Santé publique France

TABAC

MAI 2024

ÉTUDES ET ENQUÊTES

PRÉVALENCES DÉPARTEMENTALES DU TABAGISME QUOTIDIEN : ESTIMATIONS SUR PETITS DOMAINES À PARTIR DES DONNÉES DU BAROMÈTRE DE SANTÉ PUBLIQUE FRANCE 2021

Résumé

Prévalences départementales du tabagisme quotidien : estimations sur petits domaines à partir des données du baromètre de santé publique France 2021

La prévalence du tabagisme quotidien est estimée au niveau national et régional depuis près de vingt ans à travers les données du Baromètre de Santé publique France. Au niveau départemental, compte tenu de la taille d'échantillon insuffisante des différentes éditions de cette enquête, jamais aucun résultat n'a été publié. Dans ce travail, afin de fournir les premières estimations départementales de la prévalence du tabagisme quotidien, la méthode d'estimation sur petits domaines de Fay-Herriot (FH) a été utilisée et comparée à la méthode directe.

Les données sur le tabagisme quotidien proviennent du Baromètre de Santé publique France de 2021. Ces données ont concerné 22 625 individus âgés de 18 à 75 ans en France hexagonale. Des données sociodémographiques (âge, sexe, diplôme, catégorie socioprofessionnelle, type de ménage, taux d'activité, revenu médian…) provenant du recensement de la population de 2019 ont été également utilisées comme variables auxiliaires prédictives du tabagisme quotidien dans le modèle de FH.

Les prévalences départementales du tabagisme quotidien estimées par la méthode directe varient entre 12,6 % et 41,4 %. Elles varient entre 17,7 % et 32,5 % pour le modèle de FH. La précision des prévalences estimées varie entre 7,4 % et 53,1 % pour la méthode directe et est inférieure à 15 % pour la méthode de FH, quel que soit le département.

Les résultats de ce travail montrent la faisabilité de produire des estimations départementales du tabagisme quotidien en utilisant les données du Baromètre de Santé publique France 2021 prévu initialement pour produire des estimations au niveau national et régional. Cependant, la sélection de variables auxiliaires bien corrélées à la variable d'intérêt est primordiale pour réduire le biais des estimations du modèle de FH.

MOTS-CLÉS : ESTIMATIONS SUR PETITS DOMAINES,
ESTIMATIONS DÉPARTEMENTALES,
TABAGISME, MÉTHODE DE FAY-HERRIOT.

Citation suggérée : Zeghnoun A, Richard JB. Prévalences départementales du tabagisme quotidien : estimations sur petits domaines à partir des données du baromètre de santé publique France 2021. Saint-Maurice : Santé publique France, 2024. 30 p. Disponible à partir de l'URL : www.santepubliquefrance.fr

ISSN : 2609-2174 - ISBN-NET : 979-10-289-0911-6 - RÉALISÉ PAR LA DIRECTION DE LA COMMUNICATION, SANTÉ PUBLIQUE FRANCE - DÉPÔT LÉGAL : MAI 2024

Abstract

Prevalence of daily smoking at departmental level in France: estimates on small areas using data from the 2021 French Public Health Barometer

The prevalence of daily smoking at the national and regional levels in France has been estimated for nearly twenty years using data from the Health Barometer survey. Results at departmental level (the French administrative division between region and municipality) have never been published due to the insufficient sample size of the various editions of this survey. This report presents results obtained using the Fay-Herriot (FH) small area estimation method compared to the direct method, in order to provide the first departmental estimates for the prevalence of daily smoking.

Data on daily smoking come from the 2021 Health Barometer conducted by Santé publique France, the national public health agency. These data concerned 22,625 individuals aged 18 to 75 years in mainland France. Sociodemographic data (age, sex, education, socioeconomic category, household type, activity rate, median income, etc.) from the 2019 population census were also used as predictive auxiliary variables for daily smoking in the FH model.

Estimations for departmental prevalences of daily smoking range from 12.6% to 41.4% using the direct method. They range from 17.7% to 32.5% using the FH model. The precision of the estimated prevalences varies from 7.4% to 53.1% for the direct method and is less than 15% for the FH method, regardless of the department.

The results of this work demonstrate the feasibility of producing departmental estimates of daily smoking using data from the 2021 Health Barometer that was originally intended to produce estimates at the national and regional levels. However, selecting auxiliary variables that correlate well with the variable of interest is crucial to reducing the bias of the FH model estimates.

KEY WORDS: SMALL AREA ESTIMATES, DEPARTMENTAL ESTIMATES, SMOKING, FAY-HERRIOT METHOD."

Auteurs

Abdelkrim Zeghnoun, Chargé d'analyse et d'expertise, Direction Appui, Traitements et Analyses de données (Data)

Jean-Baptiste Richard, Responsable de l'unité Appui à la conception, à la mise en place et à l'exploitation des enquêtes, Data

Remerciements

Nous tenons à remercier les relecteurs Abdessattar Saoudi, Raphaël Andler, Romain Guignard, Anne Pasquereau, Viet Nguyen Thanh et Yann Le Strat qui ont contribué, grâce à leurs commentaires et conseils, à l'amélioration de cette note.

Table des matières

1. INTRODUCTION

Lors de l'exploitation de données collectées via des enquêtes par sondage, les domaines ou sous-populations, sur lesquels une moyenne ou une prévalence sont estimées, sont souvent des caractéristiques sociodémographiques telles que le sexe, l'âge, les niveaux d'éducation ou des caractéristiques géographiques comme les régions ou les départements de résidence. Des difficultés peuvent alors être rencontrées dans la production d'estimations lorsque les sous-populations d'intérêt sont de taille trop petite dans l'échantillon pour obtenir une précision suffisante. On parle alors de petits domaines. En effet, les disparités démographiques du territoire sont telles qu'en France métropolitaine, la part de population résidant dans les différents départements varie de 0,1 % en Lozère à 4,0 % dans le Nord, si bien qu'en l'absence de stratification, dans une enquête de 20 000 participants menée en France métropolitaine, on observera en moyenne 200 individus par département et certains « petits » départements n'incluront qu'une vingtaine d'individus, ce qui est insuffisant pour garantir une précision suffisante des estimations.

Une solution peut être trouvée lors de la conception de la stratégie d'échantillonnage. Par exemple, pour atteindre l'objectif de publication d'estimations régionales, un sur-échantillonnage régional peut être réalisé afin de disposer d'un nombre de répondants minimum par région et garantir une précision suffisante des estimations. Toutefois, pour atteindre un objectif de publication au niveau infrarégional, par exemple départemental, l'application d'une telle stratégie implique alors des tailles d'échantillons, des coûts et des délais de collecte très importants. À titre d'exemple, des objectifs de publication d'estimations départementales ont été visés dans le cadre des enquêtes EpiCOV, Vie Quotidienne et Santé ou Vécu et Ressenti en matière de Sécurité, et ont nécessité des échantillons tirés au sort de plus de 200 000 individus ou ménages.

Une autre solution est de s'appuyer sur des méthodes d'estimations sur petits domaines, qui font référence à un ensemble de méthodes permettant d'estimer un paramètre d'intérêt (une moyenne, une prévalence…) dans un domaine plus petit en ce qui concerne la taille d'échantillon que celui pour lequel une enquête a été conçue, en combinant des données d'enquête et des données auxiliaires issues de sources externes comme le recensement.

Dans le cadre du Baromètre de Santé publique France, la méthode de sélection des individus repose sur la génération aléatoire de numéros de téléphone fixes et mobiles et n'offre pas de possibilité de surreprésentation selon le lieu de résidence[1] (sauf à n'utiliser que des numéros de téléphone fixes, les numéros de téléphone mobiles ne comprenant par exemple aucun indicatif géographique). Aussi, lorsque des objectifs de publication d'estimations au niveau régional sont visés, la taille de l'échantillon est calculée afin d'interroger un minimum de 1 000 répondants par région de France métropolitaine, exception faite de la Corse. De nombreuses valorisations régionales ont ainsi été produites, notamment dans le champ des addictions (Beck *et al.*, 2008; Beck *et al.*, 2013; Pasquereau *et al.*, 2022; Andler *et al.*, 2023). Concernant plus spécifiquement la consommation de tabac, première cause de mortalité prématurée évitable en France, des Bulletins de santé publique et des Points épidémiologiques régionaux[2] ont été réalisés en 2019, 2021 et 2023, combinant différentes sources de données relatives aux habitudes de consommation et à la morbimortalité associée au tabac. Dans cet état des

[1] À compter de 2024, la méthode d'échantillonnage, reposant sur un tirage d'individus dans la base Fidéli, permettra de réaliser de telles surreprésentations.

[2] https://www.santepubliquefrance.fr/regions/grand-est/documents/bulletin-regional/2019/bulletin-de-sante-publique-tabac-dans-le-grand-est.-janvier-2019 ; Publication des premiers bulletins de santé publique dédiés au tabac pour chaque région de France (santepubliquefrance.fr)
https://www.santepubliquefrance.fr/regions/grand-est/documents/bulletin-regional/2021/bulletin-de-sante-publique-tabac-dans-le-grand-est.-fevrier-2021
https://www.santepubliquefrance.fr/regions/grand-est/documents/bulletin-regional/2023/tabac-dans-la-region-grand-est.-donnees-regionales-du-barometre-2021

lieux détaillé a été soulignée la grande variabilité entre femmes et hommes, entre générations, entre les régions et pour certaines d'entre elles, entre les départements les composant. Alors que les données de mortalité sont disponibles à l'échelon départemental, celles concernant les habitudes de vie ne sont généralement pas diffusées à ce niveau géographique, exception faite des données publiées récemment à partir de la vague 2 de l'enquête EpiCov[3]. Pourtant, ces informations permettraient de contribuer à mieux définir, cibler, voire évaluer les actions locales de santé publique en matière de prévention de l'entrée dans le tabagisme, d'incitation à l'arrêt du tabac et d'accompagnement des fumeurs souhaitant arrêter, notamment dans le cadre des programmes régionaux de réduction du tabagisme.

L'objectif de ce travail est d'explorer l'utilisation des méthodes d'estimation sur petits domaines pour produire des estimations à l'échelon départemental. Nous présenterons dans ces travaux des estimations des prévalences du tabagisme quotidien en comparant la méthode directe et la méthode de Fay et Herriot (Fay et Herriot, 1979).

[3] https://www.epicov.fr/publications/

2. DONNÉES UTILISÉES

Les données du Baromètre de Santé publique France ont été utilisées pour estimer la prévalence départementale du tabagisme quotidien. L'édition de 2021 a permis d'enquêter 22 625 individus âgés de 18 à 75 ans en France métropolitaine. Des extensions ont par ailleurs été réalisées dans les DROM (départements et régions d'outre-mer) hors Mayotte, avec pour objectif de produire des estimations régionales/départementales, et ne sont pas incluses dans cette analyse. Une description détaillée du protocole de cette enquête et de la méthode de génération aléatoire de numéros de téléphone utilisée est présentée dans un rapport méthodologique (Soullier *et al.*, 2022). Pour ce travail, nous utilisons à titre d'illustration les données sur le tabagisme quotidien. Un individu est considéré comme « fumeur quotidien » s'il déclare fumer tous les jours ou déclare une fréquence de consommation quotidienne pour au moins un type de tabac, chicha comprise.

Le modèle de Fay et Herriot, décrit ci-après, nécessite des variables auxiliaires à l'échelle des petits domaines, c'est-à-dire les départements dans ce travail. Ces variables ont été obtenues au niveau départemental sur le site de l'Insee (Institut national de la statistique et des études économiques). Il s'agit principalement de variables sociodémographiques (âge, sexe, diplôme, catégorie socioprofessionnelle, type de ménage, taux d'activité, revenu médian…) déclinées au niveau départemental (proportion d'hommes et de femmes par département…) et provenant des données du recensement de la population de 2019[4]. Ces variables ont été sélectionnées car elles sont corrélées au tabagisme quotidien (Pasquereau *et al.*, 2022).

[4] https://www.insee.fr/fr/statistiques

3. MÉTHODES UTILISÉES

3.1 Méthode directe

L'estimateur de Horvitz et Thompson a été utilisé pour produire l'estimation de la prévalence départementale du tabagisme quotidien. Seuls les individus ayant participé à l'enquête dans chacun des départements sont impliqués dans cette estimation et aucune information auxiliaire n'est utilisée. Cette méthode pose problème lorsque la taille de l'échantillon du département est petite, la variance de l'estimation étant inversement proportionnelle à cette taille.

Des pondérations redressées par calage sur marges ont été utilisées pour produire les estimations départementales. Les variables de calage, issues des enquêtes emploi réalisées par l'Insee, sont la région de résidence (12 modalités, avec un regroupement des régions Provence-Alpes-Côte d'Azur (PACA) et Corse) croisée avec le sexe et l'âge en classes décennales, la taille de l'unité urbaine, le niveau de diplôme et la taille du foyer. Notons ici qu'un calage sur des marges départementales n'a pas été employé du fait d'échantillons départementaux de tailles insuffisantes qui auraient conduit à une importante dispersion des poids.

3.2 Modèle de Fay et Herriot (FH)

Le modèle proposé par Fay et Herriot est un modèle linéaire à effets aléatoires. C'est un modèle composite avec deux sous modèles. Le premier appelé « modèle d'échantillonnage » s'écrit comme suit :

$$\hat{\theta}_d^{Dir} = \theta_d + e_d, \qquad d = 1, \dots, D \qquad (1)$$

où $\hat{\theta}_d^{Dir}$ est l'estimation directe non biaisée du paramètre d'intérêt θ_d dans le petit domaine d, par exemple l'estimateur de la prévalence du tabagisme quotidien dans la population des adultes du département d et e_d est l'erreur due à l'échantillonnage également dans le petit domaine d.

Le second modèle appelé « modèle synthétique » est un modèle linéaire mettant en relation le paramètre d'intérêt dans la population avec les p variables auxiliaires x_d observées au niveau du domaine *d*.

$$\theta_d = x_d^T \beta + u_d, \qquad d = 1, \dots, D \qquad (2)$$

Avec u_d des effets aléatoires. La combinaison de ces deux modèles définit le modèle de Fay et Herriot et s'écrit sous la forme :

$$\hat{\theta}_d^{Dir} = x_d^T \beta + u_d + e_d, \qquad d = 1, \dots, D \qquad (3)$$

u_d et e_d sont supposés indépendants et normalement distribués avec une moyenne égale à zéro et des variances égales à σ_u^2 et σ_e^2, respectivement. L'effet aléatoire u_d représente l'effet des petits domaines et permet de tenir compte de la variabilité inter-domaines non expliquée par les variables auxiliaires x_d. β est le vecteur des coefficients de régression à estimer. Les variances σ_u^2 et σ_e^2 doivent également être estimées afin d'obtenir l'estimateur final du

modèle de FH. Cet estimateur peut s'exprimer comme une moyenne pondérée des estimateurs direct et synthétique :

$$\hat{\theta}_d^{FH} = \hat{\gamma}_d \, \hat{\theta}_d^{Dir} + (1 - \hat{\gamma}_d) \, x_d^T \hat{\beta}, \qquad d = 1, \dots, D \qquad (4)$$

Le paramètre $\hat{\gamma}_d = \hat{\sigma}_u^2 / (\hat{\sigma}_u^2 + \hat{\sigma}_e^2)$ est spécifique à chaque petit domaine et permet de donner plus de poids à l'estimateur direct du domaine lorsque la variance de l'erreur d'échantillonnage dans ce domaine est petite et inversement, lorsque celle-ci est grande, l'estimateur direct est instable et moins de poids lui est donné. Pour les domaines où aucune donnée n'est disponible, l'estimateur de FH se limite à la partie synthétique.

Le modèle de FH construit est un modèle de régression dans lequel la variable dépendante est l'estimation directe de la prévalence départementale du tabagisme quotidien et les variables prédictives au niveau départemental sont :

- l'âge (% de personnes âgées entre 18 et 24 ans, 25 et 34 ans, 35 et 44 ans, 45 et 54 ans, 55 et 64 ans, 65 et 75 ans) ;
- le sexe (% d'hommes et de femmes) ;
- le diplôme en 5 catégories (% de personnes sans diplômes ou peu diplômés, avec BEPC ou brevet, avec CAP ou BEP, avec BAC, avec un diplôme de l'enseignement supérieur) ;
- la catégorie socioprofessionnelle en 8 catégories (Agriculteurs exploitants, Artisans-commerçants-chefs entreprise, Cadres et professions intellectuelles supérieures, Professions intermédiaires, Employés, Ouvriers, Retraités, Autres personnes sans activité professionnelle) ;
- le type de ménage en 5 catégories (Ménages dont la famille principale est un couple avec enfant, Ménages dont la famille principale est un couple sans enfant, Ménages dont la famille principale est une famille monoparentale, Ménages d'une personne, Autres ménages sans famille) ;
- le taux d'activité par tranches d'âge en 3 catégories (Taux d'activité des 15 à 24 ans, 25 à 54 ans et 55 à 64 ans) ;
- et la médiane du niveau de vie issue du recensement 2019 de la population au niveau du département.

La variabilité inter-département non prise en compte par ces variables a été modélisée par un effet aléatoire sur l'ordonnée à l'origine. Ainsi, les variables auxiliaires du recensement sont utilisées comme des effets fixes et la variable indicatrice du département est utilisée comme effet aléatoire. Un modèle complet a été testé dans un premier temps. Une analyse préalable de l'autocorrélation spatiale a par ailleurs été menée, afin d'évaluer la pertinence d'en tenir compte dans le modèle utilisé. Les paramètres de ce modèle ainsi que l'analyse des résidus sont présentés en annexe. Enfin, dans ce travail, compte tenu de leur proximité géographique et des faibles échantillons, les deux départements de Corse ont été regroupés.

Le package sae (Molina *et al.*, 2015 ; Molina et Marhuenda, 2015) implémenté sous le logiciel R (R journal, 2015) a été utilisé pour estimer les prévalences départementales.

3.3 Cohérence interne des résultats : le « Benchmarking »

La plupart des enquêtes sont conçues pour fournir une estimation directe fiable au niveau national, voire régional. C'est le cas du *Baromètre de Santé publique France 2021* mené auprès d'un grand échantillon permettant des estimations précises à l'échelle nationale et régionale (la taille de l'échantillon a été calculée de façon à inclure 1 000 participants minimum par région).

La méthode de FH nécessite de vérifier la cohérence interne des estimations sur petits domaines. En effet, l'agrégation de ces estimations sur l'ensemble des domaines peut différer de l'estimation directe, jugée fiable, obtenue au niveau supérieur. Par exemple, l'agrégation des prévalences du tabagisme quotidien obtenues par une méthode de FH sur l'ensemble des départements d'Île-de-France n'est pas nécessairement égale à l'estimation de la prévalence du tabagisme quotidien sur la région Île-de-France obtenue par la méthode directe. Ainsi, pour garder une cohérence entre les estimations sur les petits domaines et les estimations directes du niveau supérieur, un ajustement (*benchmarking*) des estimations obtenues par la méthode de FH est nécessaire (Molina, Marın and Rao, 2019).

Si on note :

$Y_d, d \in A$, le total de la variable d'intérêt dans le petit domaine d de la région A

$\hat{Y}_A^{DIR}$, l'estimateur direct non biaisé de la variable d'intérêt disponible au niveau de la région A

$\tilde{Y}_d, d \in A$, l'estimateur FH de ce total dans le petit domaine d de la région A

Alors nous avons : $\sum_{d \in A} \tilde{Y}_d \neq \hat{Y}_A^{DIR}$

Le *benchmarking* assure la cohérence interne des résultats, c'est-à-dire : $\sum_{d \in A} \tilde{Y}_d^{bench} = \hat{Y}_A^{DIR}$. Pour cela l'estimateur de FH $\tilde{Y}_d, d \in A$ est ajusté par le facteur $f_A^{bench} = \frac{\hat{Y}_A^{DIR}}{\sum_{d \in A} \tilde{Y}_d}$. Nous obtenons ainsi l'estimateur ajusté :

$$\tilde{Y}_d^{bench} = \tilde{Y}_d \times f_A^{bench} = \tilde{Y}_d \times \frac{\hat{Y}_A^{DIR}}{\sum_{d \in A} \tilde{Y}_d} \qquad (5)$$

4. RÉSULTATS

4.1 Résultats obtenus avec le modèle de Fay-Herriot

L'objectif est d'illustrer l'utilisation du modèle de FH pour l'estimation des prévalences départementales du tabagisme quotidien. L'estimation de la prévalence à l'échelle du département est d'abord obtenue en utilisant l'estimateur direct de Horvitz-Thompson, ensuite par le modèle de FH. Les paramètres estimés du modèle final retenu ainsi que l'analyse de la normalité des résidus et des effets aléatoires sont présentés en annexe.

Les prévalences départementales du tabagisme quotidien estimées par la méthode directe varient entre 12,6 % (La Lozère, n=29) et 41,4 % (Les Pyrénées-Orientales, n=143). Elles sont plus resserrées vers la moyenne, variant entre 17,7 % (Les Yvelines, n=509) et 32,5 % (La Corse, n=80) pour le modèle de FH (**Tableau 1**, **Figure 2**, **Carte 1** et **2**). Quant à la précision des prévalences estimées (**Tableau 1** et **Figure 3**), elle varie entre 7,4 % (Le Nord, n=871) et 53,1 % (La Lozère, n=29) pour la méthode directe et est inférieure à 15 % pour la méthode de FH quel que soit le département (CV max=13,1% pour la Lozère).

Tableau 1. Prévalences départementales du tabagisme quotidien estimées par les méthodes directe et de Fay-Herriot à partir des données du Baromètre de Santé publique France 2021

			Méthode directe		Modèle Fay-Herriot	
Région	**Département**	**n**	**Prévalence (IC95 %)**	**CV*(%)**	**Prévalence (IC95 %)**	**CV(%)**
Auvergne-Rhône-Alpes	Rhône (69)	641	28,1 (23,4 - 32,7)	8,4 %	26,7 (22,9 - 30,4)	7,2 %
	Isère (38)	510	23,2 (18,2 - 28,2)	11,0 %	22,9 (19,3 - 26,5)	8,0 %
	Haute-Savoie (74)	297	28,3 (21,9 - 34,8)	11,6 %	26,4 (21,9 - 31)	8,7 %
	Loire (42)	272	21,5 (14,9 - 28)	15,6 %	24,0 (19,6 - 28,5)	9,5 %
	Puy-de-Dôme (63)	258	23,9 (17 - 30,8)	14,8 %	22,9 (18,9 - 26,9)	9,0 %
	Ain (1)	249	19,0 (12,2 - 25,7)	18,2 %	22,3 (18,2 - 26,5)	9,4 %
	Savoie (73)	172	22,5 (13,5 - 31,5)	20,5 %	25,2 (20,6 - 29,8)	9,3 %
	Drôme (26)	165	25,7 (16,7 - 34,7)	17,9 %	25,7 (21,5 - 29,9)	8,3 %
	Ardèche (7)	115	25,2 (15,5 - 34,8)	19,6 %	25,5 (21,2 - 29,7)	8,5 %
	Allier (3)	114	30,1 (18,5 - 41,7)	19,6 %	25,4 (20,5 - 30,4)	10,0 %
	Haute-Loire (43)	89	21,2 (10,5 - 31,9)	25,8 %	23,5 (18,6 - 28,3)	10,5 %
	Cantal (15)	42	33,8 (11,4 - 56,2)	33,8 %	23,3 (18,3 - 28,3)	10,9 %
Bourgogne-Franche-Comté	Saône-et-Loire (71)	215	30,5 (21,9 - 39)	14,3 %	25,8 (21,6 - 30)	8,3 %
	Côte-d'Or (21)	165	21,8 (14 - 29,6)	18,3 %	22,3 (18 - 26,7)	9,9 %
	Doubs (25)	163	21,5 (12,8 - 30,3)	20,8 %	22,9 (18,9 - 27)	9,0 %
	Yonne (89)	123	36,5 (25,5 - 47,5)	15,4 %	26,8 (22,4 - 31,1)	8,3 %
	Haute-Saône (70)	106	24,9 (15,3 - 34,5)	19,7 %	25,1 (20,9 - 29,3)	8,5 %
	Jura (39)	100	27,5 (16,6 - 38,4)	20,2 %	25,2 (20,8 - 29,5)	8,8 %
	Nièvre (58)	90	23,9 (13,1 - 34,7)	23,1 %	26,0 (20,7 - 31,2)	10,3 %
	Territoire de Belfort (90)	48	22,2 (5,9 - 38,4)	37,3 %	25,2 (20,6 - 29,7)	9,2 %

			Méthode directe		Modèle Fay-Herriot	
Région	**Département**	**n**	**Prévalence (IC95 %)**	**CV*(%)**	**Prévalence (IC95 %)**	**CV(%)**
Bretagne	Ille-et-Vilaine (35)	451	27,2 (21,6 - 32,8)	10,5 %	24,4 (20,5 - 28,3)	8,2 %
	Finistère (29)	343	22,6 (16,8 - 28,3)	13,0 %	22,5 (18,4 - 26,6)	9,3 %
	Morbihan (56)	314	29,4 (22,5 - 36,3)	11,9 %	25,2 (21,1 - 29,2)	8,3 %
	Côtes-d'Armor (22)	213	22,3 (15,5 - 29,1)	15,5 %	22,4 (18,3 - 26,5)	9,4 %
Centre-Val-de-Loire	Indre-et-Loire (37)	238	17,4 (10,8 - 23,9)	19,3 %	21,9 (18 - 25,8)	9,1 %
	Loiret (45)	223	13,2 (7,4 - 18,9)	22,2 %	20,2 (16,5 - 24)	9,4 %
	Loir-et-Cher (41)	125	31 (19,7 - 42,2)	18,6 %	26,3 (22 - 30,6)	8,3 %
	Eure-et-Loir (28)	125	31,5 (19,9 - 43)	18,8 %	23,9 (19,5 - 28,3)	9,3 %
	Cher (18)	94	24,3 (13,4 - 35,2)	22,8 %	24,9 (20,1 - 29,7)	9,8 %
	Indre (36)	73	26,7 (14,3 - 39,1)	23,7 %	25,9 (21 - 30,8)	9,7 %
Corse	Corse (20)	80	32,1 (19,1 - 45,1)	20,7 %	32,5 (26,7 - 38,3)	9,1 %
Grand Est	Bas-Rhin (67)	418	29,6 (23,3 - 35,8)	10,7 %	26,8 (22,6 - 31,1)	8,1 %
	Moselle (57)	335	28,5 (22,3 - 34,7)	11,1 %	26,7 (22,5 - 31)	8,1 %
	Haut-Rhin (68)	247	24,1 (16,6 - 31,5)	15,8 %	25,5 (21,2 - 29,8)	8,7 %
	Meurthe-et-Moselle (54)	234	20,8 (14 - 27,6)	16,6 %	23,6 (19,4 - 27,8)	9,1 %
	Marne (51)	189	25,4 (17,3 - 33,5)	16,2 %	23,1 (18,9 - 27,3)	9,3 %
	Vosges (88)	156	29,7 (19,9 - 39,6)	16,9 %	27,4 (23,1 - 31,8)	8,0 %
	Ardennes (8)	86	30,6 (17,6 - 43,5)	21,6 %	27,8 (23,2 - 32,5)	8,5 %
	Aube (10)	80	19,2 (8,4 - 30,1)	28,8 %	24,9 (20,5 - 29,2)	8,9 %
	Meuse (55)	71	38,8 (23,4 - 54,2)	20,3 %	26,9 (22,2 - 31,6)	8,9 %
	Haute-Marne (52)	67	25,2 (12,8 - 37,5)	25,0 %	26,5 (21,8 - 31,2)	9,0 %
Hauts-de-France	Nord (59)	871	27,7 (23,7 - 31,7)	7,4 %	26,2 (23 - 29,5)	6,4 %
	Pas-de-Calais (62)	440	24,1 (18,9 - 29,2)	10,9 %	25,5 (21,7 - 29,3)	7,6 %
	Oise (60)	266	33,1 (25 - 41,2)	12,5 %	26 (21,7 - 30,2)	8,3 %
	Somme (80)	175	16,8 (10,4 - 23,1)	19,4 %	22,8 (18,8 - 26,8)	9,0 %
	Aisne (2)	175	23,2 (14,3 - 32,1)	19,5 %	25,6 (21,3 - 29,9)	8,6 %
Île-de-France	Paris (75)	829	24,2 (20,1 - 28,2)	8,5 %	24,2 (20,1 - 28,2)	8,5 %
	Hauts-de-Seine (92)	614	19,7 (15,5 - 24)	11,0 %	20,1 (16,4 - 23,8)	9,3 %
	Yvelines (78)	509	17,9 (13,3 - 22,5)	13,2 %	17,7 (13,7 - 21,7)	11,6 %
	Essonne (91)	426	19,7 (14,8 - 24,5)	12,6 %	21,3 (17,6 - 25)	8,8 %
	Val-de-Marne (94)	426	26,8 (20,9 - 32,8)	11,3 %	25,1 (21,2 - 28,9)	7,9 %
	Seine-et-Marne (77)	425	24,2 (18,6 - 29,8)	11,8 %	22,4 (18,4 - 26,3)	9,0 %
	Seine-Saint-Denis (93)	393	25,3 (19,4 - 31,2)	11,8 %	27,3 (22,6 - 32)	8,8 %
	Val-d'Oise (95)	366	20,7 (14,5 - 26,9)	15,2 %	21,5 (16,9 - 26,1)	10,9 %
Normandie	Seine-Maritime (76)	434	22,8 (17,5 - 28,2)	11,9 %	24,2 (20,6 - 27,8)	7,7 %
	Calvados (14)	264	26,9 (19,7 - 34,1)	13,6 %	24,3 (20,1 - 28,5)	8,9 %
	Eure (27)	210	28,1 (19,9 - 36,4)	14,9 %	24,9 (20,8 - 29,1)	8,5 %
	Manche (50)	169	23,1 (13,9 - 32,3)	20,3 %	26,1 (21,4 - 30,8)	9,2 %
	Orne (61)	105	29,4 (18,7 - 40,2)	18,6 %	27,8 (23,2 - 32,3)	8,4 %

			Méthode directe		Modèle Fay-Herriot	
Région	**Département**	**n**	**Prévalence (IC95 %)**	**CV*(%)**	**Prévalence (IC95 %)**	**CV(%)**
Nouvelle-Aquitaine	Gironde (33)	599	26,9 (22,2 - 31,5)	8,9 %	26 (22,4 - 29,6)	7,1 %
	Pyrénées-Atlantiques (64)	244	24,4 (17,5 - 31,4)	14,5 %	26 (21,8 - 30,1)	8,1 %
	Charente-Maritime (17)	243	22,4 (15,9 - 29)	14,9 %	24,1 (19,7 - 28,6)	9,3 %
	Landes (40)	153	17,5 (10,4 - 24,5)	20,6 %	23,7 (19,4 - 28)	9,3 %
	Vienne (86)	151	24,7 (15,2 - 34,2)	19,7 %	24,6 (19,9 - 29,2)	9,6 %
	Dordogne (24)	144	33,2 (23 - 43,5)	15,8 %	29,8 (25,1 - 34,6)	8,2 %
	Deux-Sèvres (79)	137	29,6 (19,8 - 39,5)	16,9 %	25,5 (20,9 - 30,1)	9,3 %
	Charente (16)	128	23,1 (14 - 32,1)	20,0 %	27,1 (22,6 - 31,6)	8,5 %
	Haute-Vienne (87)	127	28,2 (18 - 38,5)	18,5 %	25,1 (20,6 - 29,6)	9,2 %
	Lot-et-Garonne (47)	118	20,6 (10,8 - 30,4)	24,2 %	27,3 (23 - 31,5)	8,0 %
	Corrèze (19)	97	27,5 (16,3 - 38,8)	20,8 %	25,4 (21,2 - 29,7)	8,5 %
	Creuse (23)	39	27,5 (8,3 - 46,8)	35,7 %	26,4 (21 - 31,8)	10,4 %
Occitanie	Haute-Garonne (31)	533	26,4 (21,2 - 31,6)	10,0 %	26 (22 - 29,9)	7,7 %
	Hérault (34)	408	30 (24 - 35,9)	10,1 %	28,4 (24,3 - 32,4)	7,4 %
	Gard (30)	249	26,3 (18,8 - 33,7)	14,4 %	28,1 (23,8 - 32,4)	7,8 %
	Tarn (81)	151	33 (22,2 - 43,8)	16,7 %	27,7 (23,5 - 31,9)	7,8 %
	Pyrénées-Orientales (66)	143	41,4 (30 - 52,8)	14,1 %	32,4 (27,3 - 37,6)	8,1 %
	Aude (11)	119	23,3 (12,7 - 34)	23,3 %	27,6 (22,8 - 32,3)	8,8 %
	Hautes-Pyrénées (65)	90	25,6 (14,5 - 36,7)	22,2 %	26,2 (21,5 - 30,8)	9,1 %
	Aveyron (12)	86	22 (12 - 32,1)	23,3 %	25,7 (21,3 - 30,1)	8,7 %
	Tarn-et-Garonne (82)	85	36,4 (21,7 - 51,1)	20,6 %	28,7 (23,8 - 33,6)	8,7 %
	Gers (32)	74	25,8 (11,8 - 39,8)	27,7 %	25,8 (21 - 30,6)	9,5 %
	Lot (46)	65	28,6 (15 - 42,2)	24,3 %	26,5 (21,1 - 31,9)	10,4 %
	Ariège (9)	57	23,4 (8,4 - 38,3)	32,7 %	28,7 (23,5 - 34)	9,3 %
	Lozère (48)	29	12,6 (-0,5 - 25,7)	53,1 %	21,3 (15,8 - 26,7)	13,1 %
Pays de la Loire	Loire-Atlantique (44)	597	18,9 (14,9 - 23)	11,0 %	20,6 (17,3 - 23,9)	8,1 %
	Maine-et-Loire (49)	304	24,9 (17,5 - 32,3)	15,1 %	23,2 (19 - 27,5)	9,4 %
	Vendée (85)	261	24,2 (17,5 - 30,9)	14,1 %	22 (16,9 - 27)	11,7 %
	Sarthe (72)	182	29,5 (20,5 - 38,6)	15,6 %	25,2 (20,9 - 29,4)	8,6 %
	Mayenne (53)	106	19,5 (10,1 - 29)	24,7 %	22,5 (17,9 - 27)	10,3 %
Provence-Alpes-Côte d'Azur	Bouches-du-Rhône (13)	688	28,8 (24,4 - 33,2)	7,7 %	28,6 (25 - 32,1)	6,4 %
	Var (83)	357	32,7 (26,2 - 39,1)	10,1 %	29,8 (25,5 - 34,1)	7,3 %
	Alpes-Maritimes (6)	356	29,4 (22,9 - 35,8)	11,2 %	28 (23,8 - 32,2)	7,7 %
	Vaucluse (84)	194	25,2 (17,4 - 33)	15,8 %	27,9 (23,5 - 32,3)	8,0 %
	Alpes-de-Haute-Provence (4)	63	30,2 (15,4 - 45,1)	25,1 %	28 (23 - 33,1)	9,1 %
	Hautes-Alpes (5)	55	16,6 (5,1 - 28,1)	35,3 %	26 (20,3 - 31,7)	11,1 %

*CV : Coefficient de variation

Les prévalences estimées par la méthode directe et le modèle de FH sont relativement proches (**Tableau 1** et **Figure 2**) lorsque la taille de l'échantillon du département est grande. Dans ce cas, le modèle FH donne en effet plus de poids à la méthode directe, relativement plus précise. La moyenne des écarts en valeur absolue entre les deux estimations est de 2,9 %. La différence est cependant plus importante pour certains départements comme

La Meuse et Le Cantal pour lesquels la taille des échantillons est petite. La corrélation reste cependant assez bonne avec un coefficient de corrélation égal à 0,69 (**Figure 4**).

Les prévalences départementales estimées par le modèle de FH sont moins dispersées que celles estimées par la méthode directe. Cela est particulièrement visible sur la **Figure 3** qui compare la précision obtenue par les 2 méthodes pour chaque département. Ainsi, les CV obtenus par le modèle de FH sont systématiquement plus petits et peu dispersés par rapport à ceux de la méthode directe quelle que soit la taille de l'échantillon par département. Ce qui n'est pas le cas pour la méthode directe où on peut observer l'augmentation des CV lorsque la taille de l'échantillon par département diminue. On observe également que l'écart entre les précisions obtenues par le modèle de FH et l'estimation directe s'accentue lorsque la taille d'échantillon par département diminue.

Figure 1. Prévalences départementales du tabagisme quotidien estimées par les méthodes directe et de Fay-Herriot à partir des données du Baromètre de Santé publique France 2021

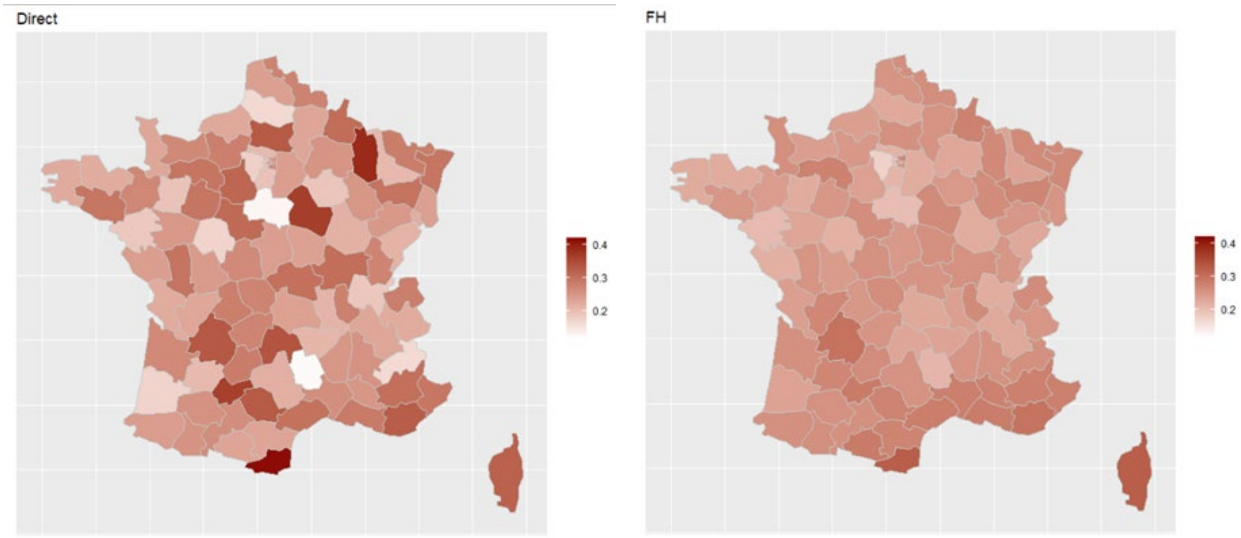


Figure 2. Prévalences départementales du tabagisme quotidien estimées par les méthodes directe et de Fay-Herriot à partir des données du Baromètre de Santé publique France 2021 (départements triés par tailles d'échantillon décroissants)

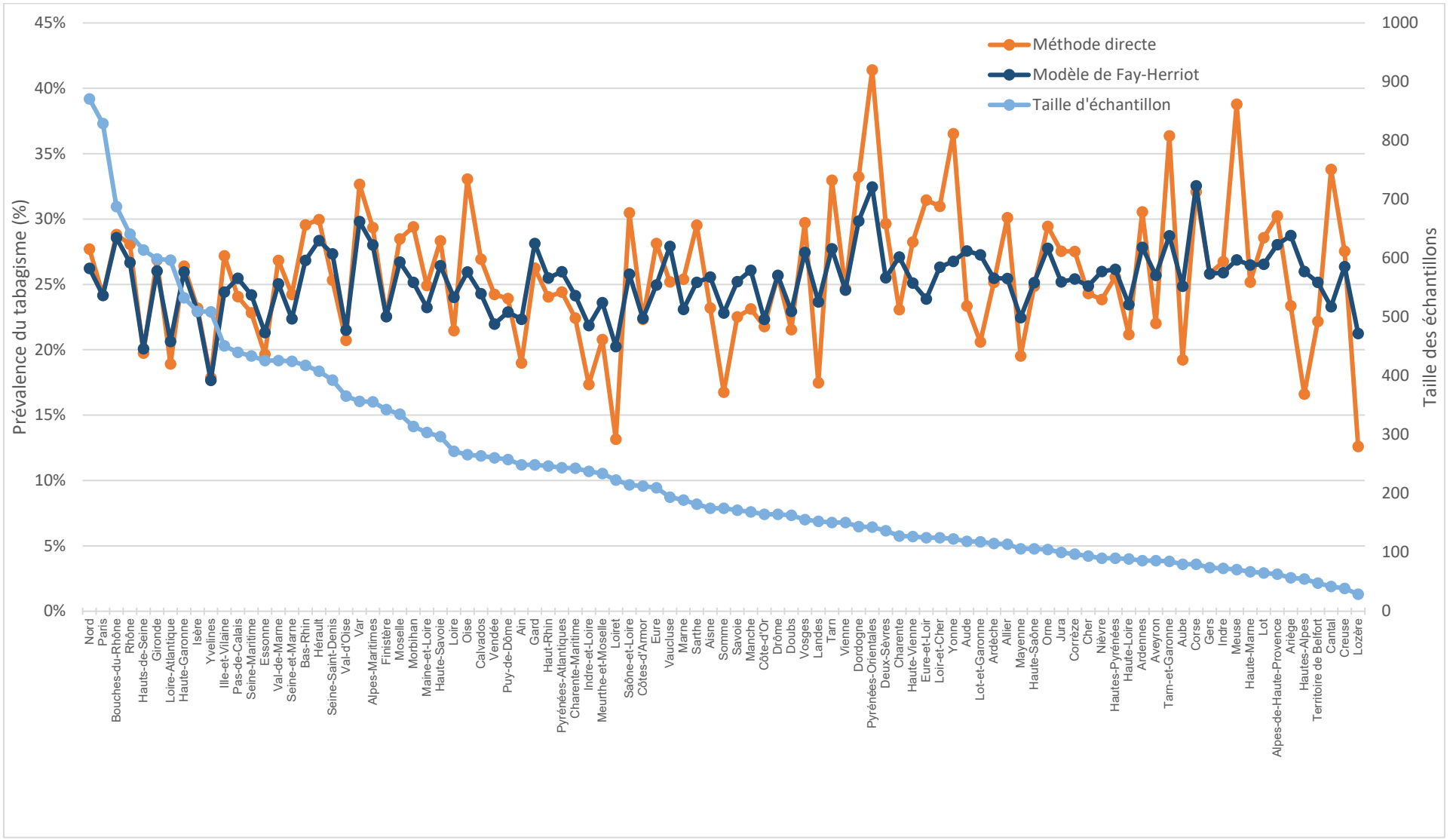


Figure 3. Coefficient de variation (CV) des prévalences départementales du tabagisme quotidien estimées par les méthodes directe et de Fay-Herriot à partir des données du Baromètre de Santé publique France 2021 (départements triés par tailles d'échantillon décroissants)

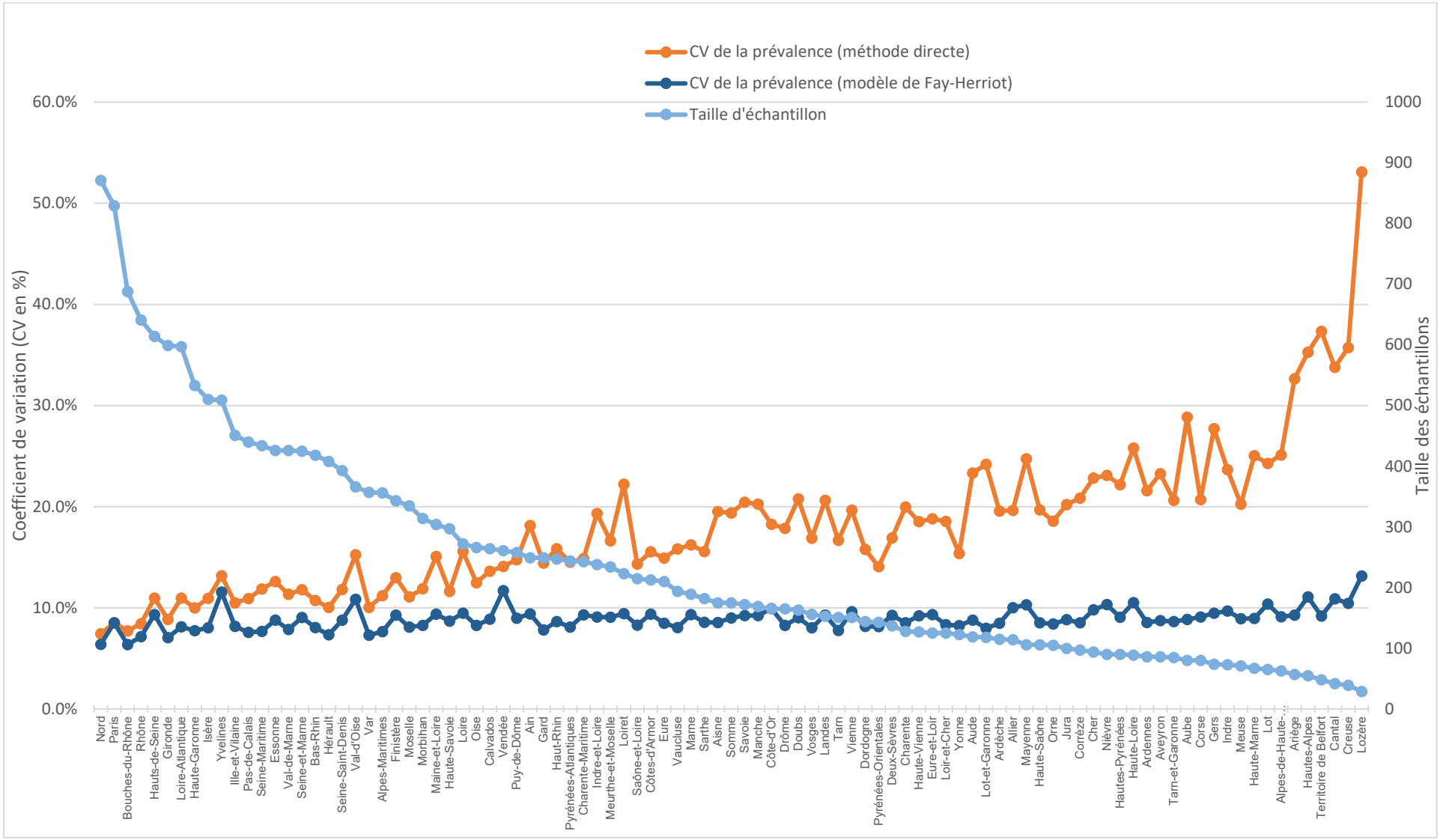
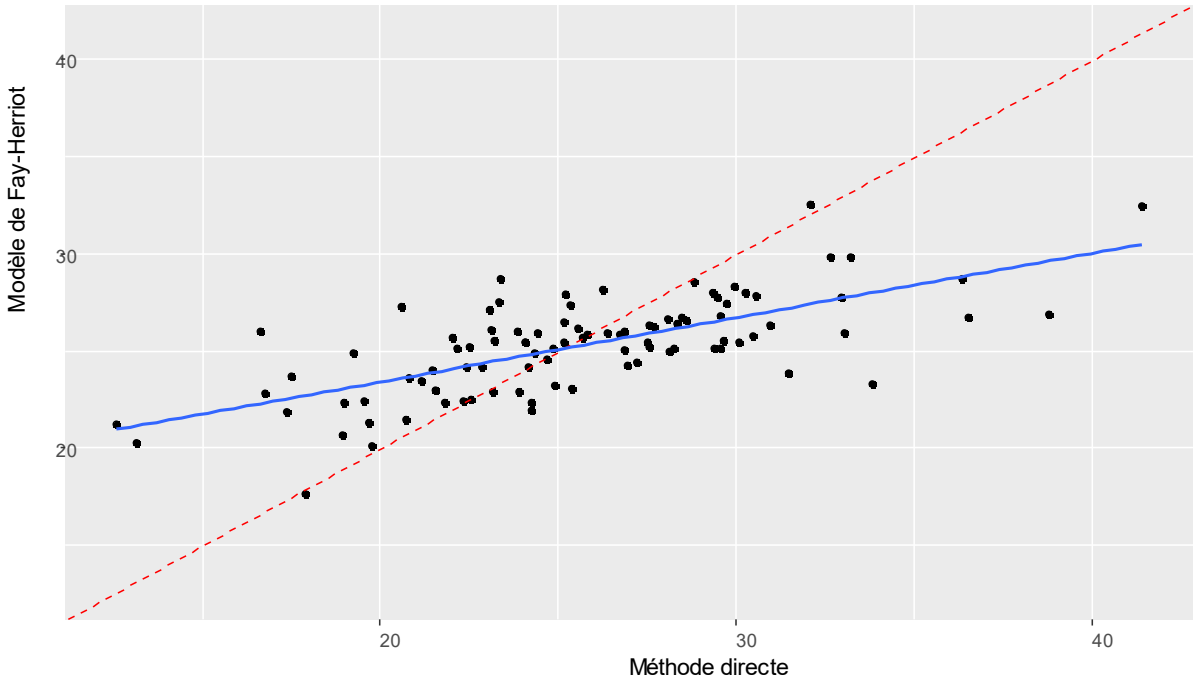


Figure 4. Corrélation entre les prévalences départementales obtenues par les méthodes directes et de Fay-Herriot à partir des données du Baromètre de Santé publique France 2021

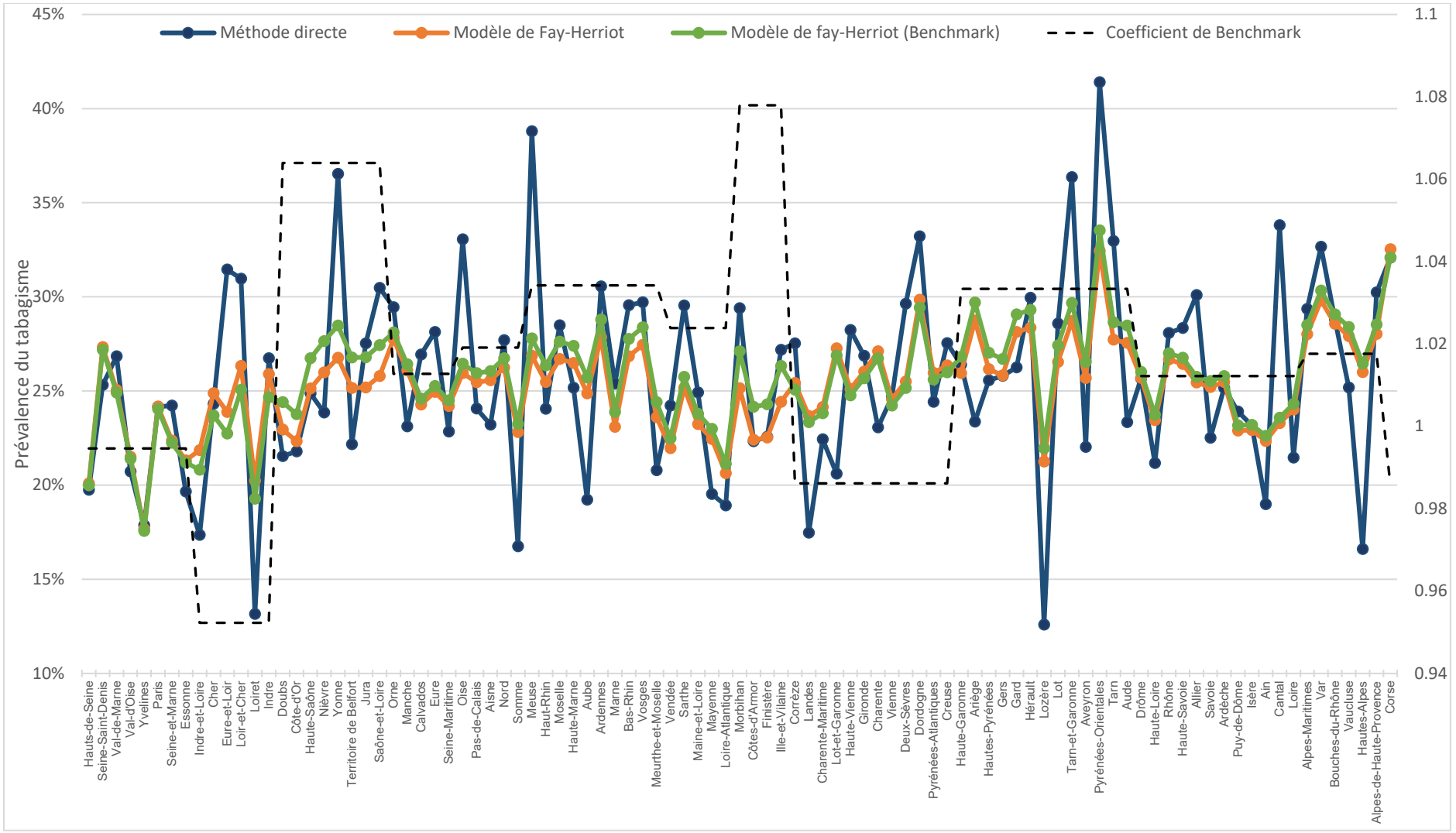


4.2 Mise en cohérence interne des résultats : le *benchmarking*

L'estimateur obtenu avec la méthode directe respecte la propriété de l'additivité. En d'autres termes, le total des estimations directes départementales au niveau d'une région est égal à l'estimation directe de cette région et le total des estimations régionales directes est égal au total de l'estimation directe nationale. Les prévalences estimées par la méthode de FH ne respectent pas cette propriété d'où l'intérêt de mettre en cohérence les résultats nationaux/régionaux et départementaux en utilisant le *benchmarking*. Celui-ci est réalisé sur les estimations directes régionales qui sont précises compte tenu de l'effectif important par région de l'édition 2021 du Baromètre de Santé publique France.

Le *benchmarking* est réalisé en utilisant l'équation 5 qui permet de respecter les estimations des totaux de fumeurs à chaque niveau (national, régional et départemental). Les estimations départementales de la prévalence du tabagisme quotidien obtenues par le modèle de FH avant et après *benchmarking* sont présentées dans la **Figure 5** et le tableau en annexe. On peut observer que lorsque l'estimation régionale de FH (total des estimations départementales de FH pour chaque région) sous-estime l'estimation régionale directe (coefficient de *benchmarking* >1) les estimations départementales de FH sont augmentées à hauteur du coefficient de *benchmarking* et inversement.

Figure 5. Prévalences départementales du tabagisme quotidien par la méthode directe et FH avant et après *benchmarking* (départements triés par région)



5. CONCLUSION

La prévalence du tabagisme quotidien est estimée au niveau national et régional depuis près de vingt ans à travers les données du Baromètre de Santé publique France. Au niveau infrarégional, compte tenu de la taille d'échantillon des différentes éditions du Baromètre, aucun résultat n'est disponible. Dans ce travail, afin de fournir les premières estimations départementales de la prévalence du tabagisme quotidien, la méthode d'estimation sur petits domaines de Fay-Herriot a été utilisée et comparée à la méthode directe. Le recours à cette méthode 1) est facilité par la disponibilité des données auxiliaires au niveau des départements ; 2) permet d'améliorer les estimations directes qui ne sont précises que lorsque la taille de l'échantillon dans les départements est suffisamment grande ; 3) est également motivé par le souhait de produire des estimations sur petits domaines sans pour autant augmenter de façon significative la taille de l'échantillon initialement prévu pour l'enquête nationale.

Les prévalences départementales du tabagisme quotidien estimées par la méthode directe varient entre 12,6 % et 41,4 %. Les valeurs extrêmes concernent surtout les départements avec une petite taille d'échantillon (la Lozère, le Loiret, la Meuse, les Pyrénées-Orientales…). En termes de précision, les résultats sont acceptables (CV < 16,5%)[5] dans près de la moitié des départements. Si on considère un seuil de 20 %, plus élevé mais qui reste acceptable pour un petit domaine, près de 77 % des estimations directes ont un coefficient de variation inférieur à 20 %.

La méthode de FH emprunte de l'information auxiliaire à d'autres départements et permet ainsi un gain de précision pour les petits départements par l'intermédiaire d'un modèle statistique. Elle donne ainsi de meilleures précisions par rapport à la méthode directe pour tous les départements. Les prévalences prédites par le modèle de FH sont également assez bien corrélées à celles obtenues avec la méthode directe. En effet, l'estimateur de FH est un estimateur composite qui donne plus de poids à l'estimateur direct du département lorsque la variance de l'erreur d'échantillonnage dans ce département est petite et inversement lorsque l'estimateur direct est instable moins de poids lui est affecté. Il utilise ainsi l'estimateur direct lorsqu'il est raisonnablement fiable. Cette caractéristique est particulièrement intéressante puisque l'estimateur direct est relativement sans biais lorsque la taille de l'échantillon du domaine est suffisamment grande.

Malgré l'amélioration de la précision, l'estimateur de FH est basé sur un modèle et présente le risque d'introduire un biais, en particulier s'il est mal spécifié. Il a ainsi tendance à resserrer les prédictions des prévalences extrêmes vers la moyenne. Ainsi, les grandes prévalences peuvent être sous-estimées et les petites prévalences peuvent être surestimées (**Figure 6** et **Figure 6**). Ce resserrement des prévalences est limité et aurait été plus important si le modèle était purement synthétique (et non composite). L'appréciation du biais introduit par le modèle de FH est difficile en l'absence de résultats fiables sur le tabagisme au niveau départemental issus d'enquêtes externes. Il est à noter cependant que les prévalences extrêmes estimées par la méthode directe peuvent être biaisées particulièrement lorsque l'échantillon obtenu dans les départements concernés est de petite taille. Le modèle de FH, s'il est bien spécifié, peut ainsi améliorer l'estimation des prévalences dans ces petits départements.

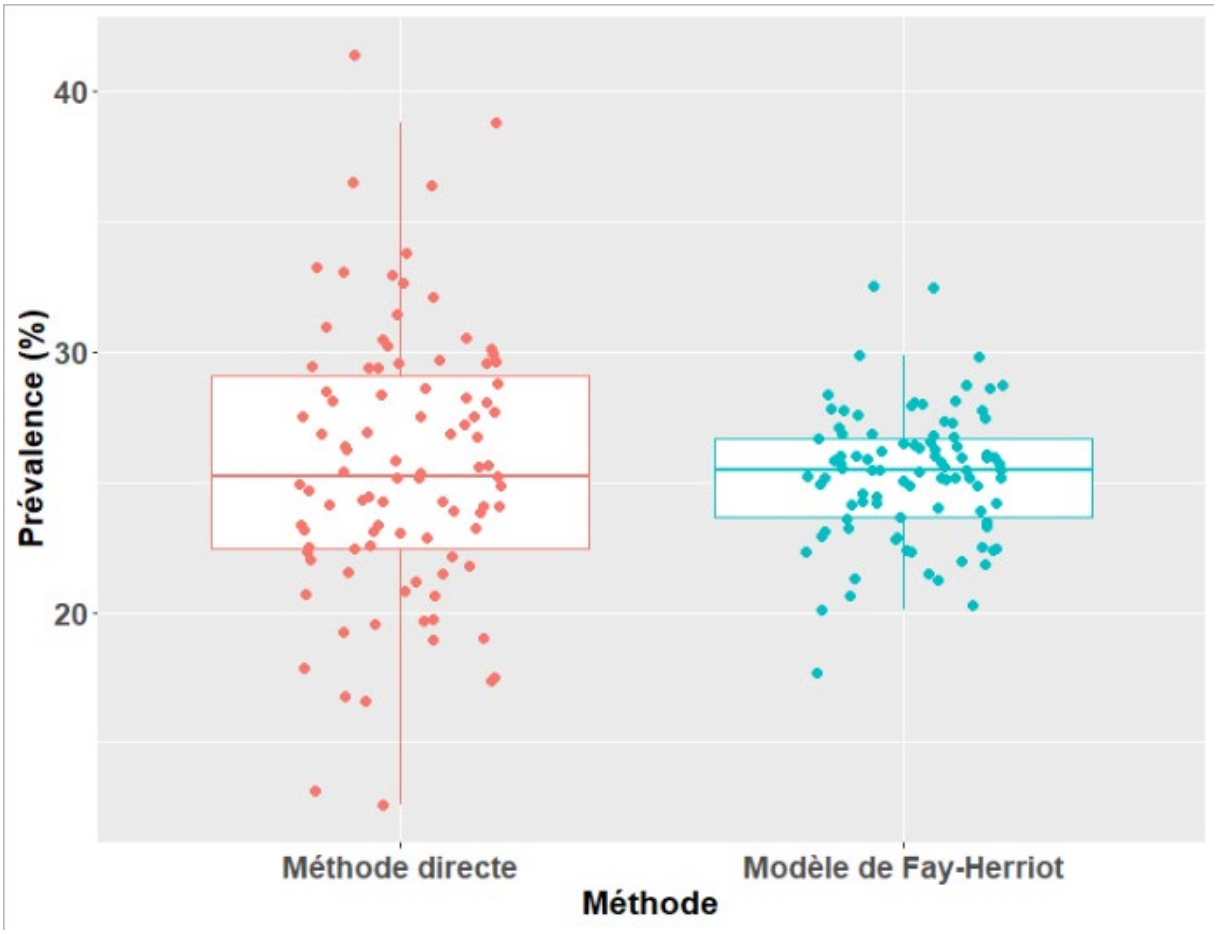
Le modèle de FH par sa nature composite entre l'estimation directe et l'estimation synthétique, semble être une piste pour l'exploitation des futurs Baromètres de Santé publique France, pour estimer les prévalences départementales du tabagisme quotidien ou d'une autre variable d'intérêt mesurée sur l'ensemble de l'échantillon (et non sur un sous-échantillon). Il tire profit de l'estimation non biaisée de la méthode directe (lorsque n est suffisamment grand) et la

[5] https://www.statcan.gc.ca/fr/programmes-statistiques/document/3226_D7_T9_V8#a11

variance faible de l'estimation synthétique. L'utilisation de ce modèle deviendra encore plus intéressante du fait de tailles d'échantillon plus conséquentes des futurs Baromètres de Santé publique France (au moins 20 000 individus) pour chaque édition, avec une stratification régionale qui permettra de disposer d'effectifs plus importants dans les départements les moins peuplés.

Au final, les résultats de ce travail montrent la faisabilité de produire des estimations départementales du tabagisme quotidien en utilisant les données du Baromètre de Santé publique France 2021 prévu initialement pour produire des estimations au niveau national et régional. Cependant, malgré la taille suffisante de l'échantillon national et des échantillons régionaux, plusieurs départements avaient une petite taille d'échantillon qui impliquait de donner plus de poids à l'estimateur synthétique. Ainsi, dans la construction du modèle de FH, la sélection de variables auxiliaires bien corrélées à la variable d'intérêt est primordiale pour réduire le biais des estimations. D'autre part, le modèle de FH utilisé dans ce travail ne mobilisait que des variables sociodémographiques disponibles au niveau du département et potentiellement corrélées au comportement tabagique. L'utilisation d'autres variables auxiliaires directement liées à la consommation tabagique au niveau départemental (par exemple, données de vente de tabac, de remboursement de traitement de substitution…) mériterait sans doute d'être explorée afin d'améliorer les prédictions du modèle. Enfin, une autre méthode permettant de produire des estimations départementales précises est la combinaison d'enquêtes Baromètres successives, approche particulièrement intéressante pour les indicateurs dont l'évolution dans le temps est modérée (Thomas et Wannell, 2009).

Figure 7. Distribution des prévalences départementales obtenues par les méthodes directes et de Fay-Herriot à partir des données du Baromètre de Santé publique France 2021

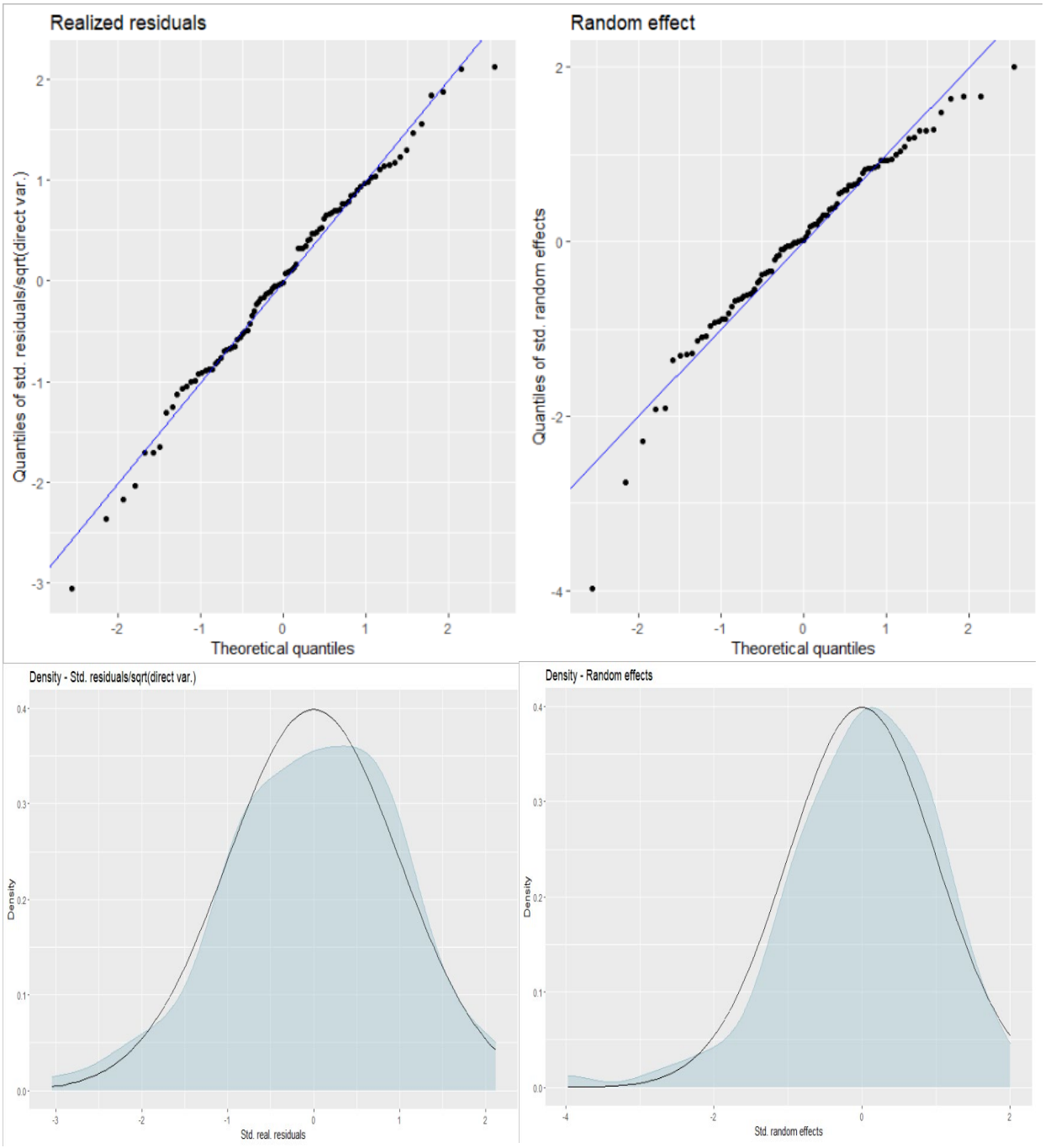


6. ANNEXE

6.1 Modèle de Fay-Herriot utilisé pour estimer les prévalences départementales

```
Call:
 fh(fixed = tab_quo ~ p_AGE1824 + p_AGE2534 + p_AGE3544 + p_AGE4554 +
    p_AGE5564 + CSP2 + CSP7 + type_menage1 + type_menage4 + taux_activite2,
vardir = "var", combined_data = fhbaro21data, domains = "DEPNUM",
interval = c(1e-04, 7e-04), MSE = TRUE)

Out-of-sample domains:  0
In-sample domains:  95

Variance and MSE estimation:
Variance estimation method:  reml
Estimated variance component(s):  0.0004195025
MSE method:  prasad-rao

Coefficients:
               coefficients std.error t.value   p.value
(Intercept)        -1.52186   1.40057 -1.0866 0.2772115
p_AGE1824           3.09291   1.46264  2.1146 0.0344628 *
p_AGE2534           4.33014   1.91366  2.2627 0.0236514 *
p_AGE3544           3.06045   1.95448  1.5659 0.1173800
p_AGE4554           3.87210   2.00604  1.9302 0.0535792 .
p_AGE5564           3.29947   1.98714  1.6604 0.0968322 .
CSP2                2.71952   1.35371  2.0089 0.0445440 *
CSP7                1.74343   0.94963  1.8359 0.0663715 .
type_menage1       -1.26786   0.76246 -1.6628 0.0963443 .
type_menage4       -0.85317   0.53827 -1.5850 0.1129591
taux_activite2     -1.26577   0.38058 -3.3259 0.0008814 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Explanatory measures:
   loglike       AIC       BIC       KIC        R2     AdjR2
1 162.0714 -300.1428 -269.4963 -288.1428 0.1356369 0.3316901

Residual diagnostics:
                         Skewness Kurtosis Shapiro_W   Shapiro_p
Standardized_Residuals -0.3457141 3.146590 0.9873414 0.498743789
Random_effects         -0.8820526 4.827892 0.9590258 0.004640387
```

6.2 Analyse de la normalité des résidus et des effets aléatoires



6.3 Exemple de programme utilisant le package emdi sous R

Pour faciliter la compréhension et l'utilisation de la méthodologie, cet exemple d'application sera publié sur le site github.com et pourra être accessible à tout public. Les données utilisées dans cet exemple (données de prévalences du tabagisme et données auxiliaires issues du site de l'Insee, agrégées au niveau département) seront également publiées. Aucune donnée individuelle directement ou indirectement identifiante ne sera publiée sur ce site.

 # Tester l'autocorrélation spatiale
L'autocorrélation spatiale est une corrélation entre les mesures géographiquement voisines d'un paramètre. La valeur de l'indice de Geary varie entre 0 et 2 : 1 signifiant aucune #autocorrélation spatiale. Une valeur < 1 (resp. >1) signifie une autocorrélation spatiale positive (resp. négative). L'indice de Geary est plus sensible à l'autocorrélation spatiale locale.
L'indice de Moran est une mesure globale de l'autocorrélation spatiale. Ses valeurs s'étendent entre -1 (dispersion parfaite) et 1 (corrélation parfaite). Une valeur nulle indique un modèle spatial parfaitement aléatoire.

 # Calcul de la matrice de proximité

```
library(maptools)
library(spdep)
library(sp)

depOM1<-st_read(system.file("shapes/New_dep_fr.shp", package="maptools")[1])
depOM=depOM1[!(depOM1$NEW_REG %in % c("01","02","03","04","06")),]
rm(depOM1)

depOM <- depOM[order(depOM$INSEE_DEP),]

data_spatial=merge(depOM,fhbaro21data,by.x="INSEE_DEP",by.y="DEPARTEMENT", all.x = F)

w <- poly2nb(data_spatial, row.names=data_spatial$INSEE_DEP)
wm <-nb2mat(w,   style='W',   zero.policy=TRUE)

# wn est la matrice de proximité et data_spatial contient la prévalence du tabagisme par département

spatialcor.tests(direct = data_spatial$tab_quo,corMatrix = wm)
```

 # Modèle FH retenu

```
library(emdi)  # Package emdi for SAE

fh_std <- fh(fixed = tab_quo ~ p_AGE1824 + p_AGE2534 + p_AGE3544 + p_AGE4554 + p_AGE5564 +
          CSP2+CSP7+type_menage1+type_menage4+taux_activite2,
vardir = "var", combined_data = fhbaro21data,
domains = "DEPARTEMENT", MSE=TRUE,interval = c(0.0001,0.0007))

summary(fh_std)
plot(fh_std)
compare_plot(fh_std, CV = TRUE, label = "no_title")
compare(fh_std)
```

 # Sauvegarde des résultats et plot

```
fhestimate=as.data.frame(estimators(fh_std, MSE = TRUE,CV=TRUE))
head(estimators(fh_std, MSE = TRUE))
summary(data.frame(estimators(fh_std, MSE = TRUE)))
fhestimate1=data.frame(fh_std$model$gamma,fh_std$model$random_effects,cbind(fh_std$model$fitted),fh_std$model$real_re
siduals,fh_std$model$std_real_residuals)
fhestimate.vf=merge(fhestimate,fhestimate1,by.x="Domain",by.y="Domain")
fhestimate.vf=merge(fhestimate.vf,taille_ech75,by.x="Domain",by.y="departement")
fhestimate.vf=merge(fhestimate.vf,fhbaro21data[,c("DEPNUM","REGION","DEPARTEMENT","nom_region","nom_departement"
)],by.x="Domain",by.y="DEPNUM")
rm(fhestimate,fhestimate1)
write.csv2(fhestimate.vf,"fhestimate_vf.csv")
```

 # plots

```
library(ggplot2)
a = min(fhestimate.vf$Direct*100, fhestimate.vf$FH*100)
b = max(fhestimate.vf$Direct*100, fhestimate.vf$FH*100)
qplot(fhestimate.vf$Direct*100, fhestimate.vf$FH*100, xlim = c(a, b), ylim = c(a, b), xlab = "Méthode directe",ylab = "Modèle de
Fay-Herriot", geom = "point") + geom_abline(color = "violet", intercept = 0, slope = 1)
```

 # boxplot

```
library(reshape2)
boxplotdata=fhestimate.vf[,c(1,2,5)]
boxplotdata=melt(boxplotdata,id="Domain")
p <- ggplot(boxplotdata, aes(x=variable, y=value*100,color=variable)) + geom_boxplot(outlier.shape = NA )
```

```
p + geom_jitter(shape=16, cex=2,position=position_jitter(0.2))+labs(x="Méthode", y = "Prévalence (%)")+ theme(axis.text.x =
element_text(face="bold", size=14), axis.text.y = element_text(face="bold",
size=14),axis.title=element_text(size=16,face="bold"))+ scale_x_discrete(labels=c("Direct" = "Méthode directe", "FH" = "Modèle
de Fay-Herriot"))+theme(legend.position = "none")

# carte par département
depOM1=as(depOM, 'Spatial')
# sans homogénéisation de l'échelle
map_plot(object = fh_std, MSE = TRUE, map_obj = depOM1, map_dom_id = "INSEE_DEP")
summary(data.frame(estimators(fh_std, MSE = TRUE)))
# ou avec échelle
map_plot(object = fh_std, MSE = TRUE, map_obj = depOM1, map_dom_id = "INSEE_DEP",
scale_points = list(Direct = list(ind = c(0.12, 0.42), MSE = c(0.00028, 0.13)), FH = list(ind = c(0.12, 0.42), MSE = c(0.00028,
0.13))))
map_plot(object = fh_std, MSE = TRUE, map_obj = depOM1, map_dom_id = "INSEE_DEP",scale_points = list(Direct = list(ind
= c(0.12, 0.42)), FH = list(ind = c(0.12, 0.42) )))
```

6.4 Résultats de Fay-Herriot après benchmarking

Région	Nom du département	Méthode directe	Modèle de Fay-Herriot	Modèle de Fay-Herriot (Benchmark)
Auvergne-Rhône-Alpes	Rhône	28,1 %	26,7 %	27,0 %
	Isère	23,2 %	22,9 %	23,2 %
	Haute-Savoie	28,3 %	26,4 %	26,8 %
	Loire	21,5 %	24,0 %	24,3 %
	Puy-de-Dôme	23,9 %	22,9 %	23,2 %
	Ain	19,0 %	22,3 %	22,6 %
	Savoie	22,5 %	25,2 %	25,5 %
	Drôme	25,7 %	25,7 %	26,0 %
	Ardèche	25,2 %	25,5 %	25,8 %
	Allier	30,1 %	25,4 %	25,8 %
	Haute-Loire	21,2 %	23,5 %	23,7 %
	Cantal	33,8 %	23,3 %	23,6 %
Bourgogne-Franche-Comté	Saône-et-Loire	30,5 %	25,8 %	27,4 %
	Côte-d'Or	21,8 %	22,3 %	23,8 %
	Doubs	21,5 %	22,9 %	24,4 %
	Yonne	36,5 %	26,8 %	28,5 %
	Haute-Saône	24,9 %	25,1 %	26,7 %
	Jura	27,5 %	25,2 %	26,8 %
	Nièvre	23,9 %	26,0 %	27,7 %
	Territoire de Belfort	22,2 %	25,2 %	26,8 %
Bretagne	Ille-et-Vilaine	27,2 %	24,4 %	26,3 %
	Finistère	22,6 %	22,5 %	24,3 %
	Morbihan	29,4 %	25,2 %	27,1 %
	Côtes-d'Armor	22,3 %	22,4 %	24,2 %
Centre-Val de Loire	Indre-et-Loire	17,4 %	21,9 %	20,8 %
	Loiret	13,2 %	20,2 %	19,3 %
	Eure-et-Loir	31,5 %	23,9 %	22,7 %
	Loir-et-Cher	31,0 %	26,3 %	25,1 %
	Cher	24,3 %	24,9 %	23,7 %
	Indre	26,7 %	25,9 %	24,7 %
Corse	Corse	32,1 %	32,5 %	32,1 %
Grand Est	Bas-Rhin	29,6 %	26,8 %	27,8 %
	Moselle	28,5 %	26,7 %	27,6 %
	Haut-Rhin	24,1 %	25,5 %	26,4 %
	Meurthe-et-Moselle	20,8 %	23,6 %	24,4 %
	Marne	25,4 %	23,1 %	23,9 %
	Vosges	29,7 %	27,4 %	28,4 %
	Ardennes	30,6 %	27,8 %	28,8 %
	Aube	19,2 %	24,9 %	25,7 %
	Meuse	38,8 %	26,9 %	27,8 %
	Haute-Marne	25,2 %	26,5 %	27,4 %

Région	Nom du département	Méthode directe	Modèle de Fay-Herriot	Modèle de Fay-Herriot (Benchmark)
Hauts-de-France	Nord	27,7 %	26,2 %	26,7 %
	Pas-de-Calais	24,1 %	25,5 %	26,0 %
	Oise	33,1 %	26,0 %	26,4 %
	Aisne	23,2 %	25,6 %	26,1 %
	Somme	16,8 %	22,8 %	23,3 %
Île-de-France	Paris	24,2 %	24,2 %	24,0 %
	Hauts-de-Seine	19,7 %	20,1 %	20,0 %
	Yvelines	17,9 %	17,7 %	17,6 %
	Essonne	19,7 %	21,3 %	21,2 %
	Val-de-Marne	26,8 %	25,1 %	24,9 %
	Seine-et-Marne	24,2 %	22,4 %	22,2 %
	Seine-Saint-Denis	25,3 %	27,3 %	27,2 %
	Val-d'Oise	20,7 %	21,5 %	21,4 %
Normandie	Seine-Maritime	22,8 %	24,2 %	24,5 %
	Calvados	26,9 %	24,3 %	24,6 %
	Eure	28,1 %	24,9 %	25,3 %
	Manche	23,1 %	26,1 %	26,4 %
	Orne	29,4 %	27,8 %	28,1 %
Nouvelle-Aquitaine	Gironde	26,9 %	26,0 %	25,7 %
	Pyrénées-Atlantiques	24,4 %	26,0 %	25,6 %
	Charente-Maritime	22,4 %	24,1 %	23,8 %
	Landes	17,5 %	23,7 %	23,3 %
	Vienne	24,7 %	24,6 %	24,2 %
	Dordogne	33,2 %	29,8 %	29,4 %
	Deux-Sèvres	29,6 %	25,5 %	25,2 %
	Charente	23,1 %	27,1 %	26,7 %
	Haute-Vienne	28,2 %	25,1 %	24,8 %
	Lot-et-Garonne	20,6 %	27,3 %	26,9 %
	Corrèze	27,5 %	25,4 %	25,1 %
	Creuse	27,5 %	26,4 %	26,0 %
Occitanie	Haute-Garonne	26,4 %	26,0 %	26,8 %
	Hérault	30,0 %	28,4 %	29,3 %
	Gard	26,3 %	28,1 %	29,1 %
	Tarn	33,0 %	27,7 %	28,7 %
	Pyrénées-Orientales	41,4 %	32,4 %	33,5 %
	Aude	23,3 %	27,6 %	28,5 %
	Hautes-Pyrénées	25,6 %	26,2 %	27,0 %
	Aveyron	22,0 %	25,7 %	26,5 %
	Tarn-et-Garonne	36,4 %	28,7 %	29,7 %
	Gers	25,8 %	25,8 %	26,7 %
	Lot	28,6 %	26,5 %	27,4 %
	Ariège	23,4 %	28,7 %	29,7 %
	Lozère	12,6 %	21,3 %	22,0 %

Région	Nom du département	Méthode directe	Modèle de Fay-Herriot	Modèle de Fay-Herriot (Benchmark)
Pays de la Loire	Loire-Atlantique	18,9 %	20,6 %	21,1 %
	Maine-et-Loire	24,9 %	23,2 %	23,8 %
	Vendée	24,2 %	22,0 %	22,5 %
	Sarthe	29,5 %	25,2 %	25,8 %
	Mayenne	19,5 %	22,5 %	23,0 %
Provence-Alpes-Côte d'Azur	Bouches-du-Rhône	28,8 %	28,6 %	29,1 %
	Var	32,7 %	29,8 %	30,3 %
	Alpes-Maritimes	29,4 %	28,0 %	28,5 %
	Vaucluse	25,2 %	27,9 %	28,4 %
	Alpes-de-Haute-Provence	30,2 %	28,0 %	28,5 %
	Hautes-Alpes	16,6 %	26,0 %	26,5 %

7. QUELQUES RÉFÉRENCES

Soullier N, Richard JB, Gautier A. Baromètre de Santé publique France 2021. Méthode. Saint-Maurice : Santé publique France, 2022 : 17 p (https://www.santepubliquefrance.fr/docs/barometre-de-sante-publique-france-2021.-methode)

Pasquereau A, Andler R, Guignard R *et al.* Prévalence nationale et régionale du tabagisme en France en 2021 parmi les 18-75 ans, d'après le Baromètre de Santé publique France. BEH 26, 470-480 (2022), (http://beh.santepubliquefrance.fr/beh/2022/26/pdf/2022_26.pdf)

Enquête emploi en continu 2016. https://www.insee.fr/fr/metadonnees/source/operation/s1415/documentation-methodologique.

Fay, R and Herriot, R (1979). Estimates of income for small places: an application of James–Stein procedures to census data. Journal of the American Statistical Association 74, 269–277.

Molina I and Marhuenda Y. R package sae: Methodology (2015). https://cran.r-project.org/web/packages/sae/vignettes/sae_methodology.pdf.

Schenker N and Raghunathan TE. Combining information from multiple surveys to enhance estimation of measures of health. Statist. Med. 2007; 26:1802–1811.

Thomas T and Wannell B. Combining cycles of the Canadian Community Health Survey. Component of Statistics Canada Catalogue no. 82-003-X. Health Reports. February 2009.

Molina I and Marhuenda Y. sae: An R Package for Small Area Estimation. The R journal, 2015, https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf

Molina I, Marın JM and Rao JNK. Small Area Estimation. https://www.eustat.eus/sem19_curso_areas_pequenas_i.pdf

Gonzalez-Manteiga, W, Lombarda, MJ, Molina, I, Morales, D and Santamara, L. (2007), Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model, Computational Statistics and Data Analysis, 51, 2720-2733.

Harter R, Vaish A, Sukasih A *et al.* À Practical Guide to Small Area Estimation illustrated Using the Ohio Medicaid Assessment Survey. JSM 2019, Survey Research Methods Section.

Andler, R, G. Quatremère, A Gautier, V Nguyen Thanh and F Beck. 2023. "Consommation d'alcool : part d'adultes dépassant les repères de consommation à moindre risque à partir des données du Baromètre de Santé publique France 2021". Bull Epidemiol Hebd 11: 178-86.

Beck, F, R Guignard, C Léon and J.B. Richard. 2013. Atlas des usages de substances psychoactives 2010. Analyses régionales du Baromètre santé de l'Inpes. Saint-Denis: Inpes

Beck, F, S Legleye, O Le Nezet and S Spilka. 2008. Atlas régional des consommations d'alcool 2005 : données INPES/OFDT. Saint-Denis: Inpes

Pasquereau, Anne, R Andler, R Guignard, A Gautier, Noémie Soullier, J. B. Richard, *et al.* 2022. "Prévalence nationale et régionale du tabagisme en France en 2021 parmi les 18-75 ans, d'après le Baromètre de Santé publique France". Bull Épidémiol Hebd. 26: 470-80.

Pascal Ardilly. Panorama des principales méthodes d'estimation sur les petits domaines. Documents de travail numéro M0602 paru le 1er septembre 2006 (https://www.insee.fr/fr/statistiques/1380679).

Ann-Kristin Kreutzmann, Sören Pannier, Natalia Rojas-Perilla, Timo Schmid, Matthias Templ, Nikos Tzavidis. The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators. Journal of Statistical Software. Vol. 91 (2019). Issue 7 (https://www.jstatsoft.org/article/view/v091i07).

Sylvia Harmening, Ann-Kristin Kreutzmann, S¨oren Schmidt, Nicola Salvati, Timo Schmid. À Framework for Producing Small Area Estimates Based on Area-Level Models in R (https://cran.r-project.org/web/packages/emdi/vignettes/vignette_fh.pdf).

Steven Thomas et Brenda Wannell. Combiner les cycles de l'Enquête sur la santé dans les collectivités canadiennes (février 2009). Statistique Canada, no 82-003-XPF au catalogue • Rapports sur la santé, vol. 20, no 1, mars 2009 (https://www150.statcan.gc.ca/n1/fr/pub/82-003-x/82-003-x2009001-fra.pdf?st=Hz1iW4tn).