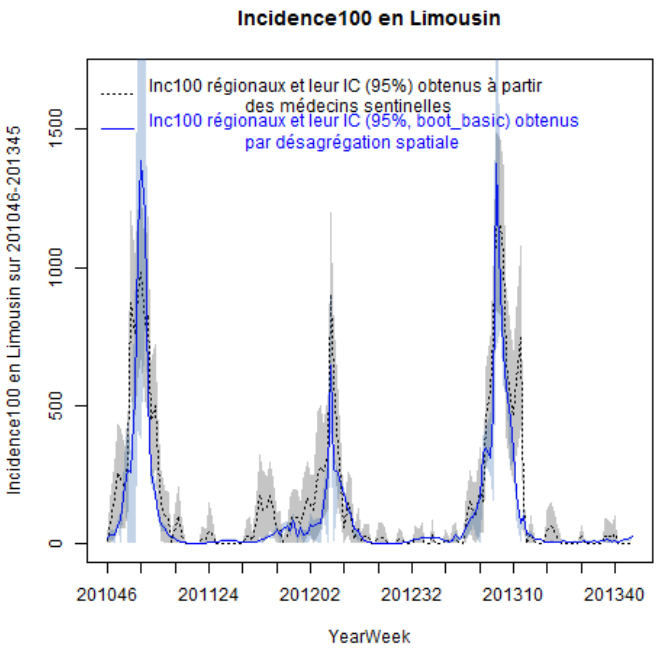


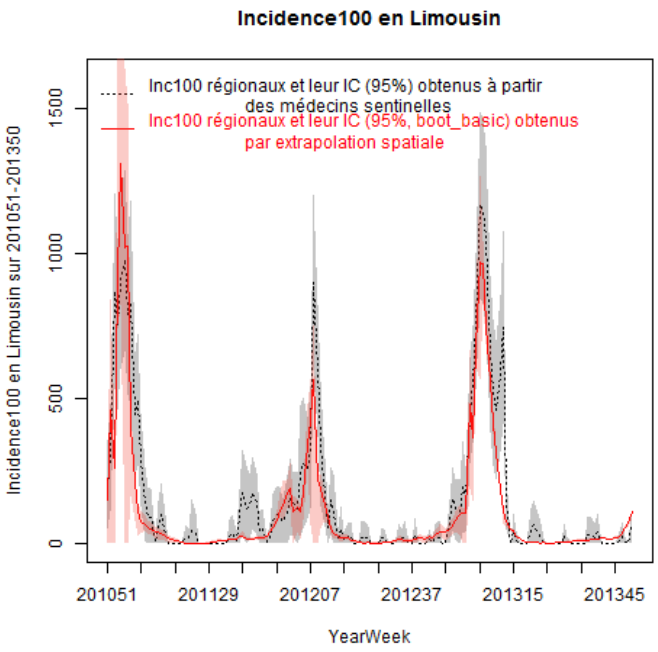


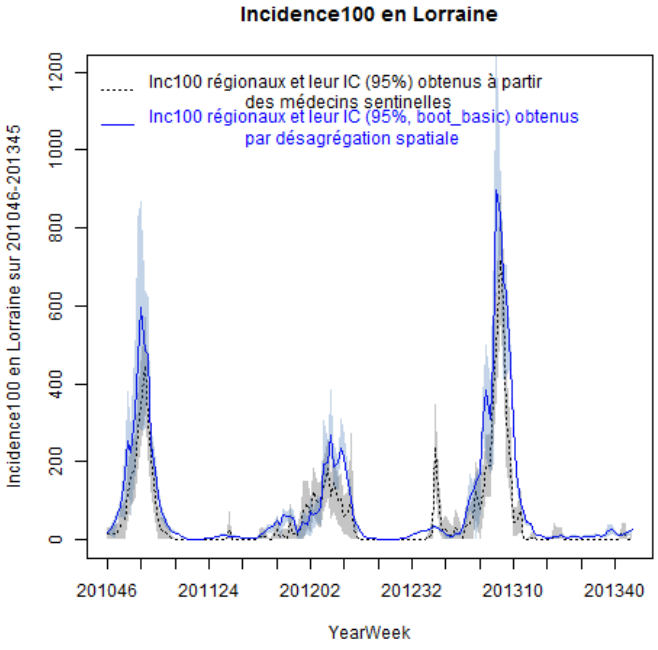


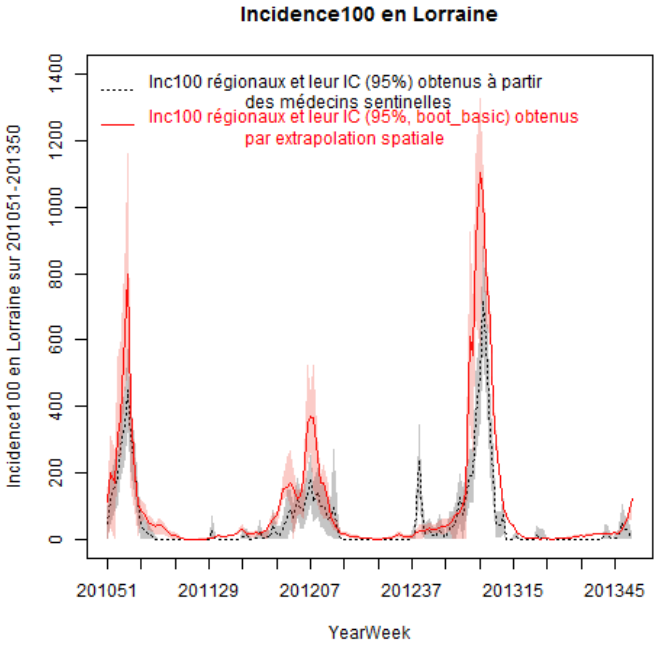


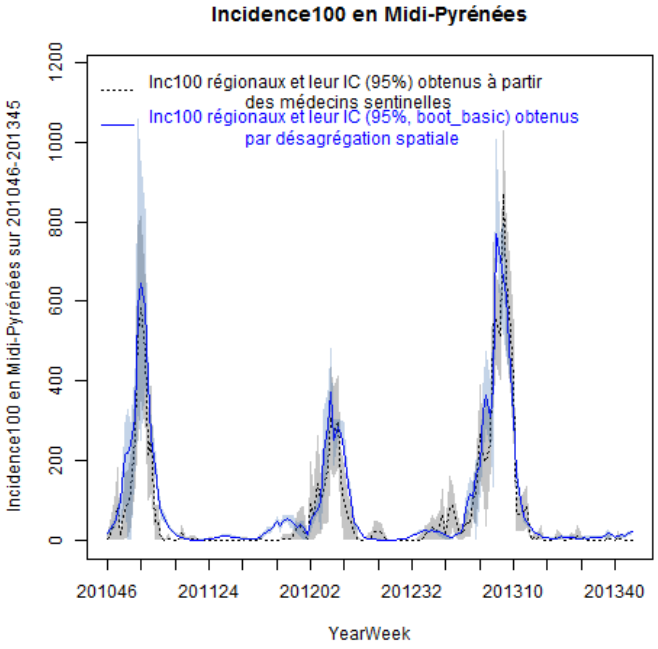
Annexe XI - Suite (5/7)

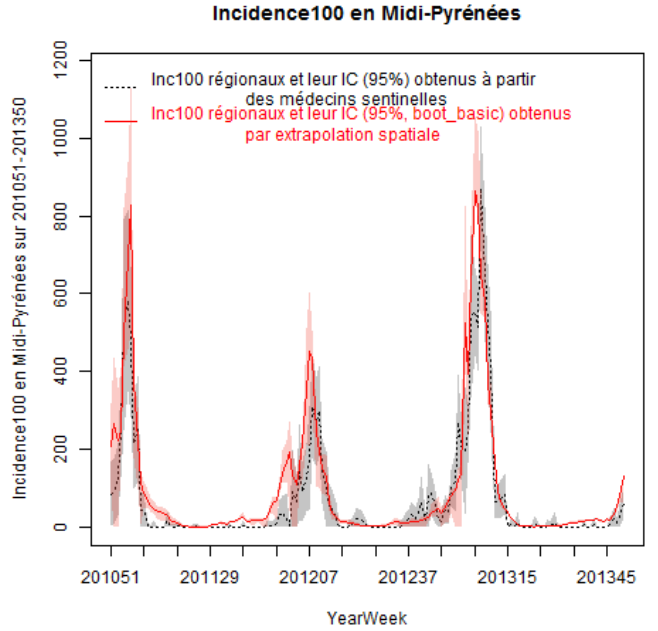




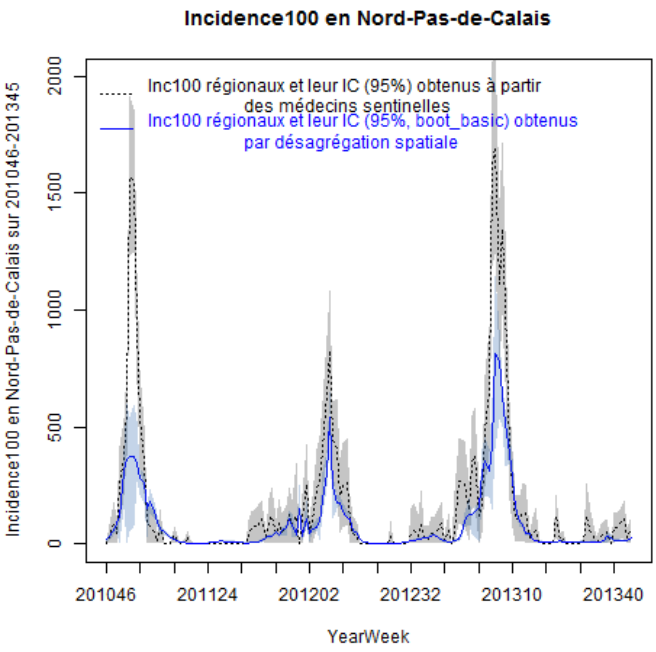


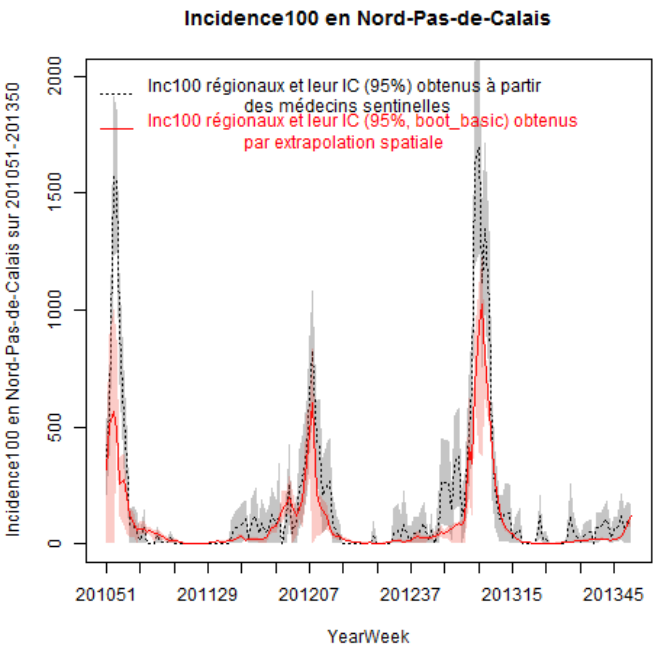


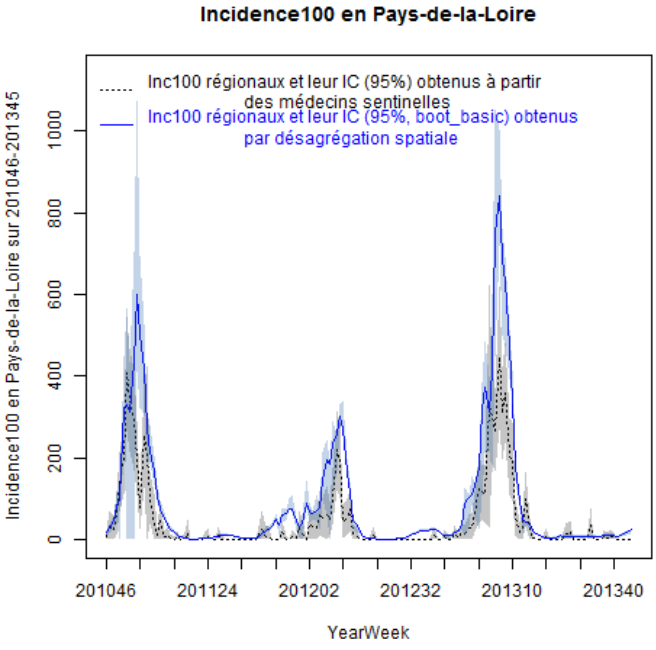


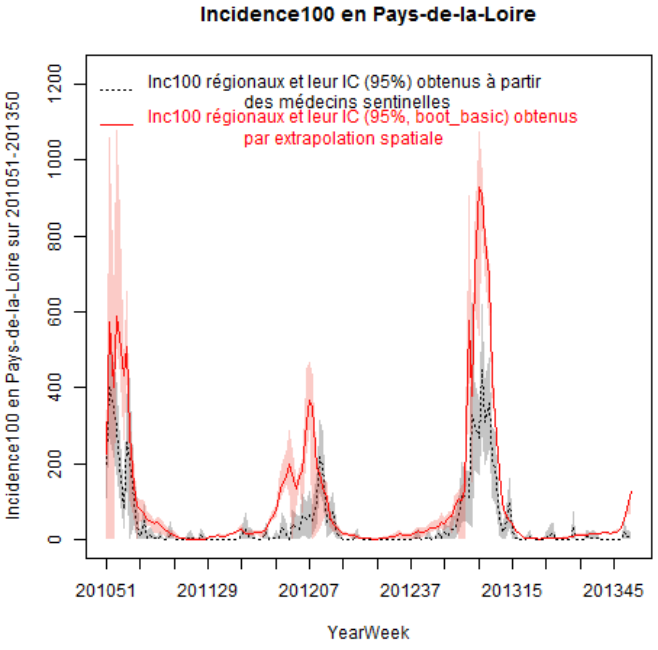


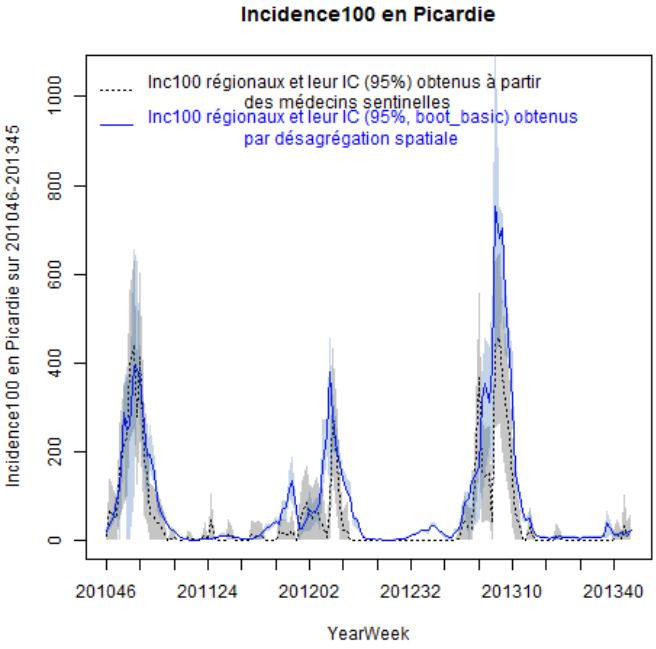
Annexe XI - Suite (6/7)

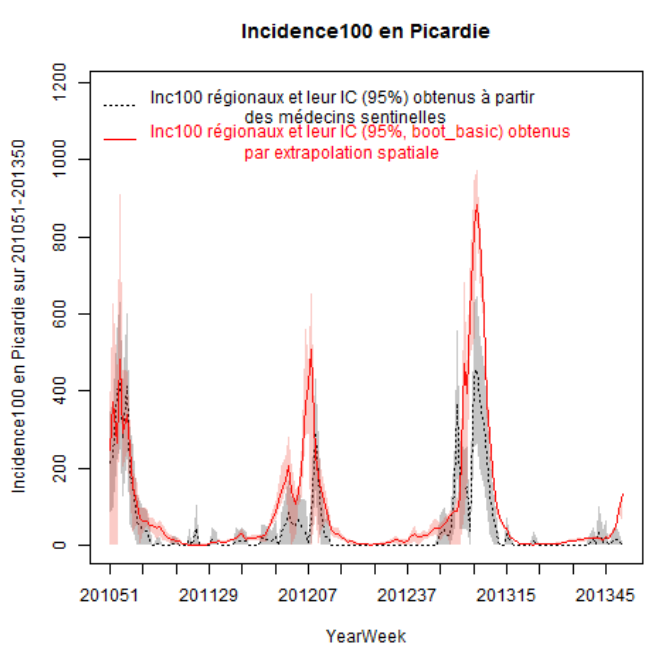




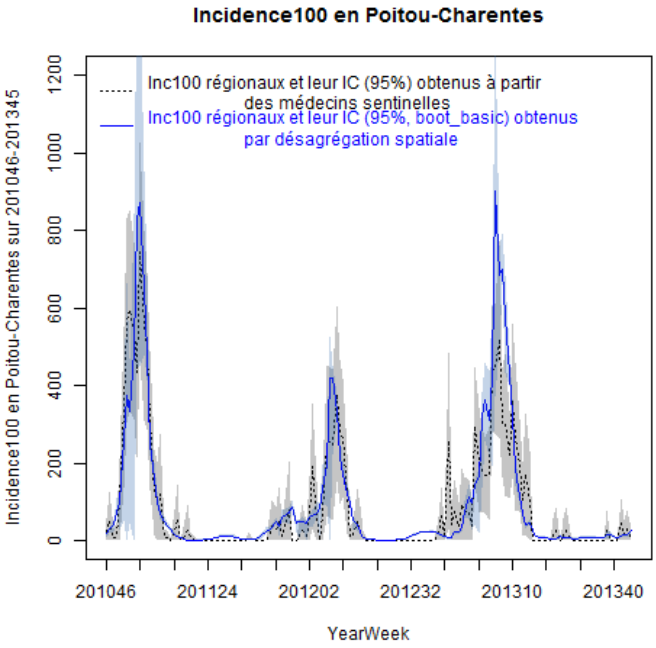


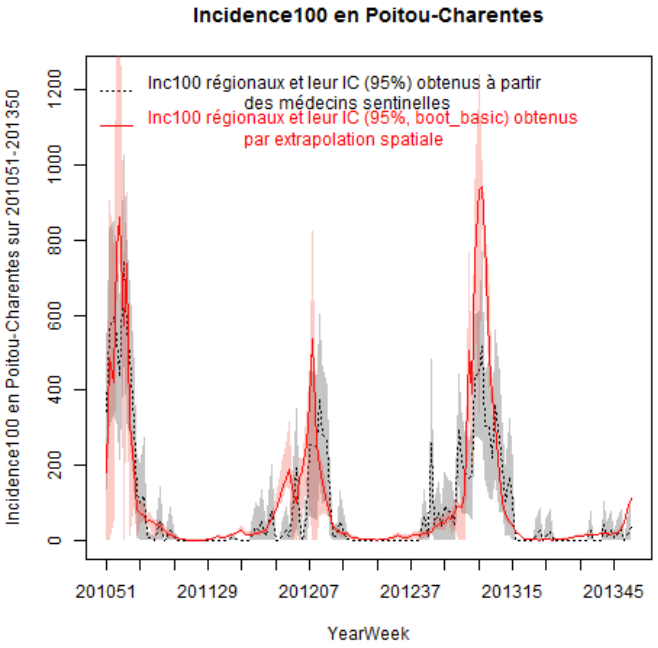


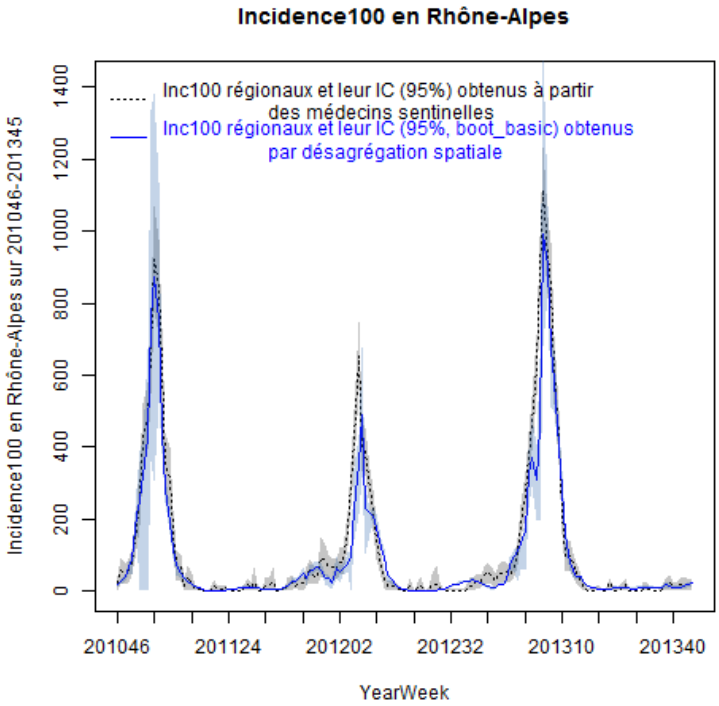




Annexe XI - Suite (7/7)







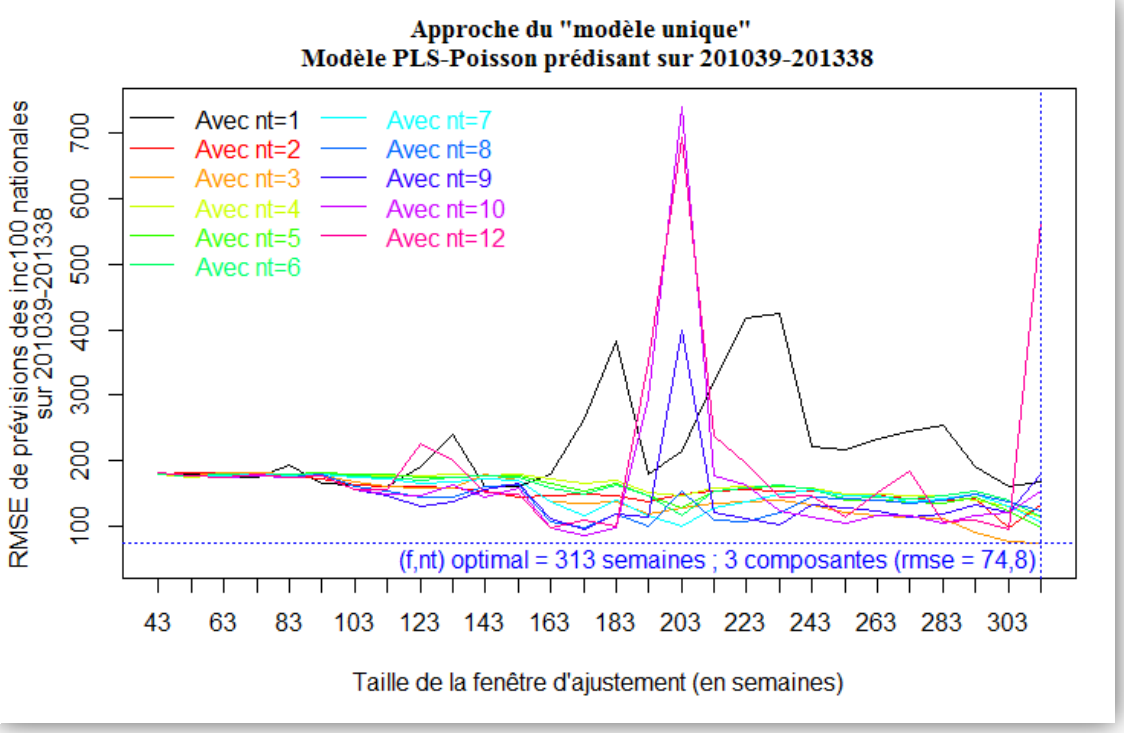
Annexe XII. La minimisation des erreurs de prédictions des inc100 nationaux par des modèles PLS-Poisson en utilisant toute la base de données médicamenteuses

« Modèle unique » PLS-Poisson sur toutes les classes disponibles

Paramètres optimaux :

'f' = 313 semaines
'nt' = 3 valeurs latentes

RMSEP = 74,8

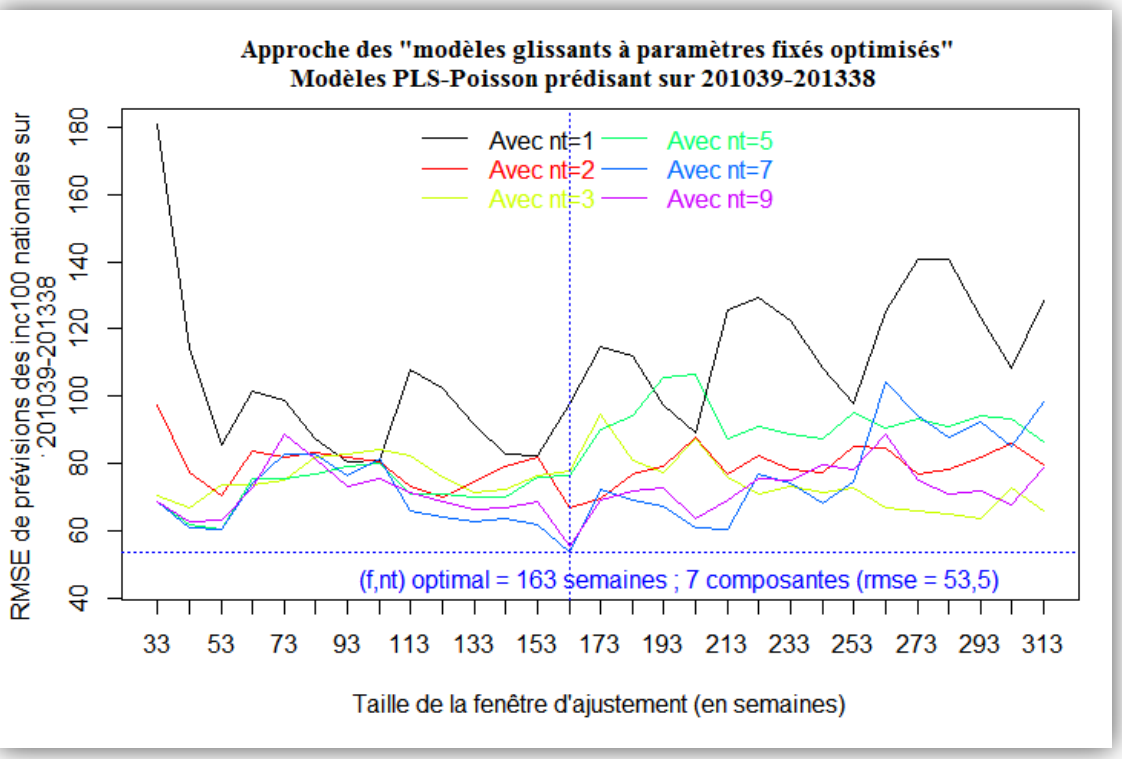


« Modèles glissants » PLS-Poisson sur toutes les classes disponibles

Paramètres optimaux :

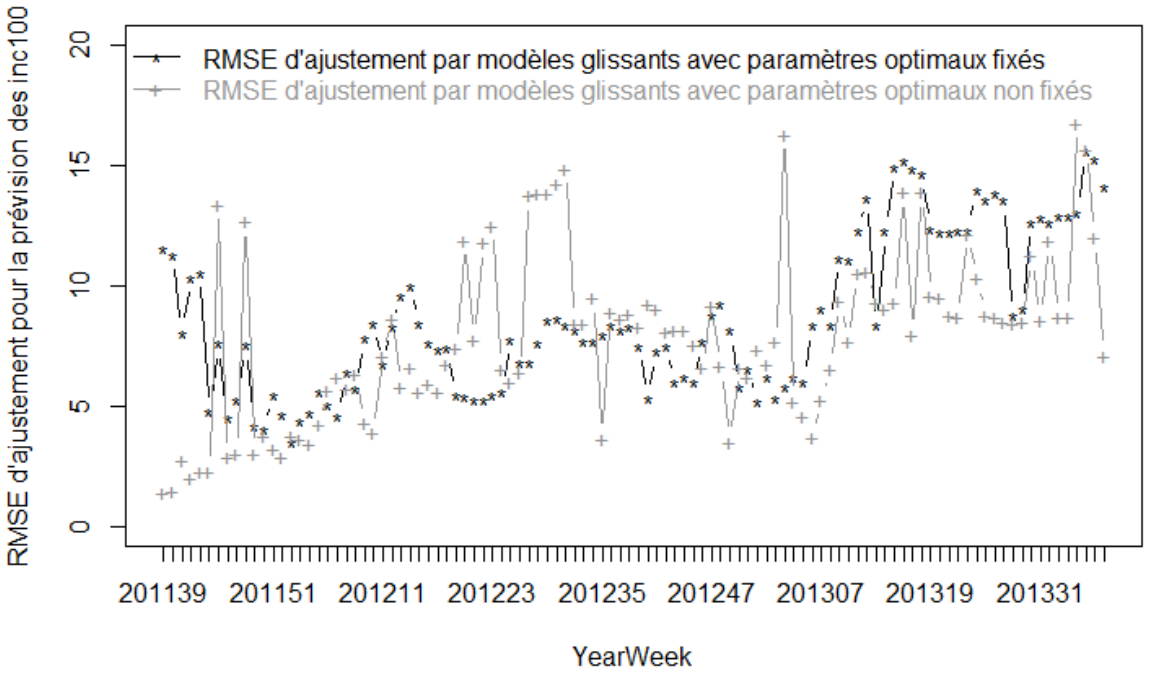
'f' = 163 semaines
'nt' = 7 valeurs latentes

RMSEP = 53,5



Que ce soit avec l'approche du « modèle unique » ou avec l'approche « des modèles glissants à paramètres fixés optimisés », les modèles PLS-Poisson appliqués sur l'ensemble de la base de données médicamenteuse ne donnent pas de meilleures prévisions des taux d'incidence des SG (les RMSEP étant supérieurs à ceux du meilleur modèle déterminé en Figure 12).

Annexe XIII. Comparaisons des erreurs d'ajustement par les approches du type « modèles glissants », entre ceux "à paramètres fixés optimisés" et ceux "à paramètres non fixés optimisés ", sur 201139-201338



A la différence des prévisions, l'ajustement par les meilleurs modèles glissants - en terme de prédictions - à paramètres fixés optimisés (approche 2) est légèrement moins bon sur 201139-201338 que par ceux des modèles où les paramètres se ré-optimisent hebdomadairement (approche 3) : RMSE moyen de 7,88 cas/100 000 pour l'approche 2, et de 8,67 cas/100 000 pour l'approche 3.

Concernant l'approche 2, chaque RMSE (en ordonnée) associé à une semaine donnée 's' (en abscisse) correspond à l'erreur quadratique moyenne d'ajustement sur les 'f' = 45 dernières semaines (et avec 'n' = 3) ayant permis la prédiction de cette semaine 's' (paramètres optimaux 'f' et 'n' déterminés pour le meilleur modèle de prédiction, cf. Annexe VII).

En revanche, pour l'approche 3, chaque RMSE correspond à l'erreur quadratique moyenne sur les 'f' dernières semaines (et avec 'n' optimal classes), où les paramètres 'f' et 'n' se ré-optimisent chaque semaine en se basant sur les 'j' = 10 pré-semaines. Les différentes valeurs prisent par 'f' et 'n' pour chaque semaine prédite sont en Annexe VI.

Diplôme : Ingénieur

Spécialité : Agronomie

Spécialisation : Statistique Appliquée

Enseignant référent : David CAUSEUR

Auteur : Cyril ESNAULT

Date de naissance : 09/09/1990

Nb pages : 35 *(10 974 mots)* **Annexe(s) :** 13

Année de soutenance : 2014

Organisme d'accueil : Réseau Sentinelles
UMR S 11 36

Adresse : 27 rue Chaligny
75012 Paris

Maître de stage : Clément TURBELIN

Titre français : Estimations intégratives des incidences régionales des syndromes grippaux en France par utilisation de données de délivrances médicamenteuses

Titre anglais : Integrative estimates of regional incidences of influenza-like illness in France by using medication sales

Résumé (français) :
Les politiques de prévention et de lutte contre les épidémies de syndromes grippaux nécessitent une surveillance épidémiologique en continue, basée sur les déclarations en temps réel de médecins généralistes bénévoles. Le calcul hebdomadaire des incidences au niveau national et régional permet d'estimer la répartition spatiale des épidémies en France. Toutefois, la stabilité de ces estimations varie au cours du temps et selon les régions. Ce travail vise donc à améliorer les estimations régionales en utilisant une source de données externe, la base des délivrances médicamenteuses. Deux axes d'études sont ici présentés : l'un consiste en la « désagrégation spatiale » d'un modèle ajusté sur les délivrances nationales, l'autre en « l'extrapolation spatiale» d'un modèle ajusté sur les délivrances en Rhône-Alpes, région choisie pour la stabilité de ses incidences. Pour cela, une procédure de présélection de classes médicamenteuses en lien avec les syndromes grippaux a été développée, par des méthodes de classification. De même, différents modèles de régression périodique (log-linéaire, log-non linéaire et PLS-poisson) et différentes approches (modèles glissants ou non, à paramètres fixés ou non fixés) ont été comparés afin de sélectionner le meilleur modèle de prédiction des incidences. La « désagrégation » et « l'extrapolation » spatiales ont alors permis de proposer de nouvelles estimations des incidences régionales, de meilleures stabilités. Enfin, un *bootstrap* a permis d'estimer un intervalle de confiance des nouvelles prédictions.

Abstract (english) :
Policies to prevent and fight against epidemics of influenza-like illness (ILI) require epidemiological surveillance continuously, based on the statements in real-time GPs volunteers. The weekly calculation of incidences at national and regional level allows us to estimate the spatial distribution of outbreaks in France. However, the stability of these estimates varies over time and across regions. This work aims to improve regional estimates using external data source, based medication sales. Two areas of study are presented here: one consists in the "spatial disaggregation" of a model fitted on national sales, the other in "spatial extrapolation" of a model fitted on sales in Rhône-Alpes, French region chosen for the stability of its incidences. For this, a preselection procedure of drug classes related to ILI was developed by classification methods. Similarly, different models of periodic regression (log-linear, log-nonlinear and PLS-Poisson) and approaches (unmoving or moving models with fixed or unfixed parameters) were compared to select the best model for predicting the incidences. The spatial "disaggregation" and "extrapolation" allowed us to propose new regional estimates of the incidences, of better stability. Finally, a bootstrap was used to estimate a confidence interval around the new predictions.

Mots-clés : Syndromes grippaux, incidences, délivrances médicamenteuses, désagrégation spatiale, extrapolation spatiale
Key Words: Influenza-like illness, incidences, medication sales, spatial disaggregation, spatial extrapolation