Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Département fédéral de l'intérieur DFI
Office fédéral de la statistique OFS

Estimation du taux d'activité au niveau communal dans le cadre du relevé structurel

Etude de la combinaison des méthodes Small Area Estimation et du pooling

EXPERIMENTAL STATISTICS

Neuchâtel, 2018

Éditeur:	Office fédéral de la statistique (OFS)	**Concept de mise en page:**	Section DIAM
Renseignements:	info.pop@bfs.admin.ch, tél. +41 58 467 25 25	**Téléchargement:**	www.statistique.ch
Rédaction:	Section METH, OFS	**Copyright:**	OFS, Neuchâtel 2018
Contenu:	Marie Dupraz, Anne Massiani, Daniel Kilchmann OFS		La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée
Domaine:	00 Bases statistiques des généralités		
Langue du texte original:	Français		
Traduction:	Services linguistiques de l'OFS		

ESPACE DE L'EUROPE 10 CH-2010 NEUCHÂTEL WWW.STATISTIQUE.CH

Table des matières

1 Introduction

Le but des travaux décrits dans ce document est d'estimer le taux d'activité au niveau communal dans le cadre du relevé structurel (RS). Comme la taille des échantillons communaux est souvent trop faible pour se baser sur l'estimateur par la régression généralisée (GREG) habituellement utilisé dans cette enquête, l'OFS a envisagé la possibilité de recourir à des méthodes de type "Small Area Estimation" (SAE). Ces méthodes se basent sur une modélisation de la variable d'intérêt en fonction d'informations auxiliaires qui doivent être connues pour toute la population.

Dans ce projet, les performances des estimateurs ont été étudiées sous une approche dite "design-based", dans laquelle la précision des résultats dépend de deux composantes, le biais et la variance. Le recours au modèle et aux informations auxiliaires a pour effet de produire des estimations SAE en général beaucoup plus stables en termes de variance que les estimateurs "traditionnels" tels que le GREG. En revanche, tandis que les estimateurs "traditionnels' sont en principe pas ou peu biaisés, un risque de biais est possible avec les méthodes SAE si le modèle n'est pas suffisamment performant. La validation du modèle est donc un point crucial qui doit faire l'objet d'une attention particulière.

Dans une première étape, un mandat a été confié à l'Universidad Carlos III de Madrid. Le cadre de ce mandat était restreint à l'année 2012 du relevé structurel. Il a permis de démontrer la possibilité d'obtenir des estimations fiables du taux d'activité et de sa précision pour les communes disposant d'un échantillon d'au moins 100 personnes. Pour ces communes, le gain médian en précision est de 78% par rapport aux estimateurs "traditionnels". La méthode utilisée est brièvement décrite à la section 2. Une description détaillée de la problématique, de la méthode, ainsi que des résultats obtenus peut être trouvée dans la documentation fournie par l'Universidad Carlos III de Madrid, qui est disponible sur le microsite "Statistiques expérimentales".

Dans une deuxième étape, ces travaux on été intégrés à l'OFS, aussi bien afin d'élargir l'étude de faisabilité que pour gagner de l'expérience dans ce domaine. L'OFS a reproduit les estimations pour l'année 2012 et produit celles des années 2013 et 2014, puis a étudié leurs évolutions (cf. section 3). Par ailleurs, l'OFS a également produit des résultats provenant de la combinaison de méthodes SAE et d'un pooling sur plusieurs années (cf. section 4). La motivation principale est que cela permet d'obtenir des estimations fiables pour un plus grand nombre de communes que dans le cas d'exploitations annuelles, étant donné que la limite de la taille d'échantillon de 100 personnes reste à priori la même pour le pooling. L'impact du pooling sur la qualité des résultats a également été étudiée.

Enfin, nous présentons à la section 5 les conclusions de cette étude, avec ses limitations et perspectives.

2 Description de la méthode

2.1 Estimateur considéré

Les quantités que l'on cherche à estimer sont les taux d'activité dans les communes et s'expriment de la manière suivante :

$$P_d = \frac{1}{N_d}\sum_{i=1}^{N_d} Y_{di} = \frac{\text{Personnes actives dans } d}{\text{Population de } d},$$

où d est la commune considérée, N_d est le total de la population dans cette commune, et Y_{di} désigne la variable d'intérêt binaire qui indique si l'individu i de la commune d est actif ou non.

Afin d'estimer ces proportions, l'étude menée par l'Universidad Carlos III de Madrid considère plusieurs estimateurs possibles. De prime abord, parmi les modèles considérés, celui qui semble le mieux adapté pour modéliser une variable d'intérêt binaire est un modèle linéaire mixte généralisé (GLMM), avec la commune comme effet aléatoire. Cependant, dans notre cas, le choix s'est porté sur un modèle plus simple, de type modèle linéaire mixte (LMM). Les résultats du mandat ont en effet montré que l'estimateur basé sur un modèle LMM atteint dans ce contexte des performances équivalentes à celui basé sur un GLMM, et présente plusieurs avantages pratiques, notamment celui d'être moins lourd en termes de temps de calcul. Le modèle LMM est défini de la façon suivante :

$$Y_{di} = \mathbf{x}'_{di}\beta + u_d + e_{di}, \tag{1}$$

où $\mathbf{x}_{di}$ est un vecteur de variables auxiliaires, β est un vecteur de P paramètres, et où :

- u_d est un effet aléatoire pour la commune d qui, dans une certaine mesure et sous certaines hypothèses, permet de tenir compte de la spécificité de chaque commune. Les hypothèses sont les suivantes : les variables aléatoires u_d sont supposées indépendantes et identiquement distribuées, de moyenne nulle et de variance σ_u^2.
- e_{di} est un terme d'erreur qui reste inexpliquée. Les variables aléatoires e_{di} sont supposées indépendantes et identiquement distribuées, de moyenne nulle et de variance σ_e^2.

Lorsque, comme dans notre cas, la taille de l'échantillon est très faible par rapport à celle de la population, l'estimateur EBLUP [1] qui découle du modèle (1) s'exprime sous la forme de la combinaison d'un estimateur assimilable à un estimateur "traditionnel" et d'un estimateur purement synthétique (i.e. basé sur un modèle linéaire "classique" sans effet aléatoire spécifique pour la commune) :

$$\hat{P}_d = \hat{\gamma}_d \underbrace{\left\{\bar{y}_d + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_d)'\hat{\beta}\right\}}_{\text{Traditionnel}} + (1-\hat{\gamma}_d) \underbrace{\bar{\mathbf{X}}'_d\hat{\beta}}_{\text{Synthétique}}, \tag{2}$$

où, pour chaque commune d, $\bar{\mathbf{X}}_d$ représente la moyenne des variable auxiliaires sur la population totale, $\bar{\mathbf{x}}_d$ celle sur l'échantillon, $\bar{y}_d$ est la proportion d'actifs dans l'échantillon, et $\hat{\beta}$ est un estimateur de β. Le facteur de combinaison $\hat{\gamma}_d$ est donné par :

$$\hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_d), \tag{3}$$

1. Empirical Best Linear Unbiased Predictor.

où $\hat{\sigma}_u^2$ et $\hat{\sigma}_e^2$ sont des estimateurs des variances σ_u^2 et σ_e^2, et où n_d désigne la taille de l'échantillon dans la commune d. Il découle de l'expression de $\hat{\gamma}_d$ que plus la taille de l'échantillon dans la commune d est grande, plus la part attribuée à l'estimateur direct est importante, diminuant ainsi le risque de biais lié aux éventuels défauts du modèle.

Finalement, il est important de pouvoir garantir la cohérence entre les estimations communales obtenues par des méthodes SAE, et celle obtenues au niveau géographique supérieur, ici le canton, par des méthodes "traditionnelles". Cette propriété n'étant pas automatiquement garantie, il est nécessaire de procéder à un ajustement appelé "benchmarking". Plus précisément, en désignant par $\hat{Y}_d^{GREG}$ l'estimateur "traditionnel" GREG du nombre d'actifs dans la commune d, le nouvel estimateur "benchmark" est donné par :

$$\hat{P}_d^{BM} = \hat{P}_d \left(\sum_{d=1}^{D} \hat{Y}_d^{GREG} / \sum_{d=1}^{D} N_d \hat{P}_d \right),$$

où D désigne le nombre de communes dans le canton considéré. C'est cet estimateur qui est utilisé dans l'ensemble du travail.

2.2 Informations auxiliaires

Les variables intervenant dans le modèle sont listées dans le tableau 1. La construction du jeu de données est décrite dans l'annexe A. A noter l'introduction d'un effet fixe pour un district particulier proche du Liechtenstein, désigné comme district "atypique" dans le tableau 1. L'étude de l'Universidad Carlos III de Madrid a en effet montré une spécificité de ce district par rapport au reste de la population, que les variables auxiliaires à disposition ne suffisent pas à expliquer. Pour les communes du district "atypique", l'introduction d'un effet fixe au niveau du district permet de limiter le risque de biais dû à "l'inadéquation" du modèle pour ces communes particulières. En contrepartie, cela a pour conséquence d'augmenter la variance pour ces mêmes communes.

2.3 Modification mineure du modèle

Une légère modification du modèle élaboré par l'Universidad Carlos III de Madrid a été effectuée afin de garantir sa cohérence dans le temps. La version initiale utilise un effet fixe sur une commune particulière qui semble mal satisfaire l'hypothèse de normalité des effets aléatoires, qui est souvent utilisée dans le contexte de l'estimation SAE afin de pouvoir disposer d'un estimateur de la précision sous l'aléa du modèle (approche dite "model-based"). Or, nous avons constaté que les communes pour lesquelles un effet fixe serait utile selon ce critère ne sont pas les mêmes d'années en années. D'un point de vue pratique, il paraît difficilement envisageable d'adapter chaque année le modèle.

Par ailleurs, comme l'hypothèse de normalité n'est pas indispensable pour obtenir des estimtions SAE et étudier leurs propriétés sous l'approche "design-based" que nous avons adoptée, nous avons décidé de supprimer cet effet fixe. Cela a eu un effet négligeable sur l'ensemble des autres communes, et peu d'effet pour cette commune particulière (l'estimation du taux d'activité passe de 54.31% à 55.2%).

Tableau 1 – Variables incluses dans le modèle.

variable dépendante	Actif	actif (1), non-actif (0)
effet aléatoire	Commune	commune dans l'échantillon du RS (2475 communes en 2012)
effets fixes	District atypique	District atypique (1), sinon (0)
	Poids strate	strate avec médiane des poids de sondage élevée (0), sinon (1)
	Age	15, [16, 20), [20, 60), [60, 64), 64, ≥ 65
	Sexe	homme (1), femme (2)
	Nationalité	non-suisse (1), suisse (2)
	Etat civil	célibataire , non-marié (1), marié, partenariat enregistré (2), veuf (3), divorcé, partenariat dissout (4)
	AVS, 1er trimestre de l'année précédente	dans le relevé AVS seulement Jan-Mars (1), sinon (0)
	Revenu de l'année précédente	inconnu (hors registre AVS), (0, 12000], (12000, 24000], (24000, 48000], (48000, 72000], (72000, 96000], (96000, 120000], $>$ 120000
	Taille du foyer	1, 2, [3,5], [6,10], >10
	Résidence secondaire	non (1), oui (2)
	Age x Sexe	
	Etat civil x Sexe	

3 Estimations annuelles du taux d'activité

3.1 Comparaison des estimations annuelles

Les estimations du pourcentage d'actifs par commune ont été réalisées pour les années 2012, 2013 et 2014. Des comparaisons communes par communes ont été effectuées, et ne montrent pas de grandes variations entre deux années successives, voir tableau 2. Il est intéressant de remarquer que deux des extrema du tableau 2 sont atteints pour des communes avec un nombre de répondants très proche du seuil de 100 individus : 101 en 2013 individus pour le minimum entre 2012 et 2013 ; 100 en 2013 individus pour le maximum entre 2013 et 2014. Les deux autres extrema (maximum entre 2012 et 2013 et minimum entre 2013 et 2014) sont observés pour deux communes ayant des tailles d'échantillon plus conséquentes (environ 160 pour l'une et 300 pour l'autre). Il s'agit de deux communes du canton de Saint-Gall, proches l'une de l'autre géographiquement et situées dans le district "atypique" discuté section 2.1. Cela peut s'expliquer par la plus grande variabilité des estimations pour les communes située dans le district "atypique", due à l'introduction d'un effet fixe pour ce district particulier.

Tableau 2 – Statistique des différences des taux d'activité par commune entre deux années d'enquête successives, pour les communes sans mutation (fusion ou subdivision) entre 2012 et 2014 et présentant au moins une fois un échantillon de plus de 100 répondants.

	de 2012 à 2013	**de 2013 à 2014**
Moyenne	-0.12%	0.14%
Médiane	-0.14%	0.20%
Min	-3.88%	-3.32%
Max	3.22%	3.53%
Q1	-0.70%	-0.43%
Q3	0.46%	0.72%

3.2 Estimation de la précision des estimations

Plusieurs mesures de précision ont été étudiées dans le cadre du mandat confié à l'Universidad Carlos III de Madrid. La mesure retenue par l'OFS est la "design Mean Squared Error" (design-MSE), qui intègre les éventuels défauts du modèle sur lequel se basent les estimations SAE. Elle est la somme de deux composantes :

— la variance, qui résulte de l'incertitude liée à l'échantillonnage,

— le carré du biais, qui découle des éventuelles imperfections du modèle.

L'estimation de la design-MSE étant une question très complexe actuellement peu développée dans la littérature, ce sujet a été approfondi dans le cadre du mandat confié à l'Universidad Carlos III de Madrid. La méthode proposée est dénommée *parametric design bootstrap (PDB)* et est décrite dans l'annexe B. Dans ce contexte, elle permet des

estimations fiables de la design-MSE dès que la taille de l'échantillon dans le domaine est supérieure à 100 individus. A noter que l'estimation de la précision d'un estimateur étant souvent plus complexe que l'estimation du paramètre d'intérêt lui même, le seuil de 100 individus retenu pour des estimations fiables de la design-MSE permet également d'obtenir des estimations fiables du taux d'activité.

La mesure de précision qui accompagne nos résultats est en fait la racine carrée de la design-MSE, appelée "design-Root Mean Squared Error" (design-RMSE). Le passage à la racine carrée permet en effet d'obtenir une mesure de précision qui s'exprime dans la même unité que la variable d'intérêt, ici le taux d'activité en pourcentages, et qui peut donc lui être directement comparée. Les design-RMSE ont été estimées pour les années 2012 à 2014, voir tableaux 3 et 4. La distribution des estimations de la design-RMSE est stable entre les différentes années d'estimations. En revanche, nous avons observé que commune par commune, des évolutions relativement importantes sont possibles. Ce point doit encore être approfondi. Les évolutions doivent également être mises en regard avec l'ordre de grandeur de l'estimation de la design-MSE, qui reste très faible, comme le montre le tableau 3.

Tableau 3 – RMSE, en pourcentages, estimées pour les années 2012 à 2014 pour des communes sans mutation, avec un nombre de répondants toujours supérieur à 100.

Level	2012	2013	2014
100% Max	3.13	2.85	2.93
99%	2.55	2.25	2.38
95%	1.94	1.85	1.89
90%	1.57	1.55	1.58
75% Q3	1.18	1.07	1.10
50% Median	0.67	0.63	0.64
25% Q1	0.42	0.36	0.38
10%	0.32	0.25	0.28
5%	0.28	0.22	0.24
1%	0.23	0.18	0.19
0% Min	0.20	0.14	0.10

Tableau 4 – Différence des RMSE entre deux années d'enquête consécutives pour les communes avec toujours plus de 100 répondants.

	de 2012 à 2013	**de 2013 à 2014**
Moyenne	-0.06	0.03
Médiane	-0.03	0.02
Min	-2.50	-2.38
Max	2.44	1.81
Q1	-0.44	-0.39
Q3	0.34	0.49

4 Pooling

Afin d'obtenir des estimations fiables pour un nombre plus important de communes, nous avons réalisé un pooling des trois années d'enquête de 2012 à 2014. Ceci permet d'avoir à disposition un nombre de réponses environ trois fois supérieur et de ce fait d'augmenter le nombre de communes qui franchissent le seuil de 100 personnes dans l'échantillon, pour lesquelles les résultats seront considérés comme publiables.

4.1 Impact du pooling sur le nombre de communes publiables

Dans un premier temps, nous avons étudié l'évolution du nombre de communes publiables à l'aide d'estimations SAE, entre plusieurs enquêtes successives. Ce nombre peut en effet varier entre deux années successives, du fait des fluctuations d'échantillonnage et de l'effet de la non-réponse qui peuvent faire basculer d'un côté ou de l'autre les communes dont l'échantillon annuel est proche du seuil de 100 personnes. Les changements dans les densifications cantonales peuvent également avoir un impact. De plus, chaque année des mutations (essentiellement des fusions) de communes ont lieu et cela engendre également des changements dans le nombre de communes publiables. Cependant, nous avons constaté que ce nombre reste assez stable, et représente environ 30% des communes. Le pooling permet d'augmenter nettement ce taux, qui passe à environ 60% pour un pooling sur trois ans, voir tableau 5. Par ailleurs, nous estimons qu'un pooling sur cinq ans permet d'augmenter ce taux à environ 70%.

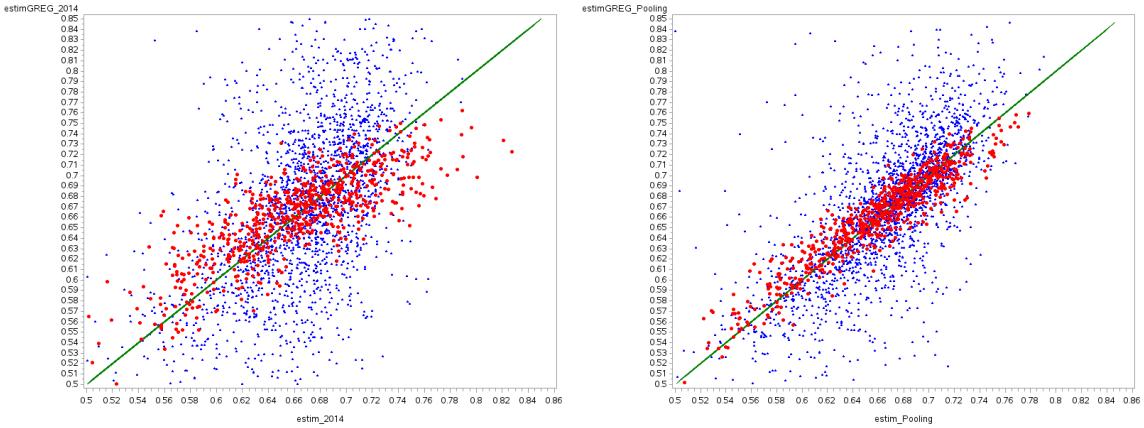
Tableau 5 – Nombre et pourcentage de communes pour lesquelles la proportion du nombre d'actifs est publiable ou non en cas de pooling, avec état des communes en 2014.

Pooling 2012 à 2014	**Fréquence**	**Pourcentage**
non-publiable	924	39.3
publiable	1427	60.7
total	2351	100

4.2 Impact du pooling sur la qualité des estimations

Nous commençons tout d'abord par nous intéresser à l'impact du pooling sur le risque de biais dû aux éventuels défauts du modèle. A cet effet, nous nous basons sur le graphique 1, qui représente, pour chaque commune, l'estimation "traditionnelle" (GREG), supposée pas ou peu biaisée, en fonction de l'estimation SAE. Ceci est réalisé pour l'année 2014, puis pour le pooling 2012-2014. Ce type de représentation graphique peut permettre de détecter un éventuel biais. En effet, en l'absence de biais, les points devraient être uniformément répartis autour de la droite d'équation $y = x$ (en vert), et ne pas montrer de tendance ou de forme particulière. Les communes avec plus de 100 répondants en 2014 sont indiquées en rouge. Il s'agit des communes pour lesquelles à la fois l'estimation annuelle (2014) et l'estimation poolée sont considérées comme sufisamment fiables pour être publiées, et pour lesquelles évaluer l'effet du pooling sur le risque de

biais a du sens. A noter également que le graphique 1 est un zoom sur les communes dont l'estimation SAE et l'estimation "traditionnelle" du taux d'activité se situent entre 50% et 85%. Cela représente la majorité des communes ayant plus de 100 répondants en 2014 (en rouge).



(a) 2014 (b) Pooling 2012-2014

Graphique 1 – Comparaison des estimations "traditionnelles" (GREG) avec les estimations SAE pour l'année 2014, puis pour le pooling 2012-2014. Les communes avec plus de 100 répondants en 2014 sont indiquées en rouge.

On constate que le pooling a pour effet de rapprocher les points de la droite d'équation $y = x$, c'est-à-dire que l'estimation SAE est plus proche de l'estimation "traditionnelle" dans le cas du pooling que dans celui de l'estimation annuelle. Cela est attribuable à deux facteurs. D'une part, l'estimateur "traditionnel" est moins instable dans le cas du pooling que dans celui des estimations annuelles, du fait de l'augmentation de la taille de l'échantillon. L'estimation SAE étant alors comparée à une valeur de référence plus stable, le diagnostic visuel permis par le graphique 1 est plus précis et plus pertinent dans le cas du pooling. En ce sens, la partie b) du graphique 1 est encourageante quant à la validité des estimations SAE. D'autre part, pour une commune donnée, le pooling a pour effet d'augmenter la part de l'estimateur assimilable à un estimateur "traditionnel" dans l'équation (2), puisque cette part augmente avec la taille de l'échantillon. Comme l'estimation "traditionnelle" n'est en principe pas ou peu biaisée, le pooling devrait donc avoir un impact positif sur le risque de biais. L'importance de cet effet ne peut pas être évaluée sur la base du graphique 1, et cette question doit encore être étudiée plus en détails.

Il est également intéressant d'évaluer l'impact du pooling sur l'estimation de la précision, en termes de design-RMSE. Tout d'abord, le tableau 6 donne des informations sur la distribution de l'estimation de la design-RMSE dans le cas du pooling 2012-2014. Il s'agit dans ce tableau de communes qui totalisent plus de 100 répondants sur les données du pooling, et pour lesquelles les résultats du pooling seraient donc publiables. On constate que la distribution de l'estimation de la design-RMSE obtenue dans le cas du pooling est similaire à celles obtenues pour les estimations annuelles (cf. tableau 3). Autrement dit, le pooling permet d'obtenir une distribution de l'estimation de la design-RMSE similaire au cas annuel, mais pour un nombre plus important de communes.

Tableau 6 – RMSE pour le pooling 2012 - 2014 pour les communes avec un nombre de répondants supérieur à 100.

Level	$100 \cdot \sqrt{MSE}$
100% Max	3.39
99%	2.70
95%	1.93
90%	1.58
75% Q3	1.08
50% Median	0.69
25% Q1	0.43
10%	0.33
0.05	0.30
1%	0.25
0% Min	0.11

Pour étudier plus en détails l'impact du pooling sur la design-RMSE, le tableau 7 donne des statistiques descriptives sur la différence, commune par commune, entre l'estimation des design-RMSE annuelles et celle du pooling, pour les communes avec un nombre de répondants supérieur à 100 dans les données annuelles. Ces résultats semblent indiquer que le pooling n'apporte pas d'amélioration flagrante sur l'estimation de la design-RMSE. L'effet du pooling sur l'estimation de la design-RMSE peut également être visualisé à l'aide du graphique 2, qui représente l'estimation de la design-RMSE du pooling en fonction des design-RMSE annuelles. Chaque année est représentée par une couleur différente : 2014 en bleu, 2013 en vert et 2012 en violet. On constate que les points sont assez dispersés autour de la droite d'équation $y = x$ (en rouge), ce qui va dans le sens du faible impact du pooling sur l'amélioration de l'estimation de la design-RMSE. Deux hypothèses peuvent être émises pour tenter d'expliquer cette observation :

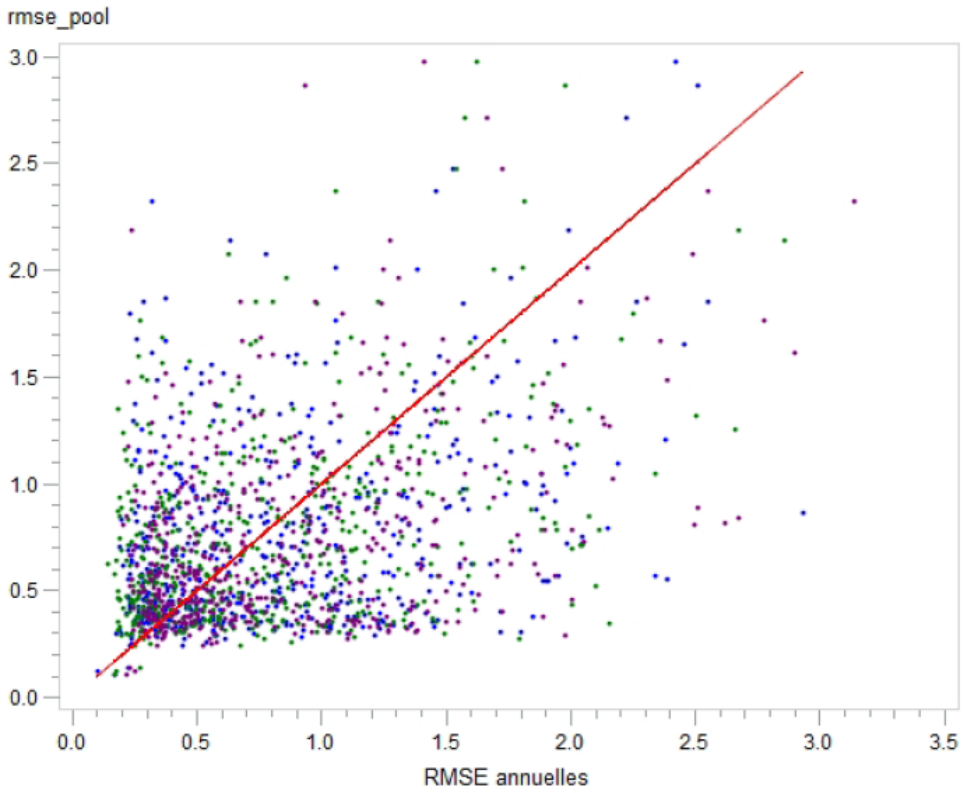
— soit le pooling ne permet pas une très grande amélioration de la précision, en termes de design-MSE,

— soit la méthode d'estimation de la design-RMSE n'est pas assez précise pour détecter l'effet du pooling sur la "vraie" design-RMSE.

Cette question doit encore être approfondie, et souligne encore davantage que l'estimation de la design-MSE est une question très complexe.

Tableau 7 – Statistiques descriptives des différences, commune par commune, entre l'estimation de la design-RMSE du pooling et la design-RMSE annuelle, pour les communes avec un nombre de répondants toujours supérieur à 100 dans le données annuelles.

	2012	**2013**	**2014**
Moyenne	-0.08	-0.02	-0.05
Médiane	-0.05	0.00	-0.01
Min	-1.82	-1.80	-2.06
Max	1.96	1.50	2.01
Q1	-0.36	-0.32	-0.37
Q3	0.24	0.27	0.26



Graphique 2 – Comparaison des estimations des design-RMSE annuelles et de celle du pooling, pour les communes avec plus toujours de 100 répondants pour les données annuelles. Chaque année est représentée par une couleur différente : 2014 en bleu, 2013 en vert et 2012 en violet.

5 Conclusions, limitations et perspectives

Ce travail a permis d'évaluer la possibilité de combiner un pooling et des méthodes SAE pour estimer le taux d'activité au niveau communal sur la base du relevé structurel. Nous avons constaté que :

- Le pooling augmente le nombre de communes qui atteignent le seuil de 100 personnes dans l'échantillon, et pour lesquelles les résultats sont jugés publiables. Ainsi, il est possible d'obtenir des estimations fiables du taux d'activité et de sa précision pour environ 60% des communes dans le cas d'un pooling sur trois ans.
- Le pooling devrait avoir un impact positif sur le risque de biais pour les communes pour lesquelles des estimations annuelles seraient possibles. L'importance de cet effet n'a pas encore été évaluée, et cette question doit encore être approfondie.
- Le pooling ne semble pas permettre d'amélioration flagrante de l'estimation de la précision, en termes de design-RMSE. Cette question doit encore être approfondie.

Les résultats obtenus sont donc encourageants, mais il est nécessaire de garder à l'esprit les limitations suivantes :

1. Limitations usuelles dues au pooling ;
   - Définition de la population : population moyenne, mutations de communes,
   - Lissage des évolutions de structure par l'utilisation d'une population moyenne,
   - Instabilité potentielle en cas d'adaptation du processus d'enquête ;
2. Limitations intrinsèques aux méthodes « petits domaines » :
   - Lissage des résultats dû à l'utilisation d'un modèle,
   - Choix et validation du modèle,
   - Complexité de l'estimation de la design-MSE ;
3. Défi pour la communication et l'interprétation des résultats.

Nous espérons que ces travaux puissent susciter l'intérêt et amener vers une publication officielle. Dans cette perspective, nous recueillons tout commentaire constructif qui nous permette de consolider ce travail.

Annexes

A Création du jeu de données

Dans le but d'avoir un maximum d'informations auxiliaires pertinentes, les données utilisées pour estimer la taux d'activité par commune sont construites sur la base d'un appariement, au niveau de l'individu, de différentes sources de données :

1. Données d'enquête : relevé structurel ;
2. Données sur l'ensemble de la population :
   (a) Statistique de la population et des ménages (STATPOP), selon la définition suivante de la population cible, qui correspond à celle du relevé structurel :
   — population résidante permanente au domicile principal ;
   — individus vivant dans des ménages privés ;
   — individus de 15 ans et plus ;
   — exclusion des diplomates, des fonctionnaires internationaux, des personnes dans un processus d'asile depuis moins d'une année, ainsi que des personnes sous tutelle.
   (b) Données AVS de l'année précédente [2] :
   — présence ou non dans le relevé AVS de Janvier à Mars ;
   — revenu annuel en francs, dans les limites de classes suivantes : 0, 12000, 24000, 48000, 72000, 96000, 120000, plus de 120000, hors registre AVS.

Cette phase d'appariement et de création des jeux de données est réalisée avec SAS. Tous les autres programmes sont développés en R.

2. Les données AVS de l'année de référence ne sont pas disponibles dans la production courante du RS.

B Procédure d'estimation de la design-MSE

La méthode proposée par l'Universidad Carlos III de Madrid est dénommée *parametric design bootstrap (PDB)* et repose sur la procédure suivante :

1. Ajustement du modèle sur l'échantillon avec la commune comme effet aléatoire du modèle et obtention de l'estimateur $\hat{P}_d$;
2. Ajustement du modèle sur l'échantillon avec la commune comme effet fixe du modèle et obtention de l'estimateur $\hat{P}_d^{FIX}$;
3. Génération d'une population bootstrap sur la base des paramètres du modèle obtenus au point 1 et répétition des étapes suivantes, $b = 1, \ldots, B$, avec $B = 250$:
   (a) Sélection d'un échantillon aléatoire simple parmi la population bootstrap ;
   (b) Ajustement du modèle LMM sur cette population ;
   (c) Estimation du taux d'activité pour cette population et obtention de l'estimateur $\hat{P}_d^{\star(b)}$.
4. Calcul de l'estimateur de la design MSE :

$$mse_{PDB}\left(\hat{P}_d\right) = \hat{\gamma}_d \frac{1}{B}\sum_{b=1}^{B}\left(\hat{P}_d^{\star(b)} - \hat{P}_d^{FIX}\right)^2 + (1-\hat{\gamma}_d)\frac{1}{B}\sum_{b=1}^{B}\left(\hat{P}_d^{\star(b)} - \hat{P}_d\right)^2,$$

où $\hat{\gamma}_d \in [0,1]$ est le facteur de combinaison donné dans l'équation (3).