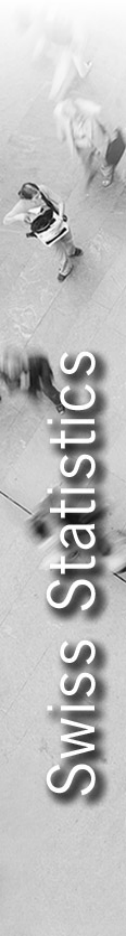
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

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



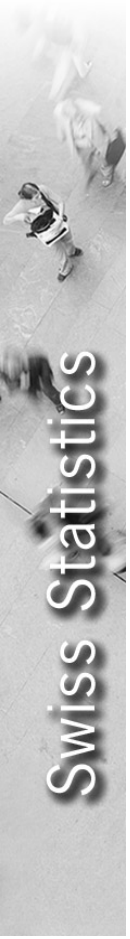
Analyse du processus de préparation statistique des données du relevé structurel du recensement fédéral de la population

Journées suisses de la statistique publique 2017, Chartreuse d'Ittingen 20-21 novembre 2017

Christian Panchard

Section des Méthodes statistiques, Office fédéral de la statistique

Plan



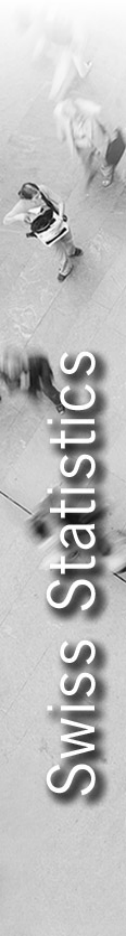
Introduction

Implémentation de l'analyse du RS-PPSD

Résultats de l'analyse

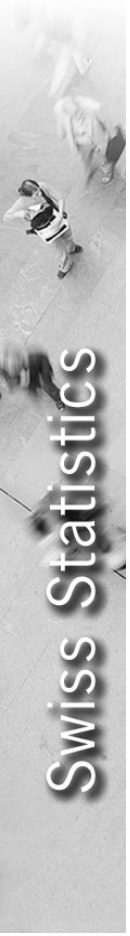
Évolution des indicateurs

Conclusions et perspectives



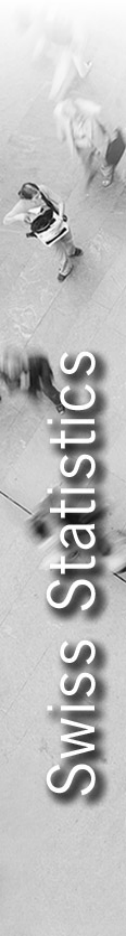
Introduction

Recensement fédéral de la population (depuis 2010)



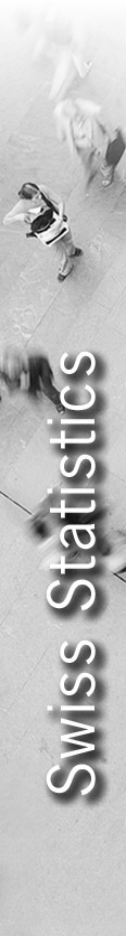
- Recensement basé sur des registres complétés par des enquêtes par échantillonage :
  - registres des habitants des communes et des cantons, registres fédéraux de personnes ;
  - registre fédéral des bâtiments et des logements ;
  - **relevé structurel (RS)** ;
  - deux enquêtes par échantillonage sur des thèmes spécifiques.
- RS, échantillon d'environ 300'000 personnes.

Processus de préparation statistique des données

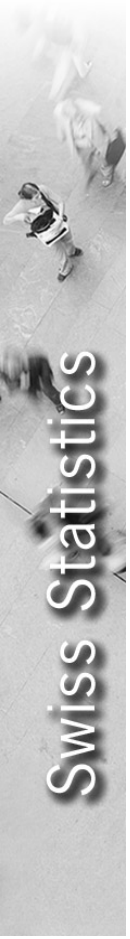


- **Processus de préparation statistique des données du RS (RS-PPSD)** suivant les recommandations du manuel EDIMBUS-RPM, ([Luzi, O. et al.(2007)]).
- Plusieurs phases.
- Plusieurs états des données produits et archivés.

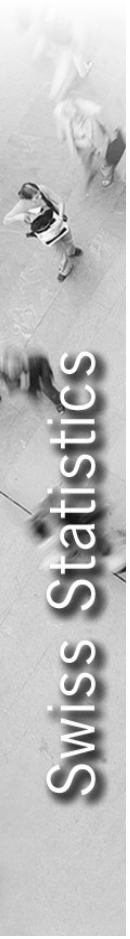
Analyse du RS-PPSD



- Indicateurs selon le manuel EDIMBUS-RPM ([Luzi, O. et al.(2007)]) et des rapports de qualité de Eurostat, [Quality team of Eurostat(2014)].
- Implémentés pour le RS 2013 en collabaration avec la Haute École Spécialisée de Suisse orientale, ([Hulliger, B. et Berdugo, J. D.(2015)]), et présentés aux Journée suisse de la statistique publique 2016, [Hulliger, B. et Kilchmann, D.(2016)].
- Implémentation appliquée au RS 2014.



- Buts :
  - évaluation de la qualité des données ;
  - évaluation de l'impact des traitements sur les résultats ;
  - détection d'améliorations envisageable pour le PPSD ;
  - mise en évidence de problèmes éventuels dans la conception du questionnaire ;
  - contrôle du PPSD au cours du temps ;
  - évaluation de la stratégie d'implémentation pour le RS.



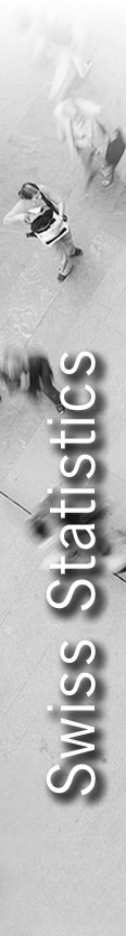
Implémentation de l'analyse du RS-PPSD

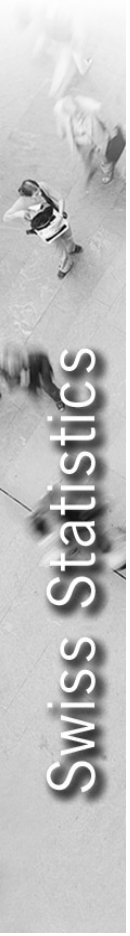
Cadre

Quatre états des données RS à analyser :

D_0		D_1		D_2		D_3
données brutes	$\rightarrow$	*données après changements consécutifs aux rappels téléphoniques*	$\rightarrow$	*données après imputations déterministes*	$\rightarrow$	*données après imputations par plus proche voisin suivies de traitements déterministes*

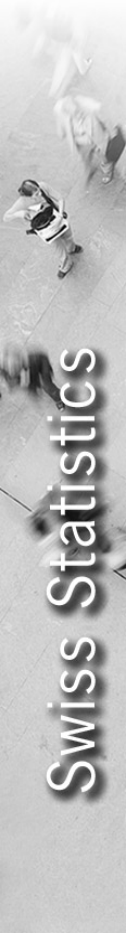
L'état D_3 est utilisé pour les publications.





- Variables se rapportant à des cases à cocher (variables catégorielles) :
  - situation professionnelle, formations achevées, statut d'activité et langue habituelle ;

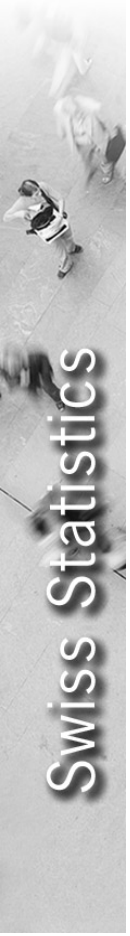
  ou à des nombres (variables numériques) :
  - loyer.
- Les imputations comprennent des imputations de valeurs manquantes et des changements de valeurs existantes.



- Indicatrices utilisées pour le calcul des indicateurs PPSD :
  - indicatrice de réponse r_{ij} (r_{0ij} pour D_0),
  - indicatrice de valeur structurellement manquante b_{ij},
  - indicatrice d'imputation g_{ij},
  - indicatrice d'imputation g_{13ij} pour toute imputation au cours du PPSD.

    où i = indice des observations et j = indice des variables.
- Ces indicateurs PPSD peuvent être pondérés ou non.

Implémentation

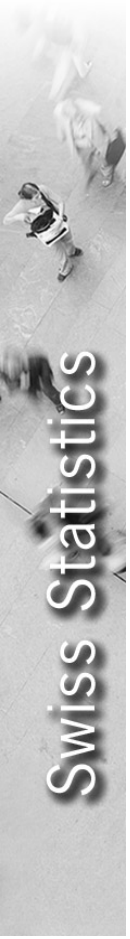


- Les indicateurs sont implémentés dans un paquet R 'sdap', développé par Hulliger et Berdugo ([Hulliger, B. et Berdugo, J. D.(2015)]).

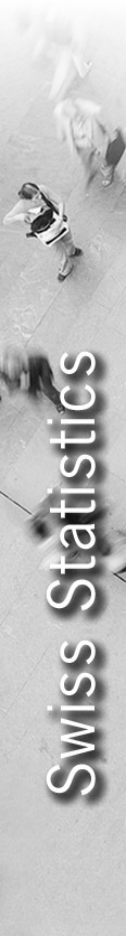
- Ce paquet peut être installé depuis la source

```
http://R-Forge.R-project.org
```

```
install.packages("sdap", repos="http://R-Forge.R-project.org").
```



Résultats de l'analyse



Dans ce qui suit on utilise :

- la version non pondérée des indicateurs pour un **suivi du processus** ;
- la version pondérée de ces indicateurs pour une **analyse de l'impact sur les résultats**.

Indicateur *IRR* (Item response rate)

- Donne la proportion de valeurs non manquantes.

État des données	D_0	D_1	D_2	D_3	D_3 avec r_{0ij}
Situation professionnelle	0.973	0.983	0.983	1.000	0.970
Formations achevées	0.968	0.992	0.995	1.000	0.971
Statut d'activité	0.960	0.993	0.996	1.000	0.963
Langue habituelle	0.990	0.997	1.000	1.000	0.993
Loyer	0.872	0.929	0.892	1.000	0.884

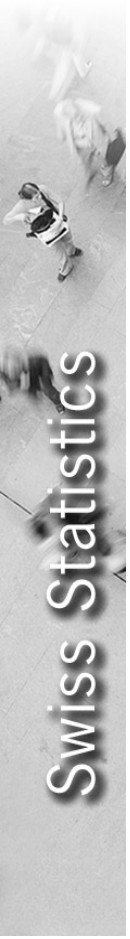
- Globallement proche de 100% et augmente au cours du PPSD.
- Plus bas pour le loyer (parfois difficilement récupérable ou même inconnu).

Indicateur *IRO* (Item response ratio)

- Donne la proportion du total provenant des valeurs manquantes, à l'exclusion de celles structurellement manquantes.

Pour le loyer :

- Environ 80.7% pour D_3 en ne considérant que les répondants initiaux $r_0ij = 1$.
- Proportion non négligeable du total provenant des imputations.
- En relation avec un taux de réponse plus bas et la détection de valeurs aberrantes.



Indicateur *IMRO* (Imputation ratio)

- Donne la proportion du total provenant des valeurs imputées, à l'exclusion de celles structurellement manquantes.

État des données	D_1	D_2	D_3	D_3 avec g_{13ij}	D_3 avec r_{0ij} et g_{13ij}
Loyer	0.061	0.013	0.174	0.251	0.058

- Environ 5.8% du total dû aux imputations provenant de la détection de valeurs aberrantes et de la correction d'erreurs structurelles comme les erreurs de scanning.
- Impact probablement surestimé (le montant total $\hat{y}$ est pris en compte pour les changements, mêmes si ces derniers sont petits).

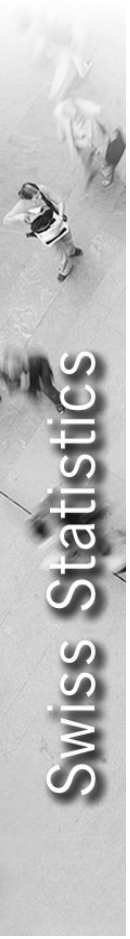
Indicateur *SMR* (Structural missingness rate)

- Donne la proportion des valeurs structurellement manquantes.

Version pondérée :

État des données	D_0	D_1	D_2	D_3
Situation professionnelle	0.394	0.385	0.380	0.367
Loyer	0.452	0.470	0.406	0.412

- Proche de 40%.
- Personnes non actives ou locataires, confirmant nos attentes en ce qui concerne D_3.



Évolution des indicateurs

Détection d'évolutions inhabituelles

- $\Delta = i_{2014} - i_{2013}$, $i_y =$ un indicateur pour l'année y.
- Première idée : calcul d'un intervalle de confiance de type Wald pour les variables catégorielles (voir [Agresti(2013)], page 71)

$$[\Delta \pm 1.96 \times \sigma(\Delta | CSS_{2013}, CSS_{2014})],$$

with $\sigma(\Delta | CSS_{2013}, CSS_{2014}) = \sigma(i_{2014} - i_{2013})$

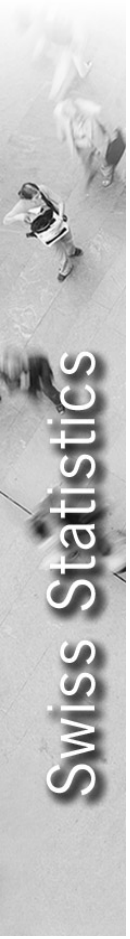
$$= \sqrt{\frac{i_{2014}(1-i_{2014})}{n_{2014}} + \frac{i_{2013}(1-i_{2013})}{n_{2013}}}.$$

- Toujours à l'étude, en particulier pour les indicateurs pondérés et les variables numériques.

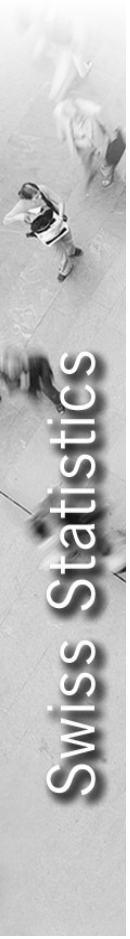
Évolution pour D_1

		2013	2014	Δ	CI	
IRR	Situation professionnelle	0.983	0.983	0.000	-0.001	0.000
	Formations achevées	0.993	0.992	-0.001	-0.001	-0.001
	Statut d'activité	0.994	0.993	-0.001	-0.001	0.000
	Langue habituelle	0.998	0.997	0.000	-0.001	0.000

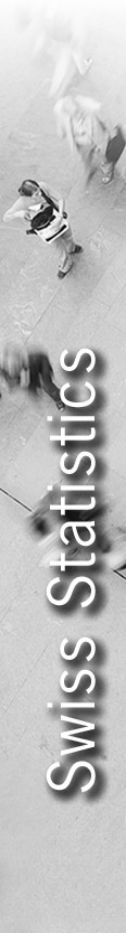
- Petites différences.
- Résultats similaires pour les autres états des données et les autres indicateurs.
- RS-PPSD stable au cours du temps.



Conclusions et perspectives



- Pour les variables autres que le loyer, la qualité du RS-PPSD peut être considérée comme bonne :
  - pour le loyer, un léger potentiel d'optimisation possible, avec des effets marginaux ;
  - pas de nécessité d'adapter le PPSD, aussi en raison de l'automatisation élevée.
- Des améliorations potentielles mineures au questionnaire et au PPSD ont pu être détectées et ont été implémentées depuis 2015.
- Le RS-PPDS est stable au cours du temps.



- Les indicateurs montrent bien quelles étapes du PPSD ont l'impact le plus important.
- Le suivi de l'évolution est encore à l'étude.
- La stratégie d'implémentation choisie pour le RS-PPSD a été un succès et a montré sa capacité à contrôler le processus, ainsi que la possibilité de surveillance de ce processus.
- L'analyse, ainsi que ses indicateurs, peut être implémentée pour d'autres PPSD.

References

A Agresti.
Categorical Data Analysis.
Wiley, 2013.
ISBN 2012009792.

Hulliger, B. et Berdugo, J. D.
Analysis of the Statistical Data Preparation Process of the Swiss Structural Survey 2013.
Technical report, Swiss Federal Statistical Office, 2015.

Hulliger, B. et Kilchmann, D.
Monitoring statistical data preparation.
Présentation au journées suisses de la statistique publique, Neuchâtel, Suisse, 2016.
URL `https://statistiktage.ch/images/pdfs/2016_konferenzbeitr%C3%A4ge/3jss2016_monitoring_statistical_data_preparation.pdf`.

Luzi, O. et al.
EDIMBUS-RPM.
Eurostat, August 2007.
URL http://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross-sectional-business-sur pdf.

Quality team of Eurostat.
ESS Guidelines for the Implementation of the ESS Quality and Performance Indicators (QPI).
European Commission, Eurostat, 2014.
URL http://ec.europa.eu/eurostat/documents/64157/4373903/02-ESS-Quality-and-performance-Indicators-2014.pdf.