Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Statistical Office FSO

Utilisation de méthodes “petits domaines” pour estimer le taux d’actifs selon l’enquête structurelle

Mandat réalisé par l’Université Carlos III de Madrid

Anne Massiani et Christoph Freymond
Office Fédéral de la Statistique

Journées suisses de la statistique publique
Neuchâtel, 16.09.2016

1. Introduction

- L'OFS souhaiterait produire des estimations du taux d'actifs au niveau des districts et / ou des communes, sur la base de l'enquête structurelle.

  Dans certains de ces domaines, la taille de l'échantillon est très petite.

  $\Rightarrow$ Estimateurs directs traditionnels pas performants.

- On dispose d'un riche jeu de variables auxiliaires, provenant de sources basées sur des données administratives.

- Mandat de l'OFS à l'Université Carlos III de Madrid : étudier la possibilié d'exploiter cette information pour estimer de façon fiable le taux d'actifs dans les districts et les communes (petits domaines).

- Au delà de cet objectif spécifique, cela a également été pour l'OFS l'opportunité de :

  - acquérir une compréhension plus profonde des méthodes "petits domaines",

  - identifier certains changements que leur utilisation impliquerait pour l'OFS
    (Communication ? Implémentation ?).

2. Principe de l'estimation "petits domaines"

Estimateurs directs

Les estimations pour un domaine donné sont produites uniquement sur la base des observations de ce domaine.

Exemple : les estimations pour la ville de Neuchâtel ne sont produites qu'avec des informations relevées pour Neuchâtel.

Estimateurs indirects (petits domaines)

- **Modèle** : relation, valable pour tous les domaines, entre la variable d'intérêt et les variables auxiliaires.
- Pour un domaine donné, **on peut utiliser toute l'information disponible pour tous les autres domaines**.

Estimateurs indirects (suite)

- **Sous les hypothèses du modèle :** gain de précision puisque plus d'information est utilisée.

  Mais risque de biais si hypothèses pas respectées !

  $\Rightarrow$ Importance de la validation du modèle.

- Nécessite de l'information auxiliaire pour la totalité de la population, soit au niveau individuel soit sous forme agrégée.

Estimateurs indirects (suite)

- Cohérence entre la somme des estimations dans les petits domaines et l'estimation eu niveau supérieur (national, cantonal) pas toujours automatiquement garantie.

  $\Rightarrow$ benchmarking (facteur d'ajustement).

  Exigence importante pour les offices nationaux.

- Très large palette de méthodes, des plus basiques aux plus sophistiquées.

Exemple 1 : estimateur synthétique post-stratifié

Sur l'exemple d'une étude sur la population active en Lombardie, Italie[1]

- Quantité à estimer : Nombre d'actifs
- Petits domaines : districts industriels
- Quantités estimées sur la base de l'échantillon :

  Proportion $\widehat{P}_g$ d'actifs au sein de post-strates âge-sexe g
- Information auxiliaire connue pour chaque district industriel d : taille N_{dg} des post-strates âge-sexe g

[1] Bartolini, E. (2008), Small Area Estimation and the Labour Market in Lombardy's Industrial Districts: a Mathematical Approach, *Scienze Regionali*, **7**, 27-54.

Exemple 1 (suite)

- Estimation du nombre d'actifs dans district industriel d :

$$\widehat{T}_d^{\text{synth}} = \sum_g \underbrace{\widehat{P}_g}_{\substack{\text{Proportion} \\ \text{dans} \\ \text{post-strate } g}} \cdot \underbrace{N_{dg}}_{\substack{\text{Taille} \\ \text{post-strate } g \\ \text{dans} \\ \text{domaine } d}}$$

- Hypothèse implicite (modèle) :

  Proportion dans post-strate g au sein du domaine d

  $=$

  Proportion dans post-strate g

- Précision
  - Biais : risque de biais si hypothèse pas respectée.
  - Variance : la seule variablité provient d'estimations au sein de post-strates pour la Lombardie toute entière $\Rightarrow$ gain de variance.

Exemple 2 : EBLUP basé sur un modèle linéaire mixte

Modèle

$$\underbrace{y_{id}}_{\text{Variable d'intérêt pour l'individu } i \text{ du domaine d}} = \underbrace{\sum x_{id}^{p}\beta_{p}}_{\text{Effet fixe : part expliquée par les variables auxiliaires } x_{id}} + \underbrace{\nu_{d}}_{\text{Effet aléatoire spécifique au domaine d}} + \underbrace{e_{id}}_{\text{Aléa qui reste inexpliqué}}$$

ν_d iid d'espérance nulle et de variance σ_ν

e_{id} iid d'espérance nulle et de variance σ_e

Estimateur

- Estimateur EBLUP qui en découle : combinaison d'un estimateur direct et d'un estimateur indirect.
- L'information auxiliaire doit être connue pour tous les individus de la population.

Exemple 2 (suite)

Intérêt par rapport à l'exemple 1

- Permet d'intégrer plus d'informations auxiliaires (structure additive remplace croisement)

- Permet de tenir compte de la spécificité des domaines.

3. Objectifs spécifiques du mandat

1. Proposer un estimateur "petits domaines" performant mais raisonnable du point de vue complexité et temps de calcul.

2. Etudier ses performances grâce à des simulations design-based avec de vraies données afin de tenir compte des défauts du modèle : l'enquête structurelle joue le rôle d'une population dont on tire des échantillons.

3. Analyser l'effet du benchmarking.

4. Developper une procédure d'estimation fiable pour la "design MSE" de l'estimateur retenu (mesure de précision tenant compte des défauts du modèle).

4. Description des données disponibles

Echantillon de l'enquête structurelle

- En 2012 : 286'015 personnes de la population résidente permanente en ménage privé agées de 15 ans et plus.
- Variable d'intérêt : variable binaire actif / non actif.
- $D = 2475$ communes représentées dans l'échantillon, 10 non représentées.
- Taille de l'échantillon dans les communes

Taille ech. en dessous de	10	20	30	40	50	100
Nb communes	356	697	964	1173	1337	1792
Prop. communes	0.14	0.28	0.39	0.47	0.54	0.73

Variables auxiliaires connues pour la population

- Statistique de la population et des ménages de 2012

  Genre, âge, état civil, nationalité, taille du ménage, jouissance d'une résidence secondaire, informations géographiques (commune, district, strate, etc.).

- Données AVS 2011

  Contribution ou non à l'AVS pour chaque mois, classes de revenus annuels pour les personnes ayant contribué.

  $\Rightarrow$ Décalage temporel entre ces données et l'échantillon.

Ces variables et certaines de leurs interactions sont utilisées pour ajuster un modèle linéaire mixte généralisé.
$\Rightarrow$ Modèle très performant (taux d'erreur de classification de 11%, pour un taux d'actifs de 68%).

5. Résultats des simulations

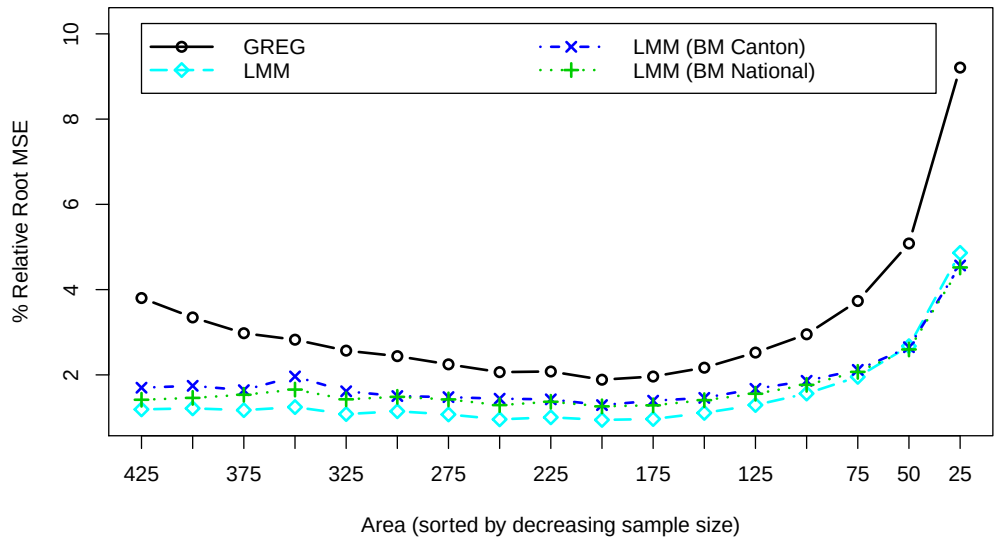
Résultats des simulations : objectif 1

- Dans notre situation, l'EBLUP basé sur un modèle linéaire mixte (exemple 2) a des performances similaires à d'autres estimateurs plus complexes (estimateurs basés sur un modèle linéaire mixte généralisé en théorie mieux adaptés à la modélisation d'une variable binaire).

- Parmi tous les estimateurs "petits domaines" considérés, l'EBLUP basé sur un modèle linéaire mixte (LMM) est choisi. Ses performances sont comparées à celles du GREG, qui est l'estimateur direct actuellement utilisé.

Résultats des simulations : objectif 2

- Les estimateurs “petits domaines” réduisent considérablement la design-MSE.



Design-RRMSE moyenne pour des classes de taille d’échantillon indiquées sur l’axe des x

- Risque limité de biais des estimateurs “petits domaines” Biais sous le plan inférieur à 5% si la taille de l’échantillon dans le domaine est supérieure à 20.

Résultats des simulations : objectifs 3 et 4

- Objectif 3 : faible perte d'efficacité à cause du benchmarking (voir graphe slide précédent).

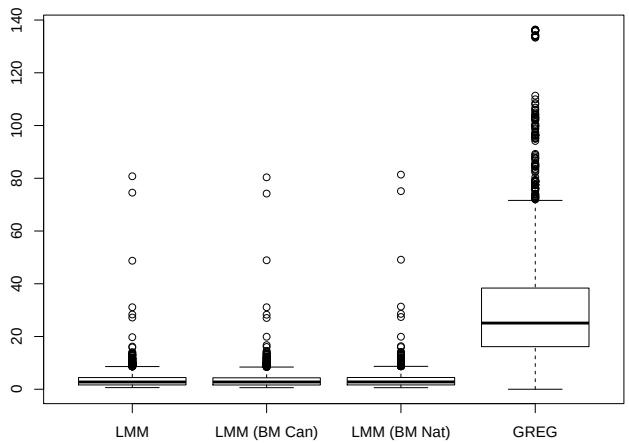
- Objectif 4 : le développement d'une procédure d'estimation fiable pour la "design MSE" a été l'objet d'une recherche très innovante de la part de l'Université Carlos III de Madrid.

  $\Rightarrow$ Procédure totalement nouvelle qui permet dans notre situation d'estimer de façon fiable la "design MSE" si la taille de l'échantillon dans le domaine est supérieure à 100.

6. Application réelle

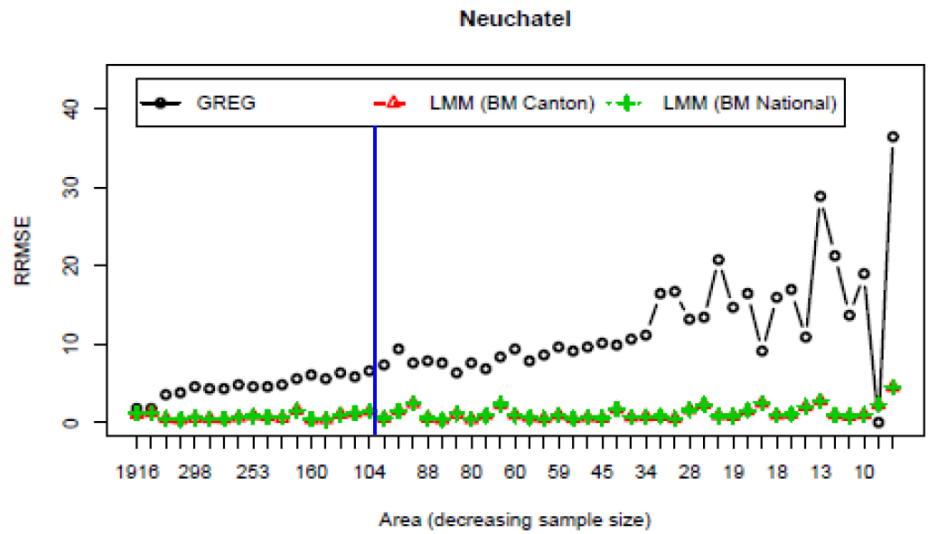
Echantillon beaucoup plus grand que dans les simulations.

$\Rightarrow$ les gains sont aussi plus importants : en comparaison avec le GREG, l'estimateur "petits domaines" proposé permet une reduction médiane de la RRMSE de 90% pour les communes, 49% pour les districts.



Boxplots de la longueur des intervalles de confiance basés sur la design-RMSE, niveau communes

L'exemple des communes du canton de Neuchâtel



RRMSE estimées pour le GREG et pour les estimateurs "petits domaines"

Ligne verticale bleue : seuil pour une estimation fiable la design-MSE des estimateurs "petits-domaines".

7. Conclusions

L'étude de l'Université Carlos III de Madrid est très utile pour l'OFS, qui mène des réflexions concernant une strategie en vue d'une éventuelle utilisation des méthodes "petits domaines".

Analyses complémentaires, par exemple évolutions.

Diffusion des résultats

- communication sur la méthode encore peu usuelle dans les statistiques officielles;
- choix d'une taille d'échantillon en dessous de laquelle aucune estimation n'est publiée, de façon à limiter le risque de biais ou de sous-estimation de la design-MSE.

Conclusions (suite)

Généralisation à d'autres variables ? D'autres enquêtes ?

Point délicat : pour chaque variable d'intérêt, il faut développer un modéle spécifique et performant.

Software: le logiciel standard à l'OFS est SAS mais certaines méthodes "petits domaines" sont implémentées en R (les deux logiciels doivent être combinés).