



# Utilisation de méthodes « petits domaines » et statistiques expérimentales

Estimation du taux d'actifs sur les données d'un « pooling » sur trois années de l'enquête structurelle

Marie Dupraz, Anne Massiani, Christoph Freymond, Daniela Wanner

Office fédéral de la statistique

Journées suisses de la statistique | 29.08.2018



# PARTIE 1 : ASPECTS METHODOLOGIQUES



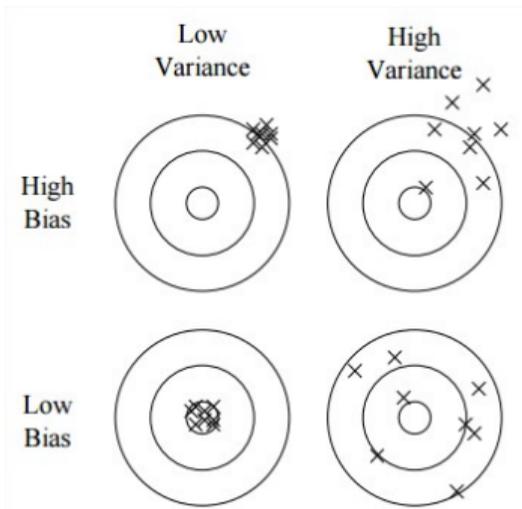
# 1. Introduction

- ▶ **But** : estimation du taux d'actifs au niveau communal pour le relevé structurel. Taille d'échantillon souvent faible  $\Rightarrow$  recours aux méthodes de type "Small Area Estimation" (SAE).
- ▶ **Première étape : mandat confié à l'Universidad Carlos III de Madrid (UC3M)**  
 $\Rightarrow$  Sous certaines conditions, estimations annuelles fiables pour 2012 (JSS 2016).
- ▶ **Deuxième étape : suite des travaux à l'OFS**
  - ▶ rapatriement du savoir-faire à l'OFS : production d'estimations annuelles (2012-2013-2014), étude de leur stabilité.
  - ▶ Pooling sur trois ans et étude de l'impact sur les résultats.



## 2. Principes des SAE et résultats du mandat confié à l'UC3M

Deux notions clés pour comprendre les enjeux des SAE : le biais et la variance d'un estimateur





## L'estimation SAE se base sur :

- ▶ des variables auxiliaires connues pour toute la population,
- ▶ un modèle entre la variable d'intérêt et ces variables auxiliaires.

Sous les hypothèses du modèle : gain de variance par rapport aux estimateurs "traditionnels" (plus d'information est utilisée).

Mais risque de biais si hypothèses pas respectées !

## Exemple : principaux résultats obtenus pour le relevé structurel

- ▶ Estimations fiables des taux d'actifs et de leur précision pour les communes ayant un échantillon d'au moins 100 personnes.
- ▶ Gain médian en précision : 78%.
- ▶ Cela équivaut à multiplier la taille de l'échantillon par 8.



## Comparaison estimation “traditionnelle” et SAE en termes de précision

	Biais	Variance
Estimateur traditionnel	En principe, biais faible	Dépend de la taille de l'échantillon dans le domaine
Estimateur SAE	Risque plus ou moins important en fonction de la validité du modèle	En général faible

Mesure de précision considérée pour évaluer les estimateurs :

$$\text{Design Mean Squared Error (MSE)} = \text{Variance} + \text{Biais}^2$$

ou

$$\text{RMSE} = \sqrt{\text{MSE}} \text{ (même unité que l'estimation ponctuelle)}$$

Estimation de la précision extrêmement complexe !



### 3. Modèle

#### ► Sources de données

1. Données d'enquête : relevé structurel ;
2. Informations auxiliaires : Statistique de la population et des ménages (STATPOP) et données AVS de l'année précédente.

#### ► Modèle : Linear Mixed Model avec effet aléatoire pour la commune.

#### ► Estimateur qui en découle : combinaison d'un estimateur de type "traditionnel" et d'un estimateur synthétique (i.e. basé sur un modèle linéaire "classique" sans effet spécifique pour la commune).

Plus la taille de l'échantillon dans la commune est grande, plus la part attribuée à l'estimateur "traditionnel" est importante (diminution du risque de biais).



## 4. Impact du pooling sur le nombre de communes publiables

La fiabilité des résultats (estimation et précision) est jugée satisfaisante pour les communes disposant d'une taille d'échantillon d'au moins 100 personnes.

⇒ Limitation du nombre de communes pour lesquelles les résultats sont publiables.

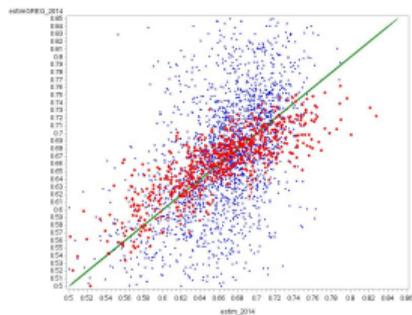
- ▶ Estimations annuelles : environ 30% des communes.
- ▶ Pooling sur trois ans : environ 60% des communes.
- ▶ Pooling sur cinq ans : environ 70% des communes ?



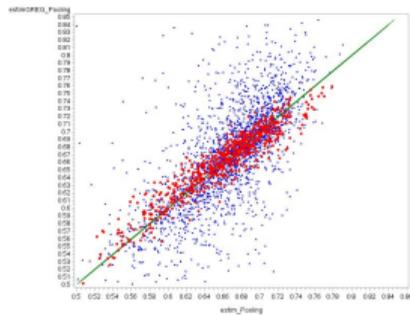
## 5. Impact du pooling (3 ans) sur le risque de biais

Graphiques : estimations (ponctuelles) SAE versus estimations traditionnelles.

- ▶ Permet de détecter un éventuel biais des estimations SAE si les points ne sont pas uniformément répartis autour de la droite  $y = x$ .
- ▶ Points rouges : communes dont l'échantillon annuel atteint les 100 personnes.



2014



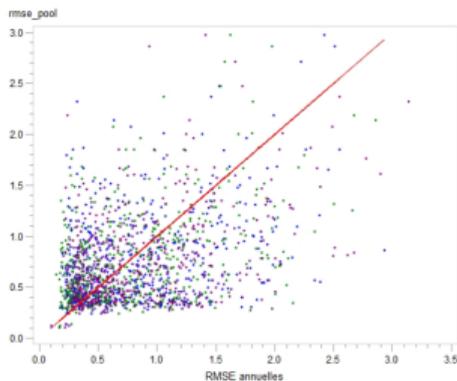
Pooling 2012-2014

- ▶ Estimation "traditionnelle" plus stable avec le pooling.
- ▶ Plus la taille de l'échantillon augmente, plus l'estimateur SAE se rapproche d'un estimateur "traditionnel".



## 6. Impact du pooling (3 ans) sur l'estimation de la précision

RMSE annuelles vs RMSE pooling. Une couleur par année.



Cette question :

- ▶ doit encore être approfondie,
- ▶ souligne la complexité de l'estimation de la (design)-RMSE.

Les points sont répartis autour de la droite  $y = x$  : pas d'amélioration flagrante pour l'**estimation** de la précision.

Hypothèses :

- ▶ soit le pooling améliore peu la précision,
- ▶ soit méthode d'estimation de la RMSE n'est pas assez précise pour détecter l'effet du pooling sur la RMSE.



## 7. Avantages et limitations de l'utilisation d'un pooling

### Avantages

- ▶ Augmentation du nombre de communes pour lesquelles les résultats sont publiables.
- ▶ Impact positif sur le risque de biais.

### Inconvénients

- ▶ Définition de la population : population moyenne, mutations de communes.
- ▶ Lissage des évolutions par l'utilisation d'une population moyenne.

A cela s'ajoute les avantages et limitations relatifs aux méthodes SAE : gain conséquents possibles en termes de précision, mais choix et la validation d'un modèle, complexité de l'estimation de la design-RMSE, défi en termes de communication, etc.



## 8. Perspectives méthodologiques

- ▶ Pooling sur 5 ans ?
- ▶ Améliorations ou adaptations du modèle, par exemple intégration de nouvelles variables auxiliaires ?
- ▶ Investigations de l'effet du pooling sur l'estimation de la design-RMSE.
- ▶ Prise en compte des feed-backs reçus suite à la publication des statistiques expérimentales.

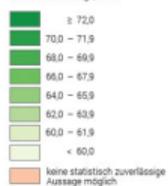


# PARTIE 2 : UTILISATIONS POTENTIELLES, CONCLUSIONS ET PERSPECTIVES

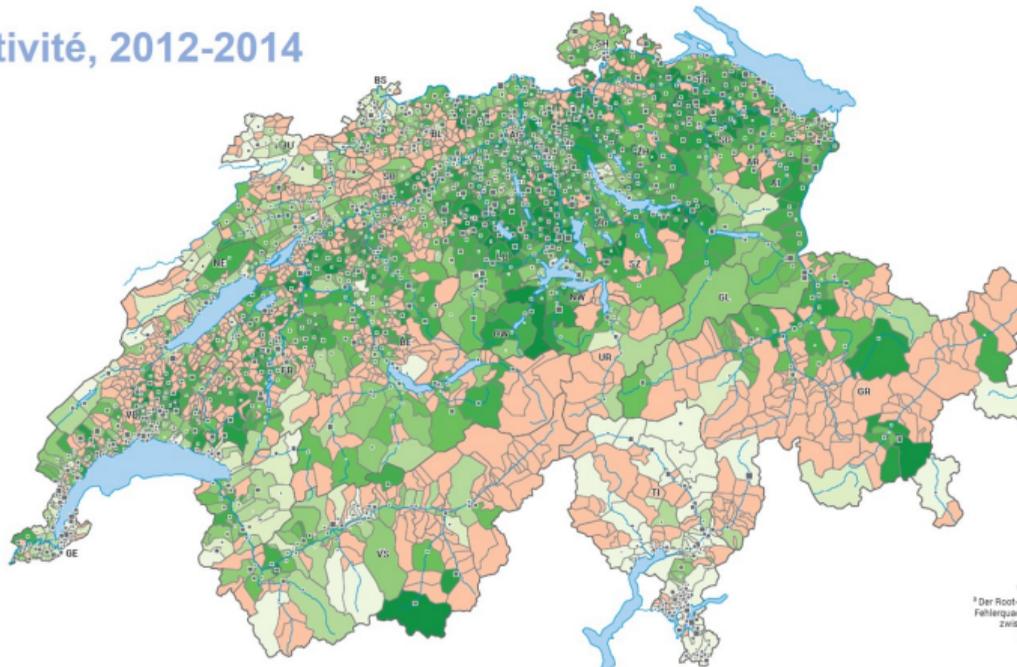


## Taux d'activité, 2012-2014

Anteil Erwerbspersonen an  
der Bevölkerung<sup>1</sup>, in %

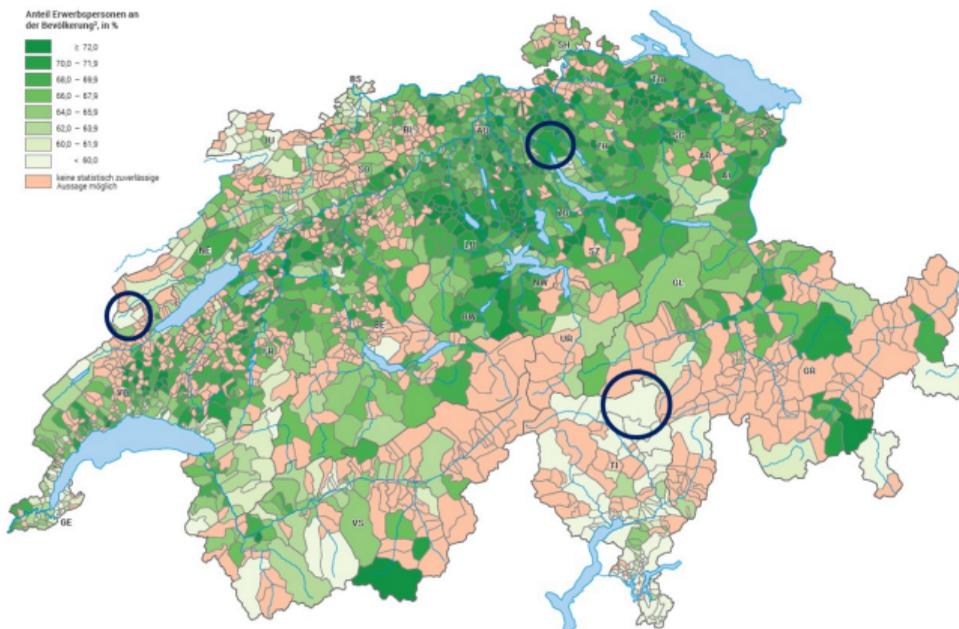


Design Root-mean-squared  
error<sup>2</sup> (\*100)



<sup>1</sup> Ständige Wohnbevölkerung ab 15 Jahren  
<sup>2</sup> Der Root-mean-squared Error (Wurzel der mittleren Fehlerquadratsumme) ist ein Mass, um die Differenz zwischen den Schätzungen eines Modells und beobachteten Messungen auszudrücken.





	Blenio	Sainte-Croix	Zürich
Taux d'activité standardisé estimé (en %) selon modèle	54	54	71
design-RMSE	1.11	0.52	0.11
Habitants 2014	1'726	4'766	391'359
dont (en %)			
0-14 ans	12	15	13
15-64 ans	62	61	71
65 ans et plus	27	24	16
Observations pooling 2012-2014 du RS	269	711	28'558
Surface totale (en km <sup>2</sup> )	202.2	39.3	87.9
dont surface habitable (en %)	1.3	7.0	61.9



## Statistiques expérimentales

### De quoi s'agit-il ?

- ▶ à côté de la production statistique normale
- ▶ nouvelles méthodes / sources de données
- ▶ pas encore entièrement mûres

### Buts ?

- ▶ implication de nos partenaires plus tôt
- ▶ prise en compte optimale des besoins
- ▶ réagir de manière plus agile

### Signification pour ce projet ?

1. Tests avec un pooling sur 5 ans
2. des résultats et recommandations supplémentaires
3. Transfert dans la production statistique courante ?



## Le voyage continue ...

[www.experimental.bfs.admin.ch](http://www.experimental.bfs.admin.ch)

... avec vous tous

**EXPERIMENTAL STATISTICS**