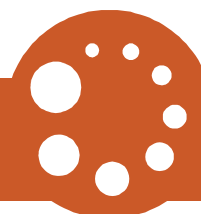




Estimation du taux d'activité au niveau communal dans le cadre du relevé structurel

Analyse de l'impact du pooling

EXPERIMENTAL STATISTICS



Neuchâtel, 2021

Éditeur: Office fédéral de la statistique (OFS)
Renseignements: info.pop@bfs.admin.ch, tél. +41 58 467 25 25
Rédaction: Djalel-Eddine Meskaldji ; Anne Massiani ;
Daniel Kilchmann
Domaine: 00 Bases statistiques des généralités
Langue du texte original: Français

Concept de mise en page: Section DIAM
Téléchargement: www.statistique.ch
Copyright: OFS, Neuchâtel 2021
La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée

Table des matières

1	Contexte	2
2	Impact du pooling	2
3	Étude théorique du gain dû au pooling sur la RMSE	5
4	Instabilité de l'estimation de la RMSE par le PDB	9
5	Conclusions	12
6	Annexe	13
7	Références	14

Résumé Les analyses présentées dans ce document ont été menées dans le cadre de l'élargissement de la publication du taux d'activité par commune, passant d'un pooling de trois à un pooling de cinq ans du relevé structurel (RS). Le but principal était de publier le taux d'activité pour plus de communes que ce n'était possible avec un pooling sur trois ans.

Les analyses se sont focalisées d'abord sur le gain théorique, en termes de la racine de l'erreur quadratique moyenne (Root Mean Square Error (RMSE), en anglais) sous le plan, dont peuvent profiter les estimations petits domaines, en plus de l'augmentation du nombre de communes publiables, c.à.d. qui atteignent le seuil de 100 personnes dans l'échantillon poolé, dans le cas d'un pooling du RS. Ensuite, la stabilité de l'estimation de la RMSE par une technique du bootstrap paramétrique a été analysée. Sur la base de ces analyses on peut conclure que dans le cas d'un pooling de 5 ans le gain peut s'élever à environ 55% de la RMSE pour les grandes communes et à environ 15% pour les communes dont la taille d'échantillons poolés avoisine 100 observations. Néanmoins, ce gain est rarement perceptible, d'une part à cause de l'ordre de grandeur faible de la RMSE et d'autre part parce que l'estimation de la RMSE présente une certaine instabilité, comme le montre la deuxième partie des analyses effectuées.

1 Contexte

Le relevé structurel est reconduit annuellement, ce qui permet de procéder à un pooling (un regroupement d'échantillons) sur plusieurs années. Les premiers résultats d'estimations du taux d'activité par commune à l'aide d'un pooling de trois ans du relevé structurel (RS) et des estimations sur petits domaines (Small area estimation - SAE) ont été publiés en 2018 sur le "microsite statistiques expérimentales". Cette publication combinait le pooling des années 2012, 2013 et 2014 du RS avec des résultats de recherche, confiée à la Prof. I. Molina de l'« Universidad Carlos III de Madrid ». Dans le cadre de ce projet, le pooling s'effectue en regroupant les échantillons de plusieurs années en un seul échantillon. Ainsi, les estimations SAE se calculent sur la base de cet échantillon poolé. Par conséquent, le taux d'activité a pu être publié pour environ 60% des communes, couvrant environ 94% des personnes âgées de 15 ans et plus de la population résidente permanente de Suisse.

Suite à cette publication, il a été décidé d'élargir le nombre de communes publiables selon les mêmes critères de qualité que ceux adoptés pour la première publication, en se basant sur un pooling de cinq ans (2012 - 2016). De plus, des analyses supplémentaires ont été menées afin d'éclaircir certaines questions soulevées lors de la première publication, notamment en relation avec le gain dû au pooling, en termes de la RMSE. Les analyses dans ce rapport se sont ainsi focalisées d'abord sur le gain théorique, en termes de la vraie RMSE, grâce aux estimations sur petits domaines dans le cas d'un pooling du RS. Ensuite, la stabilité de l'estimation de la RMSE par une technique de bootstrap paramétrique a été analysée.

2 Impact du pooling

Le pooling augmente le nombre de communes publiables, et ainsi le nombre d'individus concernés par la publication, voir tables 1 et 2. L'impact du pooling des échantillons de 2012 à 2014 a été étudié dans le rapport [1] dont un bref résumé est fourni ci-dessous.

La figure 1 montre les RMSE estimées pour les années 2012 et 2014 comparées à celles de l'année 2013. Toutes les RMSE sont estimées par la technique 'parametric design bootstrap (PDB)' proposée par [2], voir Section 4 pour plus de détails. Ces estimations peuvent varier considérablement selon les années. Cependant, cette variabilité est beaucoup moins impor-

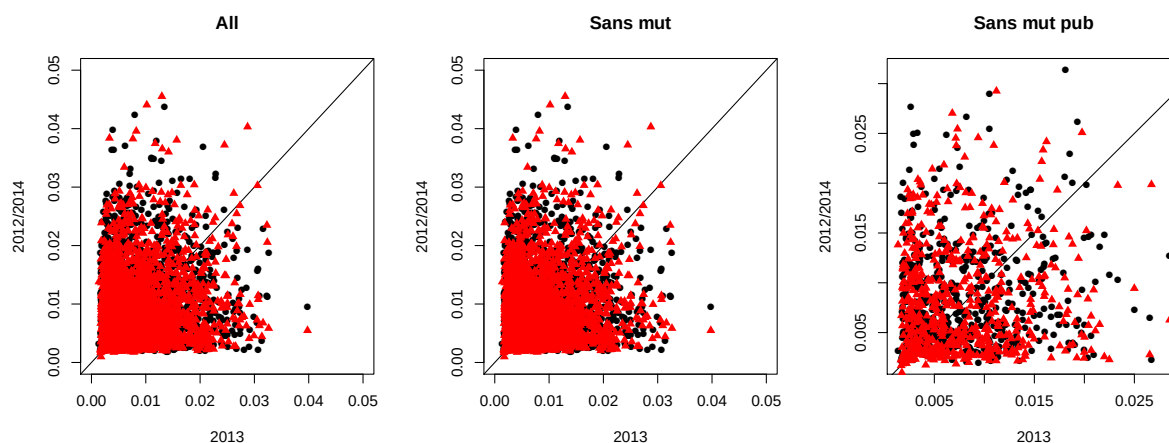


FIGURE 1 Les estimations des RMSE du taux d'activité au niveau des communes évaluées par le PDB, sans 6 "outliers". Les RMSE de 2012 (points noirs) et de 2014 (triangles rouges) sont comparées à celles de l'année 2013. Panel de gauche : toutes les communes ; panel du centre : les communes sans mutation ; panel de droite : les communes publiables sans mutation.

tante pour les communes publiables et sans mutation. Notons que dans le panel de gauche de la figure 1, six communes ont été enlevées car considérées comme des "outliers".

TABLE 1 Nombre et pourcentage de communes pour lesquelles le taux d'activité est publiable ou non en cas du pooling de 2012 à 2014 et de 2012 à 2016, avec état des communes en 2014 et 2016, respectivement.

	Pooling 2012 à 2014		Pooling 2012 à 2016	
	Fréquence	Pourcentage	Fréquence	Pourcentage
non-publiables	924	39.3%	540	23.6%
publiables	1427	60.7%	1747	76.4%
total	2351	100%	2287	100%

TABLE 2 Nombre et pourcentage d'individus dans les communes pour lesquelles le taux d'activité est publiable ou non en cas du pooling de 2012 à 2014 et de 2012 à 2016, avec état des communes en 2014 et 2016, respectivement.

	Pooling 2012 à 2014		Pooling 2012 à 2016	
	Fréquence	Pourcentage	Fréquence	Pourcentage
non-publiables	407423	6%	165108	2.4%
publiables	6304675	93.9%	6619132	97.6%
total	6712098	100%	6784240	100%

La figure 2 compare les estimations des RMSE annuelles des années de 2012 à 2014 à celles du pooling de ces trois années, toutes estimées par le PDB. Chaque année est représentée par une couleur et une forme différentes. Clairement, ce graphique ne met pas en évidence un gain « systématique » dû au pooling. De plus, l'étendue des valeurs est du même ordre de grandeur pour les estimations annuelles que pour l'estimation basée sur l'échantillon poolé. Enfin, les points ne sont pas concentrés autour de la première bissectrice, ce qui indique une forme d'instabilité d'estimation de la RMSE à travers les années. Néanmoins, ces estimations

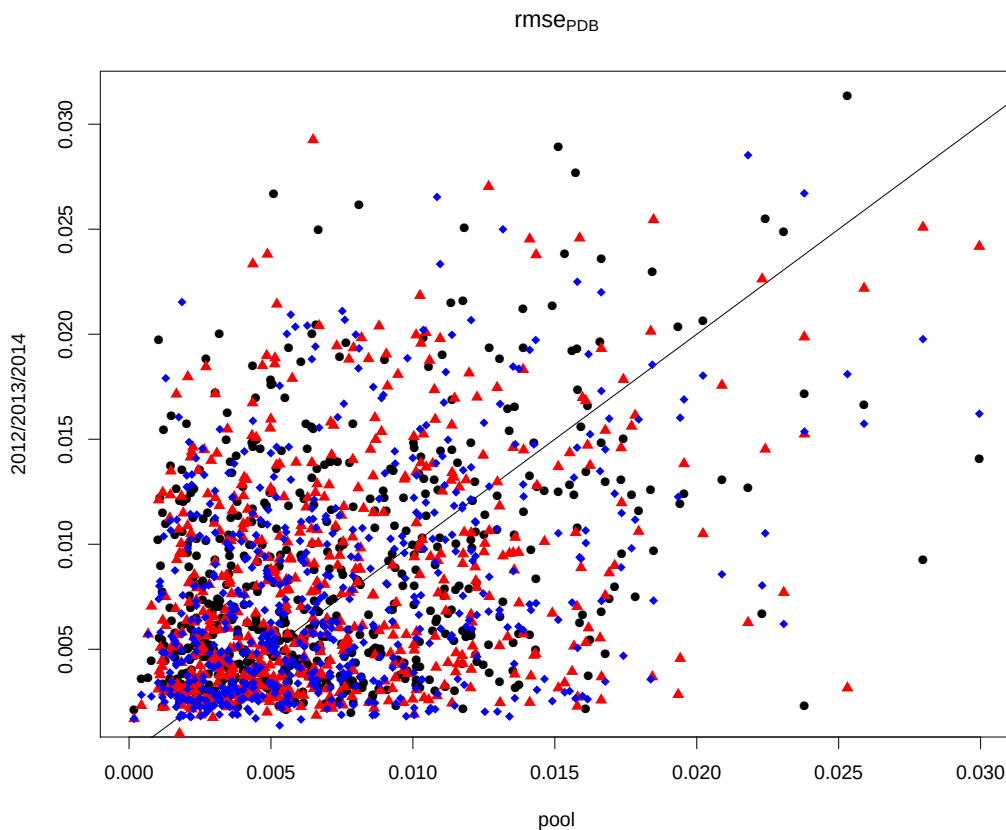


FIGURE 2 Comparaison des estimations des RMSE annuelles et de celles du pooling, pour les communes avec plus de 100 répondants pour toutes les années. Chaque année est représentée par une couleur et une forme différentes : 2012 points noir, 2013 étoiles bleues et 2014 triangles rouges.

restent relativement petites (cf. échelle). Dans la figure 3, on voit que pour chaque commune, les estimations des RMSE dans le cas des exploitations basées sur un pooling sont presque toutes inférieures au maximum des estimations des RMSE basées sur les exploitations annuelles. Le pooling permet aussi de réduire la moyenne, l'étendue et surtout la dispersion des estimations de la RMSE par rapport aux exploitations annuelles, voir figure 4.

Les deux hypothèses suivantes avaient alors été émises pour tenter d'expliquer ces observations (voir [1, p. 10]) :

- soit le pooling ne permet pas une très grande amélioration de la précision, en termes de RMSE,
- soit la méthode d'estimation de la RMSE n'est pas assez précise pour détecter l'effet du pooling sur la "vraie" RMSE.

Pour essayer de trancher entre ces deux hypothèses, nous allons procéder en deux étapes. Dans la première, nous allons étudier l'impact théorique qui devrait être apporté par le pooling sur la RMSE et ces composantes. La seconde étape consiste à étudier la stabilité de l'estimation de l'erreur par la technique du PDB.

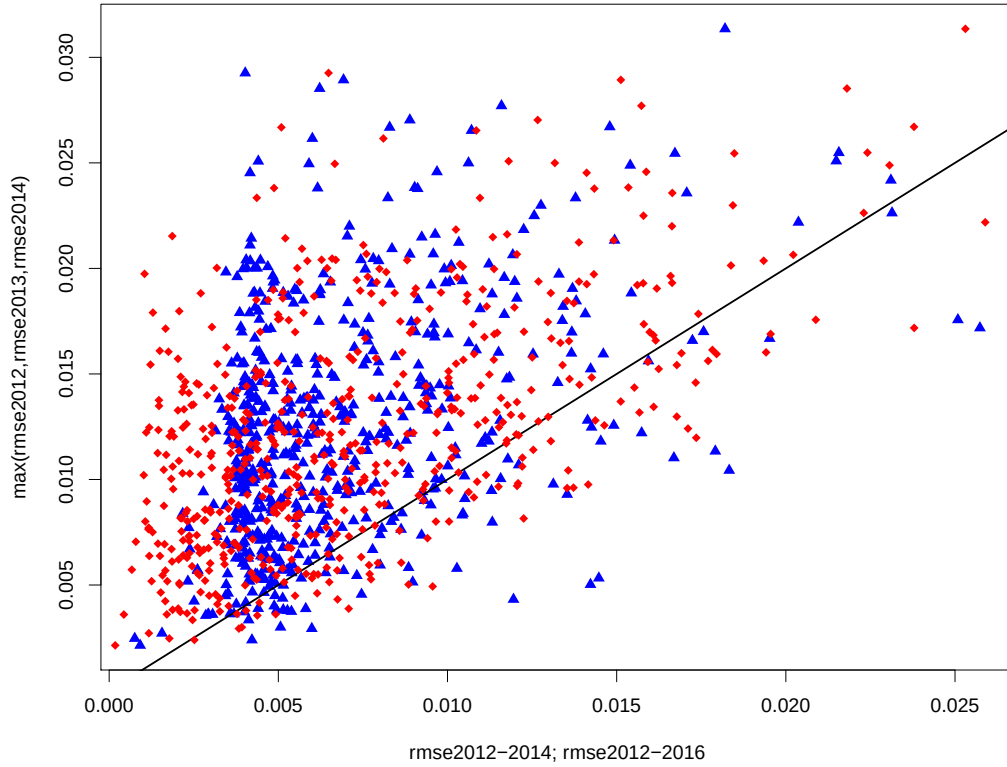


FIGURE 3 Comparaison des maximums des estimations des RMSE annuelles et de celles du pooling de 2012-2014 et du pooling de 2012-2016, pour les communes avec plus de 100 répondants. Pooling de 2012-2014 : étoiles rouges et pooling de 2012-2016 : triangles bleus.

3 Étude théorique du gain dû au pooling sur la RMSE

Le taux d'activité au sein des communes, représentant les petits domaines, est le paramètre d'intérêt de la publication. Il est donc une moyenne de variables indicatrices y_{id} qui indiquent si un individu i du domaine d est actif ou non. En se basant sur les valeurs observées y_{id} de l'échantillon et les données auxiliaires X_d (disponibles pour toute la population), le paramètre d'intérêt est estimé par la méthode SAE selon l'équation suivante :

$$\hat{P}_d = \hat{\gamma}_d \hat{y}_d^{GREG} + (1 - \hat{\gamma}_d) \bar{X}_d \hat{\beta} = \hat{\gamma}_d [\bar{y}_d + (\bar{X}_d - \bar{x}_d) \hat{\beta}] + (1 - \hat{\gamma}_d) \bar{X}_d \hat{\beta}, \quad (1)$$

où $\bar{y}_d$ et $\bar{x}_d$ désignent des moyennes sur l'échantillon, tandis que $\bar{X}_d$ est calculé sur l'ensemble de la population. Le terme $\hat{\beta}$ représente les coefficients de régression estimés selon un modèle linéaire mixte (LMM) ajusté sur tous les domaines en même temps. Le terme $\hat{\gamma}_d$ est donné par

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

où $\hat{\sigma}_u^2$ et $\hat{\sigma}_e^2$ sont des estimation de σ_u^2 et σ_e^2 qui sont la variance des effets aléatoires et la variance des résidus respectivement, dans un modèle LMM, et n_d désigne la taille de l'échantillon dans le domaine d . Le facteur $\hat{\gamma}_d$ représente la part de l'estimateur direct $\hat{y}_d^{GREG}$ dans l'estimateur SAE. Pour plus de détails, voir [1, p. 2].

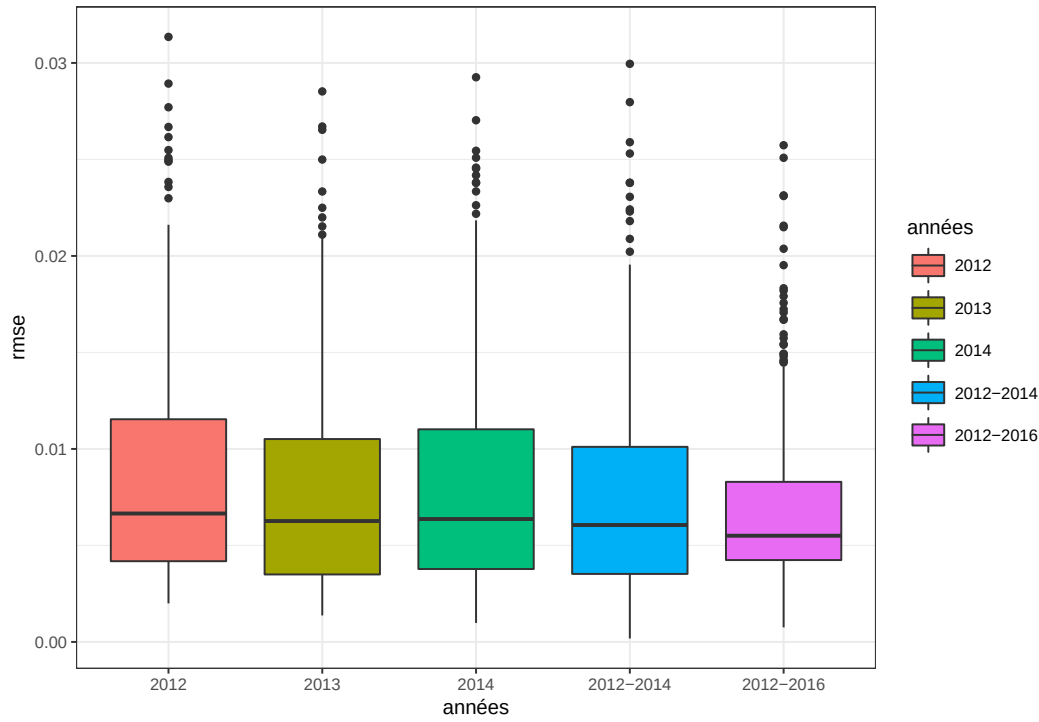


FIGURE 4 Boxplot des estimations des RMSE annuelles et de celles du pooling de 2012-2014 et du pooling de 2012-2016, pour les communes avec plus de 100 répondants.

Ici, on aimerait étudier l'impact théorique du pooling sur la "vraie" RMSE. Le regroupement des échantillons devrait réduire la variance de l'estimation puisque la taille de l'échantillon de chaque domaine ne peut qu'augmenter. Cependant, il n'est pas clair quel est l'impact du pooling sur la deuxième composante de l'erreur, c.à.d. le biais, puisque pour certains domaines, les imperfections dues au modèle pourraient se répéter à travers les années. Par conséquent, le gain dû au pooling serait différent selon que le carré du biais soit négligeable par rapport à la variance de l'estimation ou non. Dans la suite, nous nous basons sur un résultat de la Prof. I. Molina qui suppose que les paramètres sont connus.

Nous considérons T échantillons pour estimer le paramètre d'intérêt P_d^t , c.à.d. le taux d'activité dans le domaine d pour une année t avec $t = 1, \dots, T$. Nous considérons aussi l'estimateur du paramètre d'intérêt P_d basé sur l'échantillon poolé pour chaque domaine d . Soient B_d , S_d^2 et MSE_d la moyenne des résidus (directement en lien avec le biais de l'estimateur SAE dans le domaine d), la dispersion des résidus dans le domaine, et l'erreur quadratique moyenne de l'estimateur basé sur l'échantillon poolé, et soient B_d^t , S_d^{2t} et MSE_d^t les mêmes caractéristiques basées sur l'échantillon correspondant au temps t pour $t = 1, \dots, T$.

Nous supposons que les échantillons tirés au fil des années sont disjoints et ont approximativement le même taux d'échantillonnage, c.à.d. que le taux d'échantillonnage est $f_d^t \approx f_d$ pour tout $t = 1, \dots, T$. En plus, nous supposons que la moyenne et la dispersion des résidus dans chaque domaine de ces modèles ne changent pas à travers les années, c.à.d.

$$B_d = B_d^t \quad (3)$$

et

$$S_d^2 = S_d^{2t}. \quad (4)$$

TABLE 3 Les tailles des échantillons et les écart-types des effets aléatoires et des résidus estimés par les modèles LMM sur les échantillons des années 2012, 2013 et 2014, ainsi que sur les échantillons poolés de 2012 à 2014 et de 2012 à 2016.

Années	$\hat{\sigma}_u$	$\hat{\sigma}_e$	n
2012	0.00936	0.283	286015
2013	0.00902	0.282	281990
2014	0.00914	0.281	287842
2012-2014	0.00990	0.282	855847
2012-2016	0.01100	0.282	1399569

La Table 3 présente les $\hat{\sigma}_u$ et $\hat{\sigma}_e$ estimées par les modèles LMM sur les échantillons des années 2012, 2013 et 2014 ainsi que sur les échantillons poolés de 2012 à 2014 et de 2012 à 2016. On voit bien que les différentes années conduisent à des écart-types des résidus qui sont très proches, ce qui consolide notre hypothèse sur la moyenne et la dispersion des résidus. Cette hypothèse est aussi appuyée par la Table 4 qui montre les coefficients $\hat{\beta}$ des modèles LMM correspondants. La stabilité des modèles peut être expliquée par le fait que les échantillons annuels sont assez grands pour produire des modèles stables au fil des années et aussi que le phénomène auquel on s'intéresse est assez stable sur une période de quelques années. Cependant, les écart-types des effets aléatoires $\hat{\sigma}_u$ correspondants au pooling de 2012 à 2014 et de 2012 à 2016 sont légèrement plus grands que ceux correspondants aux échantillons annuels. La Table 3 contient aussi les tailles des échantillons annuels qui sont comparables d'année en année. Ainsi, les taux de sondage des domaines devraient également être stables si le même plan de sondage est utilisé chaque année.

La combinaison des échantillons peut se faire de plusieurs façons. L'option qui a été choisie est de regrouper les échantillons de plusieurs années en un seul échantillon pour estimer le paramètre d'intérêt. Nous allons donc étudier le gain théorique dû au pooling selon cette option. En supposant que les taux de sondage sont négligeables, l'erreur quadratique moyenne est donnée par (selon proposition 1 du matériel supplémentaire de [2])

$$\text{MSE}_d^t = \gamma_d^t \frac{S_d^{2t}}{n_d^t} + (1 - \gamma_d^t)(B_d^t)^2, \quad (5)$$

avec $\gamma_d^t = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2/n_d^t}$. Comme mentionné précédemment, ce facteur représente la part de l'estimateur direct dans l'estimateur SAE. Grâce à cette part basée sur l'estimateur direct (supposé sans biais), le biais de l'estimateur SAE est donc réduit, mais de peu pour des domaines ayant une petite taille d'échantillon. En remplaçant la formule de γ_d^t dans (5) on obtient

$$\text{MSE}_d^t = \frac{1}{\sigma_u^2 n_d^t + \sigma_e^2} (\sigma_u^2 S_d^{2t} + \sigma_e^2 (B_d^t)^2). \quad (6)$$

De la même façon on peut déduire que

$$\text{MSE}_d = \frac{1}{\sigma_u^2 n_d + \sigma_e^2} (\sigma_u^2 S_d^2 + \sigma_e^2 (B_d)^2). \quad (7)$$

En utilisant les équations (3) et (4), le rapport entre la sommes des carrés des erreurs pour

TABLE 4 Les coefficients β et les écart-types (std) correspondants des modèles ajustés pour les années 2012 à 2014, ainsi que le modèle ajusté sur les données poolées de 2012 à 2014.

names	$\hat{\beta}_{2012}$	$\hat{std}_{2012}$	$\hat{\beta}_{2013}$	$\hat{std}_{2013}$	$\hat{\beta}_{2014}$	$\hat{std}_{2014}$	$\hat{\beta}_{2012-14}$	$\hat{std}_{2012-14}$
(Intercept)	0.146	0.007	0.154	0.007	0.158	0.007	0.154	0.004
Strata1.f1	-0.009	0.001	-0.011	0.001	-0.009	0.001	-0.011	0.001
District1724.f1	0.078	0.010	0.097	0.010	0.068	0.010	0.080	0.007
age6.f2	0.348	0.007	0.357	0.007	0.331	0.007	0.346	0.004
age6.f3	0.294	0.007	0.295	0.007	0.291	0.007	0.293	0.004
age6.f4	0.137	0.008	0.136	0.008	0.137	0.008	0.137	0.004
age6.f5	0.008	0.010	0.015	0.010	0.000	0.009	0.008	0.006
age6.f6	-0.107	0.007	-0.115	0.007	-0.119	0.007	-0.113	0.004
sex.f2	-0.015	0.009	-0.023	0.010	-0.022	0.009	-0.020	0.005
marital.f2	0.015	0.002	0.019	0.002	0.016	0.002	0.017	0.001
marital.f3	0.008	0.006	0.006	0.006	0.006	0.006	0.007	0.003
marital.f4	0.009	0.003	0.008	0.003	0.012	0.003	0.010	0.002
nation.f2	-0.015	0.001	-0.017	0.001	-0.015	0.001	-0.016	0.001
secresid.f2	-0.065	0.005	-0.076	0.005	-0.080	0.005	-0.073	0.003
housesize.f2	-0.002	0.002	-0.003	0.002	-0.001	0.002	-0.002	0.001
housesize.f3	-0.015	0.002	-0.019	0.002	-0.016	0.002	-0.017	0.001
housesize.f4	-0.030	0.003	-0.029	0.003	-0.026	0.003	-0.028	0.002
housesize.f5	0.008	0.016	-0.039	0.018	-0.056	0.019	-0.025	0.010
Income8.f2	0.352	0.003	0.349	0.003	0.350	0.003	0.350	0.001
Income8.f3	0.481	0.002	0.480	0.002	0.476	0.002	0.478	0.001
Income8.f4	0.556	0.002	0.551	0.002	0.545	0.002	0.550	0.001
Income8.f5	0.573	0.002	0.567	0.002	0.563	0.002	0.567	0.001
Income8.f6	0.577	0.002	0.572	0.002	0.567	0.002	0.572	0.001
Income8.f7	0.581	0.002	0.574	0.002	0.573	0.002	0.576	0.001
Income8.f8	0.598	0.006	0.591	0.006	0.593	0.006	0.594	0.003
OASlonlyfirst.f1	-0.188	0.009	-0.202	0.008	-0.185	0.008	-0.192	0.005
age6.f2 :sex.f2	-0.012	0.011	-0.031	0.011	-0.019	0.011	-0.021	0.006
age6.f3 :sex.f2	0.013	0.010	0.016	0.010	0.017	0.010	0.015	0.006
age6.f4 :sex.f2	-0.005	0.011	-0.007	0.011	-0.005	0.011	-0.005	0.006
age6.f5 :sex.f2	-0.150	0.013	-0.154	0.013	-0.151	0.013	-0.151	0.008
age6.f6 :sex.f2	0.051	0.010	0.051	0.010	0.051	0.010	0.051	0.006
sex.f2 :marital.f2	-0.049	0.003	-0.044	0.003	-0.043	0.003	-0.045	0.002
sex.f2 :marital.f3	-0.047	0.007	-0.039	0.007	-0.043	0.007	-0.043	0.004
sex.f2 :marital.f4	-0.014	0.004	-0.008	0.004	-0.010	0.004	-0.010	0.002

toutes les années combinées et pour une année t , pour un domaine d , est donné par

$$R_d = \frac{\text{MSE}_d}{\text{MSE}_d^t} = \frac{\sigma_u^2 n_d^t + \sigma_e^2}{\sigma_u^2 n_d + \sigma_e^2} \quad (8)$$

$$= \frac{\sigma_u^2 n_d^t + \sigma_e^2}{\sigma_u^2 T n_d^t + \sigma_e^2} \quad (9)$$

$$= \frac{n_d^t + \sigma_e^2 / \sigma_u^2}{T n_d^t + \sigma_e^2 / \sigma_u^2} \quad (10)$$

$$= \frac{1 + \phi_d}{T + \phi_d}, \quad (11)$$

où $\phi_d = \frac{\sigma_e^2 / n_d^t}{\sigma_u^2} = \frac{1}{\gamma_d^t} - 1$, c.à.d. que ϕ_d devient plus grand quand la taille du domaine devient plus petite. De plus, le ratio R_d est une fonction croissante de ϕ_d et donc une fonction décroissante de n_d^t , et ce ratio est compris entre $1/T$ et 1. Ainsi, le gain en termes de la racine des sommes des carrés des erreurs pour un domaine d est

$$G_d = 1 - \sqrt{R_d}, \quad (12)$$

et est donc plus important pour les grandes communes car l'estimateur SAE utilise une plus grande part de l'estimateur direct pour ces communes. Il est au maximum $1 - \sqrt{1/3} \approx 42\%$ pour un pooling de 3 ans et $1 - \sqrt{1/5} \approx 55\%$ pour un pooling de 5 ans.

A titre d'illustration et en se référant à la Table 3, nous avons que $\sigma_e^2 / \sigma_u^2 \approx 1000$. Pour une commune de taille 100, la valeur de $\phi_d^t \approx 10$. Ainsi, pour une telle commune, le gain est alors d'environ 8% pour un pooling de 3 ans et d'environ 15% pour un pooling de 5 ans. Tandis que pour les 7 communes ayant les plus grandes tailles d'échantillon en 2014 (allant de 3352 à 9598 personnes), R_d est proche de $1/T$ et le gain G_d dépasse les 36% pour un pooling de 3 ans et les 50% pour un pooling de 5 ans.

Cependant, il n'est pas certain que l'estimation de la RMSE, qui se situe la plupart du temps aux alentours de 1% (voir figure 2), soit suffisamment stable et précise pour détecter le gain théorique dû au pooling étudié dans cette section. Nous allons étudier cette question dans la section suivante.

4 Instabilité de l'estimation de la RMSE par le PDB

On a vu précédemment que d'un point de vue théorique, le pooling devrait réduire la RMSE. Cependant, le gain théorique n'est pas pleinement constaté en pratique comme l'indique la figure 2. Nous avons étudié cette question en distinguant les communes sans mutation des communes avec mutation d'une part, et d'autre part, les communes publiables des autres communes.

La technique du PDB est spécifique aux problèmes d'estimations sur petits domaines et emploie un estimateur de bootstrap paramétrique, c.à.d. que le ré-échantillonnage se fait au sein d'une population bootstrap construite à partir des paramètres du modèle ajusté. L'estimateur de la MSE par le PDB pour un domaine d est donné par (voir [1, p. 14])

$$\text{MSE}_{\text{PDB}}(\hat{P}_d) = \underbrace{\hat{\gamma}_d \frac{1}{B} \sum_{b=1}^B \left(\hat{P}_d^{*(b)} - \hat{P}_d^{\text{FIX}} \right)^2}_{\text{MSE}_{II}} + (1 - \hat{\gamma}_d) \underbrace{\frac{1}{B} \sum_{b=1}^B \left(\hat{P}_d^{*(b)} - \hat{P}_d \right)^2}_{\text{MSE}_{II}}, \quad (13)$$

où $\hat{\gamma}_d \in [0, 1]$ est facteur de combinaison donné par l'équation (2). $\hat{P}_d^{\text{FIX}}$ est l'estimateur obtenu par l'ajustement du modèle sur l'échantillon poolé avec la commune comme effet fixe et donc

s'apparente à un estimateur direct, alors que $\hat{P}_d$ est obtenu par l'ajustement du modèle sur l'échantillon poolé avec la commune comme effet aléatoire. L'estimateur $\hat{P}_d^{*(b)}$ est obtenu par l'ajustement du modèle avec la commune comme effet aléatoire sur un échantillon bootstrap sélectionné par un tirage simple dans une population bootstrap générée sur la base des paramètres du modèle qui a été ajusté pour estimer $\hat{P}_d$. B est le nombre d'échantillons bootstrap sélectionnés. Pour les résultats présentés dans ce rapport, $B = 250$. Cet estimateur de la MSE est une combinaison linéaire convexe de deux estimateurs d'erreur MSE_I et MSE_{II} pondérés par $\hat{\gamma}$ et $(1 - \hat{\gamma})$, respectivement. Le premier terme MSE_I estime l'erreur même si le modèle ajusté est faux, alors que le deuxième terme MSE_{II} suppose que le modèle ajusté correspond parfaitement à la réalité. Dans ce qui suit, on va se focaliser sur les communes publiables et sans mutation.

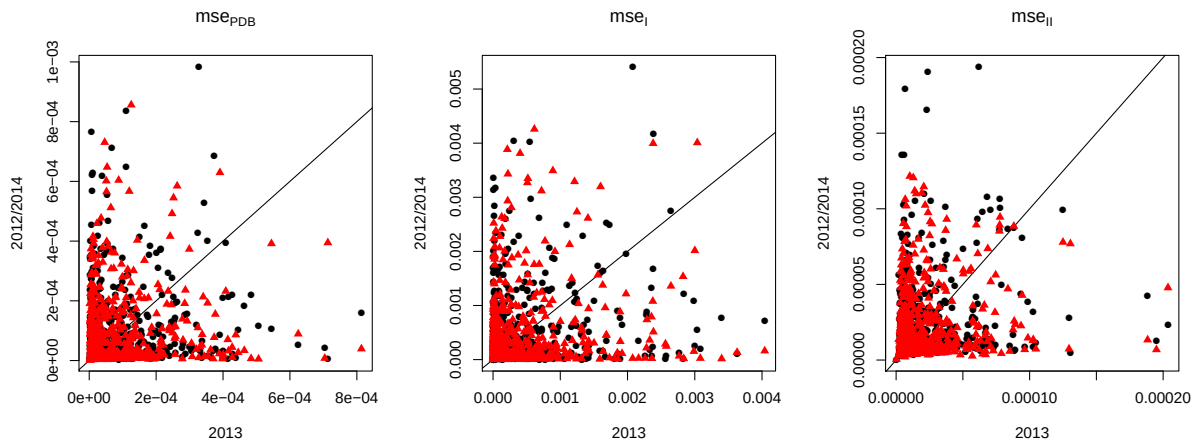


FIGURE 5 Les MSE_{PDB} du taux d'activité au niveau des communes ainsi que les deux termes MSE_I et MSE_{II} de l'équation (13). Les estimations de la MSE de 2012 (points noirs) et de 2014 (triangles rouges) sont comparées à celles de l'année 2013.

Pour mieux comprendre les raisons de l'instabilité de la MSE_{PDB} , nous allons étudier en détail tous les termes la constituant. Dans la figure 5, on observe que les deux composantes MSE_I et MSE_{II} présentent des instabilités considérables à travers les années par rapport à leur échelle. Cependant, l'ordre de grandeur de la MSE_{II} est nettement plus petit par rapport à celui de la MSE_I . La figure 6 montre que l'utilisation d'un modèle à effet aléatoire LMM (panel de droite) apporte une stabilité aux estimations par rapport à un modèle linéaire sans effet aléatoire LM (panel au centre). Ceci a une conséquence directe sur la stabilité des deux termes de la MSE_{PDB} .¹ Enfin, les $\hat{\gamma}_d$ présentent certains signes d'instabilité surtout pour des communes qui ont des petites valeurs $\hat{\gamma}_d$. Ce sont des communes avec une petite taille d'échantillon.

Pour mieux étudier l'impact de la variabilité des estimateurs directs $\hat{P}_d^{FIX}$ sur l'instabilité de la $RMSE_{PDB}$, nous avons remplacé les estimations de $\hat{P}_d^{FIX}$ des années 2012 et 2014 par les estimations de $\hat{P}_d^{FIX}$ de l'année 2013, dans les $RMSE_{PDB}$ des années 2012 et 2014, respectivement. Ensuite, nous avons comparé les $RMSE_{PDB,2013}$ résultantes pour les années 2012 et 2014 à la $RMSE_{PDB}$ de l'année 2013 qui utilise aussi par principe les estimations de $\hat{P}_d^{FIX}$ de l'année 2013, voir figure 7. On observe que les estimations de la RMSE deviennent plus stables à travers les années si les mêmes estimations de $\hat{P}_d^{FIX}$ sont utilisées.

1. Dans certains cas rares, les estimations $\hat{P}_d^{FIX}$ du taux d'activité au niveau des communes sont négatives. Ceci est dû au fait qu'un modèle linéaire (LM) a été préféré à un modèle linéaire généralisé (GLM) pour une raison liée au temps de calcul [2]. Néanmoins, les estimations pour les communes publiables et sans mutation sont toutes

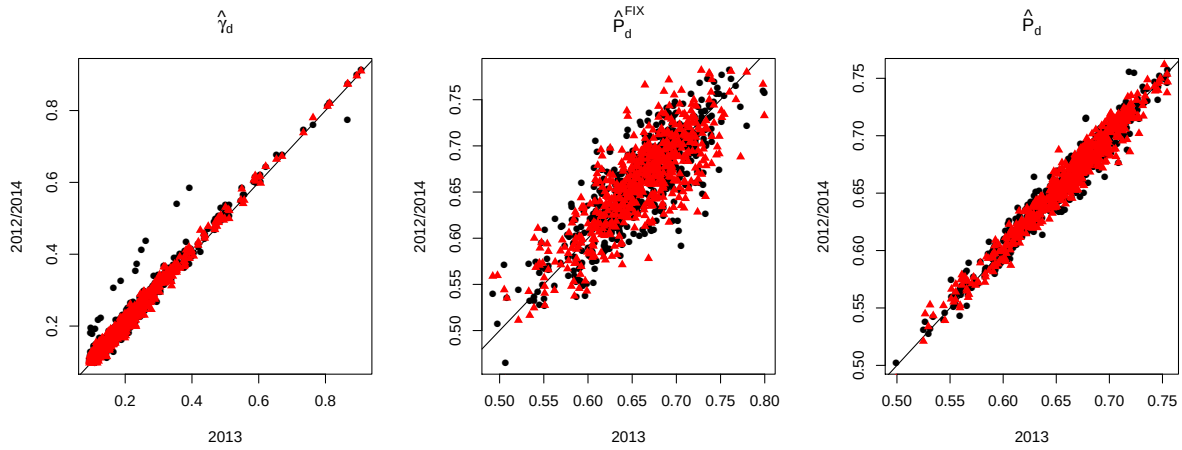


FIGURE 6 Les termes fixes pour chaque estimation bootstrap de la MSE_{PDB} (voir équation 13). Les valeurs de 2012 (points noirs) et de 2014 (triangles rouges) sont comparées à celles de l'année 2013. Panel de gauche : $\hat{\gamma}_d$; panel du centre : $\hat{P}_d^{FIX}$, panel de droite : $\hat{P}_d$.

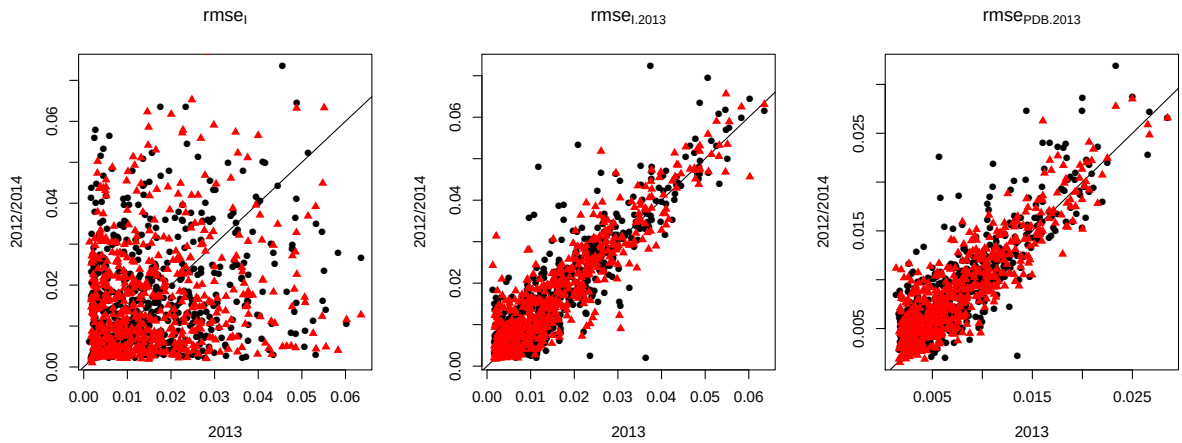


FIGURE 7 Les $RMSE_I$ des estimateurs du taux d'activité au niveau des communes (panel de gauche). Les $RMSE_{I,2013}$ obtenus en utilisant l'estimateur $\hat{P}_d^{FIX}$ de 2013 pour toutes les années 2012 à 2014 (panel du centre). Les estimateurs $RMSE_{PDB,2013}$ obtenus en utilisant comme premier terme les $RMSE_{I,2013}$ de 2013 (panel de droite). Les estimateurs des années 2012 (points noirs) et 2014 (triangles rouges) sont comparées à celles de l'année 2013.

Ces résultats suggèrent que la variance des estimateurs $\hat{P}_d^{FIX}$ qui s'apparente à un estimateur direct pour s'affranchir des éventuels défauts des modèles, est la principale cause de l'instabilité de la MSE (et ainsi de la RMSE) estimée par le PDB. Cette estimation est néanmoins utile comme ordre de grandeur ou pour détecter un problème sévère. A titre d'exemple, les RMSE estimées des communes publiables et sans mutation sont toujours inférieures à 3.5% pour un pooling de 5 ans.

positives comme le montre bien la figure 6.

5 Conclusions

Le pooling de plusieurs années permet d'augmenter le nombre de communes publiables, c.à.d. avec un échantillon qui atteint le seuil de 100 personnes. Avec un pooling de 3 ans, le pourcentage de communes publiables est de 60.7%. Ce pourcentage augmente à 76.4% avec un pooling de 5 ans. Les communes publiables couvrent 94% et 98% de la population cible pour le pooling de 3 ans et celui de 5 ans, respectivement. Par ailleurs, le pooling permet également un gain en termes de RMSE. Plus précisément, le gain théorique est au maximum $1 - \sqrt{1/T}$ pour T années poolées, et il est plus important pour les grandes communes. Cependant, il est moins important pour les petites communes. Ceci peut être expliqué par le fait que les défauts du modèle sont quasiment identiques à travers les années, et que les spécificités d'un domaine ne sont pas bien capturées par un modèle à effets aléatoires si l'échantillon correspondant est très petit. Néanmoins, le biais est quand même réduit grâce au pooling. Notons que dans ce rapport, le 'benchmarking' n'a pas été considéré dans l'étude du gain théorique car il est difficile d'obtenir une formulation analytique de la MSE dans le cas du benchmarking. En pratique, ce dernier a peu d'effet sur le biais et la variance dans ce projet. Pour plus de détails, voir [2]. Du point de vue pratique, le gain théorique dû au pooling n'est pas pleinement constaté sur les RMSE estimées par la méthode PDB proposée par [2]. En effet, les analyses empiriques ont montré une certaine instabilité des estimations de la RMSE à travers les années. Cette évaluation de la précision peut néanmoins être utilisée comme un ordre de grandeur de l'erreur de l'estimation du taux d'activité. Comme dans le cas du gain théorique, le pooling permet d'avoir des estimations plus précises et plus stables de la RMSE.

6 Annexe

Validité de la formule (5) en tenant compte du fait que les β sont estimés

On peut montrer que si l'on tient compte de l'estimation des β , la formule (5) reste valable approximativement :

$$\text{MSE}_d = \gamma_d^2 \frac{S_d^2}{n_d} + (1 - \gamma_d)^2 B_d^2 + O\left(\frac{1}{n}\right), \quad (14)$$

où n désigne la taille de l'échantillon national (et non pas dans le domaine).

L'approximation (14) se justifie de la façon suivante. On désigne par $\tilde{P}_d$ l'estimateur "petits domaines" sous l'hypothèse où les β sont connus, et par $\hat{P}_d$ le même estimateur dans lequel les β sont remplacés par leurs estimations (σ_e et σ_u sont ici encore supposés connus, mais par le même type d'arguments que ceux exposés ci-dessous l'approximation (14) devrait rester valable lorsque l'on tient compte de l'estimation de ces paramètres) :

$$\begin{aligned} \tilde{P}_d &= \gamma_d \tilde{y}_d^{GREG} + (1 - \gamma_d) \bar{X}_d \beta = \gamma_d [\bar{y}_d + (\bar{X}_d - \bar{x}_d) \beta] + (1 - \gamma_d) \bar{X}_d \beta \\ \hat{P}_d &= \gamma_d \hat{y}_d^{GREG} + (1 - \gamma_d) \bar{X}_d \hat{\beta} = \gamma_d [\bar{y}_d + (\bar{X}_d - \bar{x}_d) \hat{\beta}] + (1 - \gamma_d) \bar{X}_d \hat{\beta}, \end{aligned}$$

où $\bar{y}_d$ et $\bar{x}_d$ désignent des moyennes sur l'échantillon, tandis que $\bar{X}_d$ est calculé sur l'ensemble de la population. On a :

$$\begin{aligned} \text{MSE}_d(\hat{P}_d) &= \mathbf{E}(\hat{P}_d - P_d)^2 \\ &= \mathbf{E}\left[\left(\hat{P}_d - \tilde{P}_d\right) + \left(\tilde{P}_d - P_d\right)\right]^2 \\ &= \mathbf{E}(\hat{P}_d - \tilde{P}_d)^2 + \underbrace{\mathbf{E}(\tilde{P}_d - P_d)^2}_{\text{MSE}_d \text{ formule (5)}} + \mathbf{E}\left[\left(\hat{P}_d - \tilde{P}_d\right)\left(\tilde{P}_d - P_d\right)\right]. \end{aligned} \quad (15)$$

Par ailleurs :

$$\begin{aligned} \hat{P}_d - \tilde{P}_d &= \gamma_d (\bar{X}_d - \bar{x}_d) (\hat{\beta} - \beta) + (1 - \gamma_d) \bar{X}_d (\hat{\beta} - \beta) \\ &= [\gamma_d (\bar{X}_d - \bar{x}_d) + (1 - \gamma_d) \bar{X}_d] (\hat{\beta} - \beta). \end{aligned}$$

On a donc :

$$\mathbf{E}(\hat{P}_d - \tilde{P}_d)^2 \leq \left\{ \mathbf{E}[\gamma_d (\bar{X}_d - \bar{x}_d) + (1 - \gamma_d) \bar{X}_d]^4 \mathbf{E}(\hat{\beta} - \beta)^4 \right\}^{\frac{1}{2}}.$$

Comme $\sqrt{n}(\hat{\beta} - \beta)$ et $\sqrt{n_d}(\bar{X}_d - \bar{x}_d)$ convergent en loi vers une loi normale centrée, il vient :

$$\mathbf{E}(\hat{P}_d - \tilde{P}_d)^2 = O\left(\frac{1}{n}\right). \quad (16)$$

De même :

$$\left| \mathbf{E}\left[\left(\hat{P}_d - \tilde{P}_d\right)\left(\tilde{P}_d - P_d\right)\right] \right| \leq \left[\mathbf{E}(\hat{P}_d - \tilde{P}_d)^2 \mathbf{E}(\tilde{P}_d - P_d)^2 \right]^{\frac{1}{2}} = O\left(\frac{1}{n}\right). \quad (17)$$

Les équations (15), (16) et (17) permettent de justifier la formule (14).

7 Références

- [1] Marie Dupraz, Anne Massiani, and Daniel Kilchmann. *Estimation du taux d'activité au niveau communal dans le cadre du relevé structurel*. Number do-f-03.02-SAE-Pooling-01. www.experimental.bfs.admin.ch, Aug 2018.
- [2] Isabel Molina and Ewa Strzalkowska-Kominiak. Estimation of proportions in small areas : application to the labour force using the swiss census structural survey. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 183(1) :281–310, 2020.