Regional estimates of poverty indicators based on a calibration technique

Pascal Ardilly

Insee - Département des méthodes statistiques

Objective

To carry out yearly **regional** estimates for 6 poverty indicators (European request)

$$\hat{\theta} = \frac{\sum_{k \in s} w_k \cdot Y_k}{\sum_{k \in s} w_k}$$

Y_k (k = individual label) dummy variable associated with 4 poverty dimensions :

- At-risk-of-poverty indicator $\rightarrow$ $\hat{\theta}_1$
- Material deprivation (3 items) $\rightarrow$ $\hat{\theta}_2$
- « Severe » material deprivation (4 items) $\rightarrow$ $\hat{\theta}_3$
- Low work intensity $\rightarrow$ $\hat{\theta}_4$
  (specific household population)

And two extra synthetic indicators :

$\hat{\theta}_5$ = AROPE indicator ;

$\hat{\theta}_6$ = share of the population simultaneously in the 3 conditions linked to $\hat{\theta}_1$, $\hat{\theta}_3$ and $\hat{\theta}_4$.

$$\hat{\theta}_6 \leq \left\{\hat{\theta}_1, \hat{\theta}_3, \hat{\theta}_4\right\} \leq \hat{\theta}_5$$

Data source

EU-SILC : French version

- Complex design, rotative sample (9-years panels)
- Frame = census
- 10 500 to 11 000 responding individuals each year

All the individuals in a given household have the same weight … and the same Y_k (for the present issue).

(Very) small regional sample sizes - except maybe for Ile-de-France :

SILC : responding people sample size.
Example for a few regions

REGIONS	2009	2010
Ile-de-France	1 682	1 729
Bourgogne	315	321
Nord Pas-de-Calais	723	788
Franche-Comté	265	261
Aquitaine	637	679
Limousin	157	166
Rhône-Alpes	850	907

Where is the problem ?

Sampling error = Bias^2 + Variance

$$E(\hat{\theta}-\theta)^2 = \left(E\hat{\theta}-\theta\right)^2 + E\left(\hat{\theta}-E\hat{\theta}\right)^2$$

Traditional practice to estimate $\theta_{région}$:

Bias = 0

Variance of $\hat{\theta}_{région}$: depends on $\dfrac{1}{n_{région}}$

$n_{région}$ (too) small $\Rightarrow$ (too) large error

So, we use a MODEL to circumvent the problem : the idea is **to use the <u>national</u> sample to carry out the <u>local</u> estimates**.

Bias $\neq$ 0

Variance of $\hat{\theta}_{région}$: order of magnitude $\dfrac{1}{n_{National}}$

We hope to reduce the total sampling error !!!

Year 2010 - At-risk-of-poverty indicator

REG	Direct estimate	Complete data source RDL
11	**10.73**	**12.50**
21	14.88	14.61
22	**20.09**	**14.43**
23	13.64	12.99
24	11.15	11.78
25	**9.40**	**13.31**
26	13.86	12.52
31	18.65	18.55
41	**16.99**	**13.88**
42	12.80	11.29
43	13.94	12.85
52	**9.12**	**11.17**
53	**13.53**	**11.17**
54	14.53	13.83
72	12.86	12.93
73	14.78	13.97
74	**18.40**	**14.66**
82	**9.54**	**11.82**
83	14.05	14.00
91	18.15	18.63
93	14.56	15.75
94	**28.35**	**19.32**

Applied methodology

Model = hypothesis $\Rightarrow$ error

$\Downarrow$

The <u>relationship</u> between poverty and a set of explicative variables does not depend on the region.

Formalisation :

k : household label (identifier)

Y_k = dummy variable, associated with the poverty concept, known from SILC questionnaire

$X_k \in R^p$ = auxiliary variables, « explain » Y_k,
whose « true » regional total is known
thanks to external and complete sources
(the 'constante' variable is included).

It is **ALWAYS** possible to define B_{NAT} such that

$$Y_k = B_{NAT}^t . X_k + \varepsilon_{k,NAT}$$

with $\sum_{k \in U} \varepsilon_{k,NAT} = 0$ and $\sum_{k \in U} \varepsilon_{k,NAT}^2$ minimum

$U = U_{NAT}$ = <u>national</u> population of households.

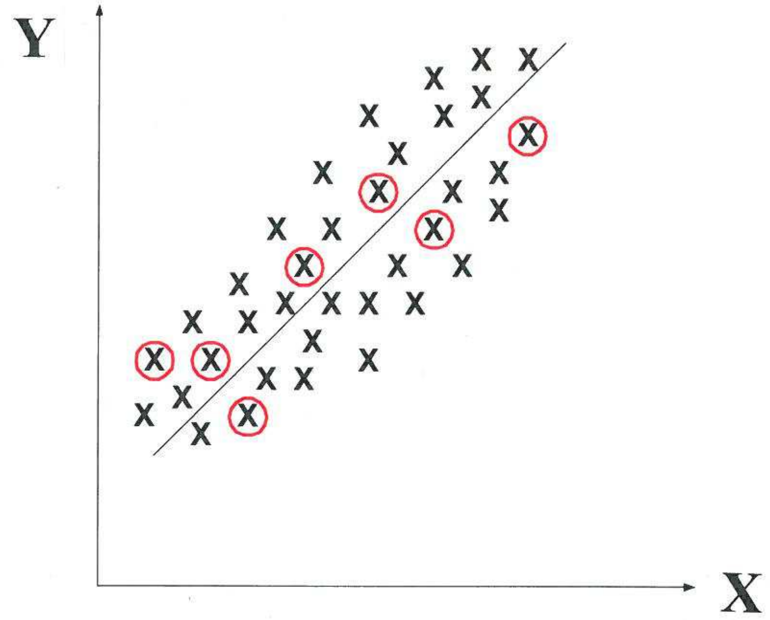
Idem to define B_{REG} when $U = U_{REG}$ = regional population of households.

Remark : fortunately, Y_k may be a 0 - 1 variable (categorical variable) !

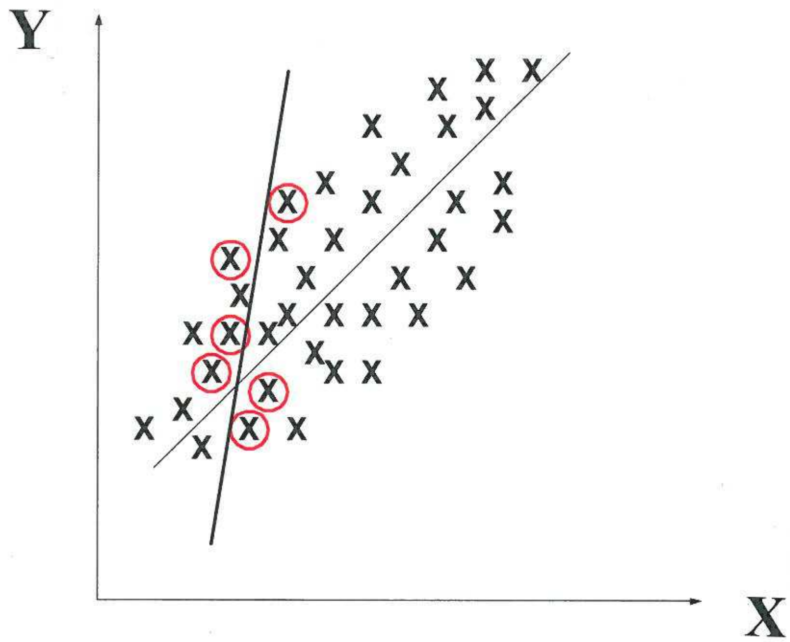
The model is equivalent to the following **hypothesis :**

This relationship is the same in each region, I mean

- either you consider $B_{NAT} = B_{REG}$
- or you consider $\sum_{k \in U_{REG}} \varepsilon_{k,NAT} = 0$



Contexte 1 : cas favorable



Contexte 2 : cas défavorable

So $$Y_{REG} = \sum_{k \in U_{REG}} Y_k = B_{NAT}^t \cdot \sum_{k \in U_{REG}} X_k$$

$$\Rightarrow \boxed{\hat{\hat{Y}}_{REG} = \hat{B}_{NAT}^t \cdot \sum_{k \in U_{REG}} X_k = \hat{B}_{NAT}^t \cdot X_{REG}}$$

with

$$B_{NAT} = \left(\sum_{k \in U_{NAT}} X_k \cdot X_k^t \right)^{-1} \cdot \left(\sum_{k \in U_{NAT}} X_k \cdot Y_k \right)$$

$$\hat{B}_{NAT} = \left(\sum_{k \in s} w_k \cdot X_k \cdot X_k^t \right)^{-1} \cdot \left(\sum_{k \in s} w_k \cdot X_k \cdot Y_k \right)$$

s = **NATIONAL** sample SILC $\Rightarrow$ $\hat{B}_{NAT}$ has a (very) small variance.

$\hat{\hat{Y}}_{REG}$ is called a « **synthetic** » estimate of Y_{REG}.

$$\boxed{BIAS = 0 \Leftrightarrow B_{NAT} = B_{REG}}$$

A *sufficient* condition to get a small bias is :
$\varepsilon_{k,NAT}$ close to zero.

But it is not a *necessary* condition, the model has not to be « good », it means that the $\varepsilon_{k,NAT}$ may be large !!!

However, there is a significant **BIAS** if you miss an important explicative variable of poverty and if the correlation between this variable and X_k depends on the region.

We can use (essentially) 2 tools to assess the bias. **Caution : they do not constitute a « proof »** - but just a suspicion : **we have anyway to take a risk !**

Which auxiliary information for the model?

The most correlated with poverty !

I used 3 data sources.

Any individual variable is first aggregated at the household level : the statistical unit is the household.

a) The census

- sex
- age (6 modalities)
- diploma (4 modalities)
- nationality (5 modalities)
- social category - CS (11 modalities)
- live / or not in a ZUS (‘sensitive’ urban area)
- urban unit category (3 modalities)
- household type (5 modalities)
- rent / or not the dwelling in a ‘HLM’ building

b) The file ' Revenus Disponibles Localisés' (RDL)

= 'Available localized income'.

Complete data source built from the fiscal (tax) files, enhanced with the social security transfers (at the moment they are imputed, unfortunately).

For each region, we calculate 19 "5%-fractiles" from the **individual** distribution of the variable 'equivalised disposable income' :

$$\frac{\textit{total household income}}{\textit{equivalised household size}}$$

Let's consider 2 consecutive 5%-fractiles : X_{REG} is the twentieth of the regional population size (all people together, without any restriction about age)

c) The number of people receiving the « Allocation de Solidarité aux Personnes Agées » (ASPA)

= ‘Solidarity transfer for the elderly’

For each region, we know how many old people - living in a traditionnal household - receive the ASPA.

We can see that, in France, the present SILC weighting system clearly underestimates this subpopulation size.

* * *

Too bad : in France, it is impossible yet to get other data sources about major social transfers able to produce a number of people receiving the corresponding transfers:

- AAH : transfers to compensate for disability;
- RSA : transfers to compensate for low work intensity.

→ we hope to get those (important) figures in the future through the project called ‘FILOSOFI’.

Implementation

Finally, we want a methodology of estimation :

- simple and (rather) easy to explain (avoid the stochastic models);

- which does not necessitate an advanced statistical procedure :

  - neither the calculation of $\hat{B}$ $\rightarrow$ the software is available but one has to launch it whenever Y changes;
  - nor the calculation of the matrix A_X, where $\hat{B} = A_X \cdot Y$ $\rightarrow$ it is 'one shot', but the software is complex.

- which is tractable for any external user who has not in the same time the **national** SILC sample file **and** the auxiliary and complete **regional** informations (Eurostat for instance…).

The asset of this methodology will be its extreme simplicity to be implemented : it must be within the reach of any non-statistician people, wheras the synthetic estimator is well-known for a long time !

The tool = a calibration technique !

- *Quick overview about the calibration technique*

Initial weights d_k without bias :

$$E\left(\sum_{k\in s} d_k \cdot Y_k\right) = \sum_{k\in U} Y_k$$

We want new weights w_k close to the d_k:

$$Min \sum_{k\in s} D(d_k, w_k)$$

$$\text{sc} \quad \sum_{k\in s} w_k \cdot X_k = \sum_{k\in U} X_k$$

We can get w_k, depending on $D(\ )$

$$D(d_k, w_k) = \frac{(w_k - d_k)^2}{d_k}$$: « linear » method

- %calmar (SAS)
- gcalib (SPSS)
- package sampling (R)

…

- *The steps*

1. consider the **national** SILC file s;

2. consider a region and its marginals X_{REG} ;

3. include the total number of <u>households</u> in X_{REG} ;

4. « normalize » the weights : if d_k = weight of the household k in the national SILC file

$$\forall k \in s \, , \quad d_k^* = d_k \cdot \frac{N_{REG}}{N} \; ;$$

5. use the **linear** method to calibrate the **national** file on the **regional** marginals X_{REG};

6. get the calibrated weights $w_k^{calé}$ (which depend on the region), for $k \in s$.

<u>Fundamental property</u>

Whatever the variable of interest Y is (quantitative or categorical) we get the synthetic estimator :

$$\boxed{\sum_{k \in s} w_k^{calé} \cdot Y_k = \hat{B}_{NAT}^t \cdot X_{REG}}$$

It never fails with the linear method.

All people in a given household have the same (calibrated) weight ;
$\Rightarrow$ statistical unit for calibration = household.

So, all the X_k represent the total number of people in the household k verifying such or such property.

The marginals X_{REG} represent a total number of people (verifying such or such property) in the given region.

Two disturbing features - but without any consequence :

1. the poverty parameters θ_λ are made with **categorical** variables Y_k ;

2. one may have a lot of **negative weights** (up to 17% of the weights for France).
   The calibration is just a useful tool, and the final estimation is anyway the good one !

Keep in mind :

- the proportion of negative weights is strongly connected to the size of the difference between the true respective structures X_{REG} and X_{NAT} (after normalization);

- any **non-linear** method of calibration no longer gives the synthetic estimator : it has no real asset, but one can find some theoretical arguments to justify its use.

  As a conclusion, it is better to restrict oneself to the use of the linear method.

The main results

We start with :

- SILC 2009 : about the incomes 2008.
- SILC 2010 : about the incomes 2009.

The equivalised disposable income of households/ individuals in RDL was calculated according to the SILC concept; however, it remains some heterogeneity about the social transfers.

The **regional** at-risk-of-poverty θ_1 is already ‘known’ thanks to the RDL source : so, it is a good benchmark to assess the calibration methodology !

At the national level

The **national** calibration was done again with the new marginals : it does not (almost) destroy the current calibration.

Current calibration : source = labor force survey
New calibration : sources = census + RDL + ASPA

In 2009, the new national calibration produced 239 negative weights ; in 2010, it produced 72 negative weights.

Ratio of weights « new calibration / current calibration »

Quantile	Ratio of weights (for 2010)
Maximum	3.96
99 %	1.86
95 %	1.48
90 %	1.36
75%	1.16
Median	**0.99**
25%	0.85
10 %	0.68
5%	0.52
1%	0.12
Minimum	- 1.66

National estimations of the population sizes, depending on the calibration used

YEAR	Number of households		Number of people	
	Current calibration	New calibration	Current calibration	New calibration
2009	26 997 136	26 866 278	60 665 878	60 997 867
2010	27 293 225	27 106 998	60 997 389	61 298 104

Year 2009

National poverty indicators, in %

Poverty indicator	Current calibration	New calibration
$\hat{\theta}_1$	12,89	12,77
$\hat{\theta}_2$	13,55	13,63
$\hat{\theta}_3$	5,56	5,82
$\hat{\theta}_4$	8,36	8,04
$\hat{\theta}_5$	18,47	18,40
$\hat{\theta}_6$	1,19	1,29

Year 2010

National poverty indicators, in %

Poverty indicator	Current calibration	New calibration
$\hat{\theta}_1$	13,28	13,78
$\hat{\theta}_2$	12,62	12,47
$\hat{\theta}_3$	5,79	5,80
$\hat{\theta}_4$	9,89	9,28
$\hat{\theta}_5$	19,17	19,45
$\hat{\theta}_6$	1,52	1,44

Year 2009

Total number of people in poverty

Consequence of a new calibration method

Poverty according to …	Current calibration	New calibration	Evolution %
θ_1	7 820 418	7 788 426	**- 0,4**
θ_2	8 219 312	8 319 704	**+ 1,2**
θ_3	3 372 182	3 549 520	**+ 5,3**
θ_4	3 873 144	3 816 054	**- 1,5**
θ_5	11 207 632	11 229 969	**+ 0,2**
θ_6	721 499	788 592	**+ 9,3**

Year 2010

Total number of people in poverty

Consequence of a new calibration method

Poverty according to …	Current calibration	New calibration	Evolution %
θ_1	8 098 613	8 448 001	**+ 4,3**
θ_2	7 700 525	7 645 298	**- 0,7**
θ_3	3 529 922	3 554 406	**+ 0,7**
θ_4	4 584 882	4 393 619	**- 4,2**
θ_5	11 692 708	11 920 864	**+2,0**
θ_6	924 246	881 022	**- 4,7**

At the regional level

‘Specific’ regions exist !!!

→ in France, mainly Ile-de-France and Corse
(the following examples concern SILC 2010) :

- in Ile-de-France, the census gives 7 629 farmers, wheras with the national file - using the d_k^* - we get 84 079 farmers : as a consequence, the weights of the farmers in the national SILC file will collapse;

- in Corse, there are 9 934 elderly people in the complete file ASPA, but we have only 927 individuals according to the national SILC file - using the d_k^*. So, the weights of people receiving ASPA in the SILC national file will seriously increase.

And in compensation :

REG	Number of NEGATIVE weights 2009	Number of NEGATIVE weights 2010
11	1632	1849
21	370	331
22	668	621
23	313	136
24	380	363
25	730	755
26	677	722
31	449	288
41	364	227
42	333	200
43	509	454
52	432	282
53	552	544
54	555	571
72	359	263
73	485	398
74	547	510
82	314	135
83	540	541
91	494	435
93	658	603
94	1926	1883

Assessement of the bias due to the model : for each poverty concept, comparaison between the **national** SILC number of people and the **sum** of the **regional** numbers of people.

Year 2009

Relative bias due to the model, in %

Poverty θ_1	Poverty θ_2	Poverty θ_3
- 0.76	**1.50**	**5.68**

Poverty θ_4	Poverty θ_5	Poverty θ_6
- 1.71	**- 0.05**	**9.45**

Year 2010

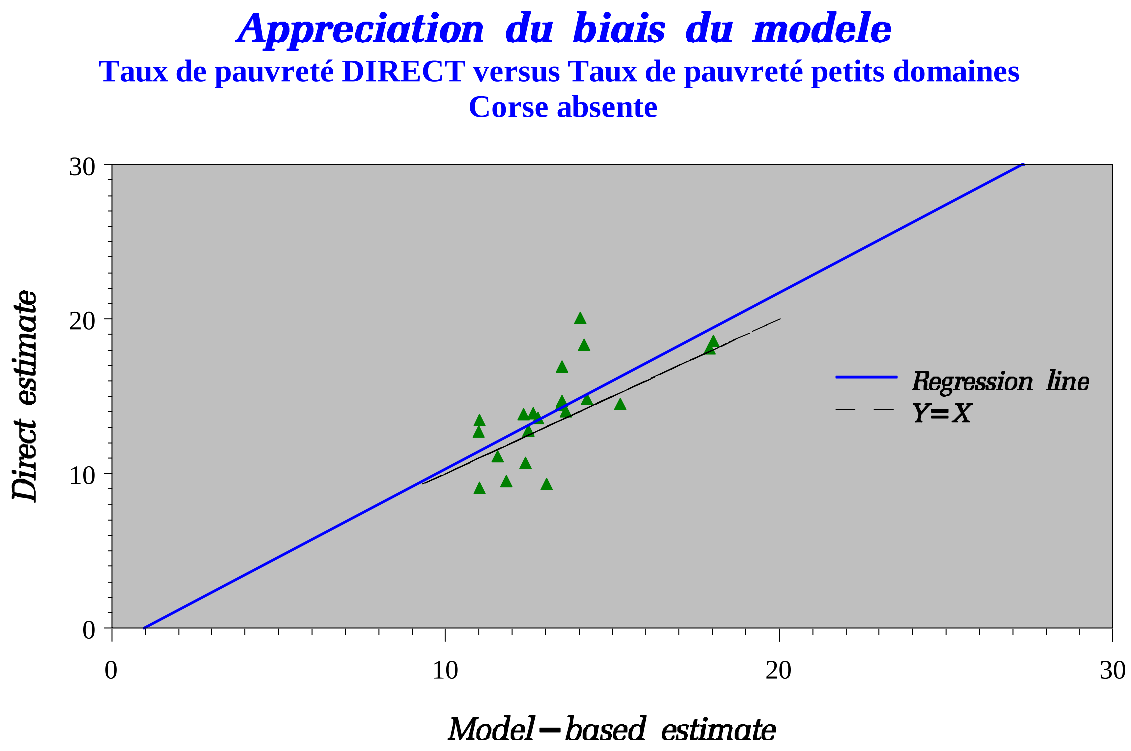
Relative bias due to the model, in %

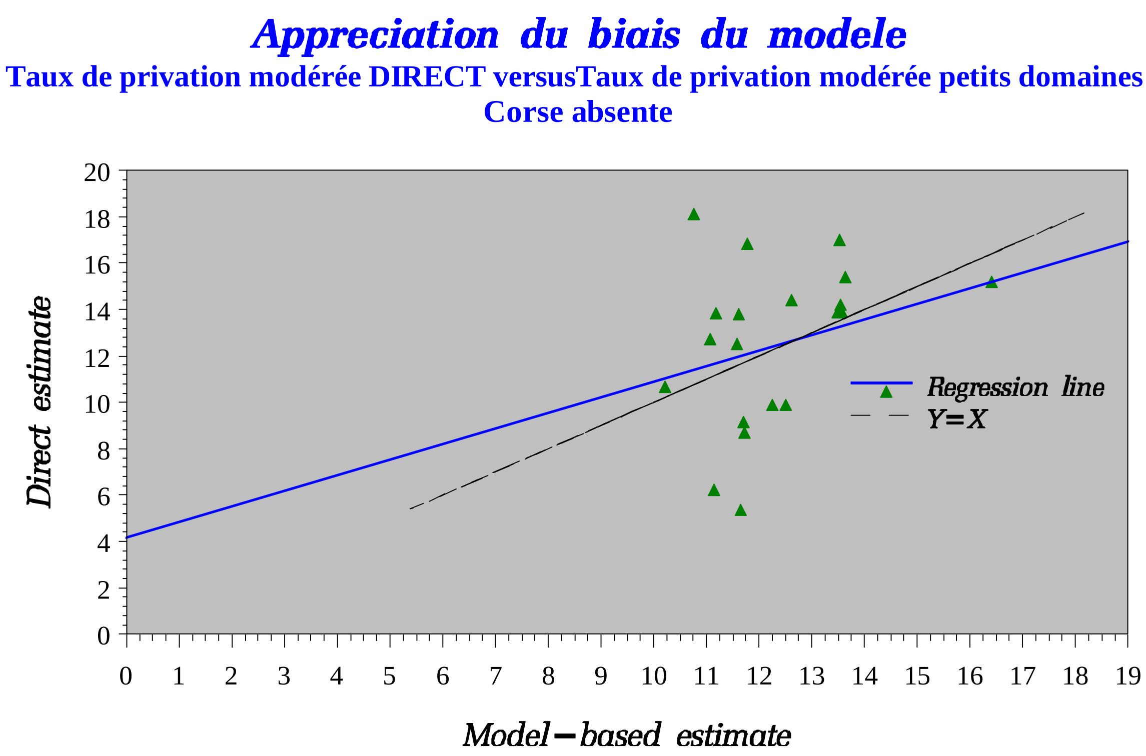
Poverty θ_1	Poverty θ_2	Poverty θ_3
4,91	**- 0,58**	**0,99**

Poverty θ_4	Poverty θ_5	Poverty θ_6
- 4,23	**2,32**	**- 4,16**

Conclusion : generally, **we get satisfactory errors** - mainly concerning the AROPE indicator $\hat{\theta}_5$.

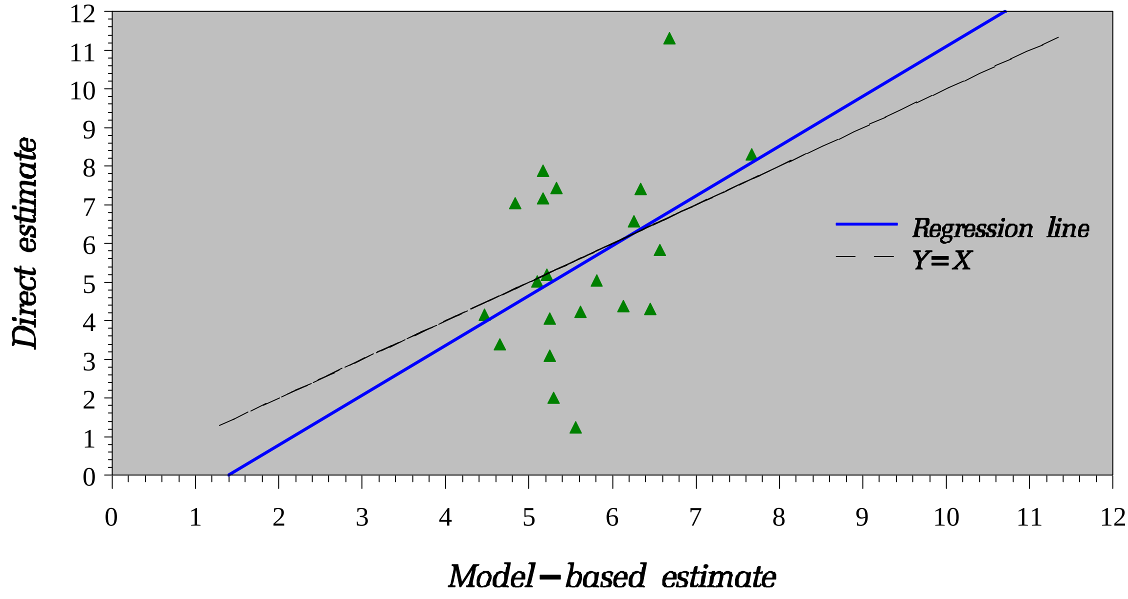
Assessement of the bias due to the model : a graphic approach (respectively for the 6 indicators).





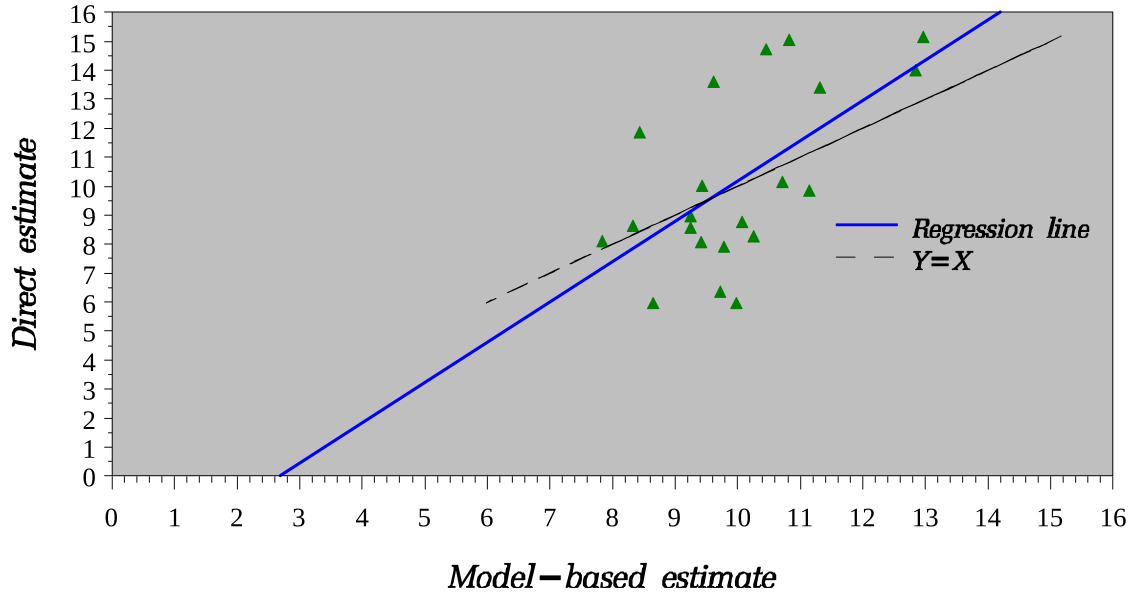
Appreciation du biais du modele

Taux de privation forte DIRECT versusTaux de privation forte petits domaines
Corse absente



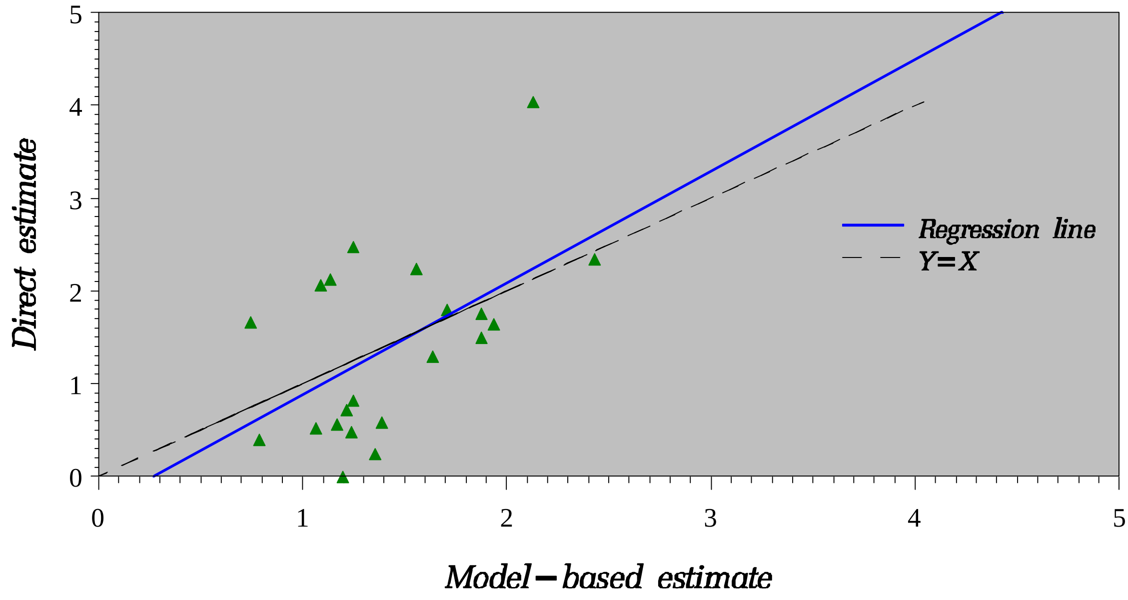
Appreciation du biais du modele

Taux de faible intensité DIRECT versus Taux de faible intensité petits domaines
Corse absente



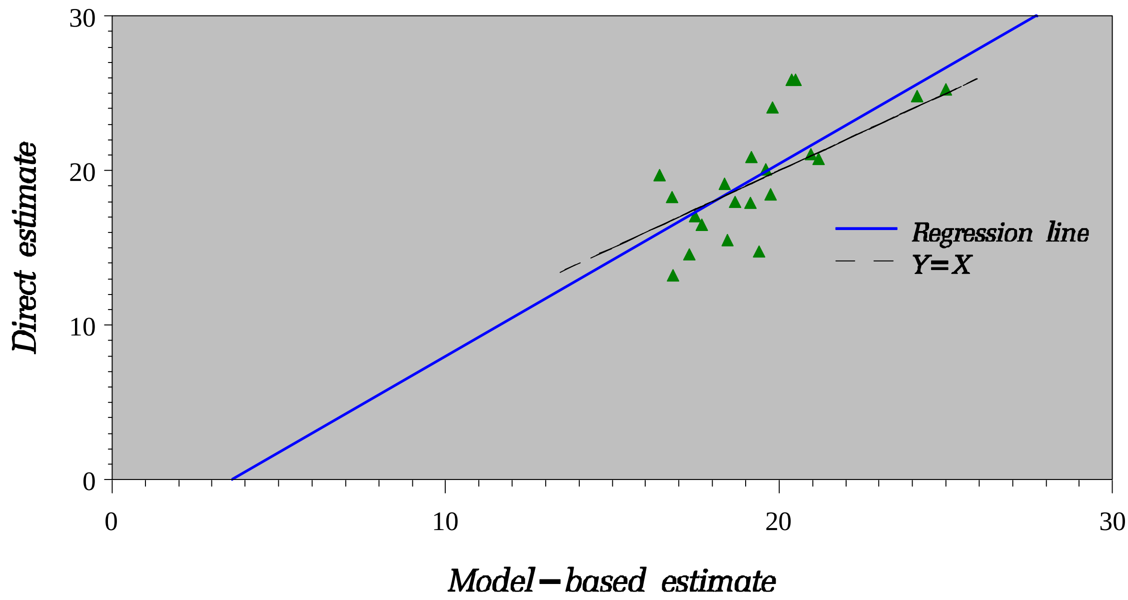
Appreciation du biais du modele

Taux de pauvreté en inter, DIRECT versus Taux de pauvreté en inter, petits domaines
Corse absente



Appreciation du biais du modele

Taux de pauvreté en UNION, DIRECT versus Taux de pauvreté en union, petits domaines
Corse absente



Bias processing :

After the diagnostic, finally, we implement a ratio estimator - called « *benchmarking* » - whose reference is the national calibrated estimation $\hat{\theta}_{NAT}$ **with the new marginals** (versus the current calibration, which would produce the current official poverty figures).

$$\hat{\hat{\theta}}_{REG} = \hat{\theta}_{REG} \cdot \frac{\hat{\theta}_{NAT}}{\sum_{REG} \hat{\theta}_{REG}}$$

$\hat{\theta}_{REG}$: **before** benchmarking

$\hat{\hat{\theta}}_{REG}$: **after** benchmarking

$\Rightarrow$ **So, we get rid of a (large) part of the bias.**

Moreover, we have now a **consistent data dissemination :**

national estimation = sum of the regional estimations

FINAL ESTIMATION

*Estimation of the regional poverty ratios - year **2010***

REG	Source RDL	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$
11	12.50	12.43	13.50	6.26	9.43	17.70	1.88
21	14.61	14.26	14.43	6.58	11.33	20.97	1.88
22	14.43	14.05	13.65	6.46	10.83	20.52	1.94
23	12.99	12.79	13.58	6.14	10.26	19.16	1.64
24	11.78	11.58	11.60	5.22	9.25	17.50	1.22
25	13.31	13.06	12.52	5.62	9.99	19.41	1.36
26	12.52	12.37	11.66	5.30	9.73	18.47	1.25
31	18.55	18.04	16.42	7.68	12.85	25.03	2.43
41	13.88	13.52	12.62	5.82	10.47	19.82	1.56
42	11.29	11.00	10.78	4.84	8.44	16.45	1.14
43	12.85	12.66	12.26	5.56	9.63	18.71	1.39
52	11.17	11.04	11.15	4.66	7.85	16.84	0.75
53	11.17	11.05	10.23	4.47	8.34	16.82	0.79
54	13.83	13.49	11.63	5.18	9.79	19.63	1.09
72	12.93	12.50	11.20	5.10	9.25	18.39	1.07
73	13.97	13.52	11.08	5.18	9.44	19.19	1.25
74	14.66	14.16	11.80	5.34	10.73	20.41	1.20
82	11.82	11.84	11.71	5.25	8.66	17.33	1.24
83	14.00	13.62	11.73	5.26	10.08	19.77	1.17
91	18.63	17.93	13.54	6.69	12.97	24.15	2.13
93	15.75	15.25	13.56	6.34	11.16	21.20	1.71
94	19.32	18.57	14.18	7.77	13.98	24.91	2.91

The official poverty indicators from RDL (cf. website *insee.fr)* are close to the estimated rates $\hat{\theta}_1$: it is a way to validate the regional weights. One can think that the regional weights behave correctly with the other estimated ratios $\hat{\theta}_\lambda$.

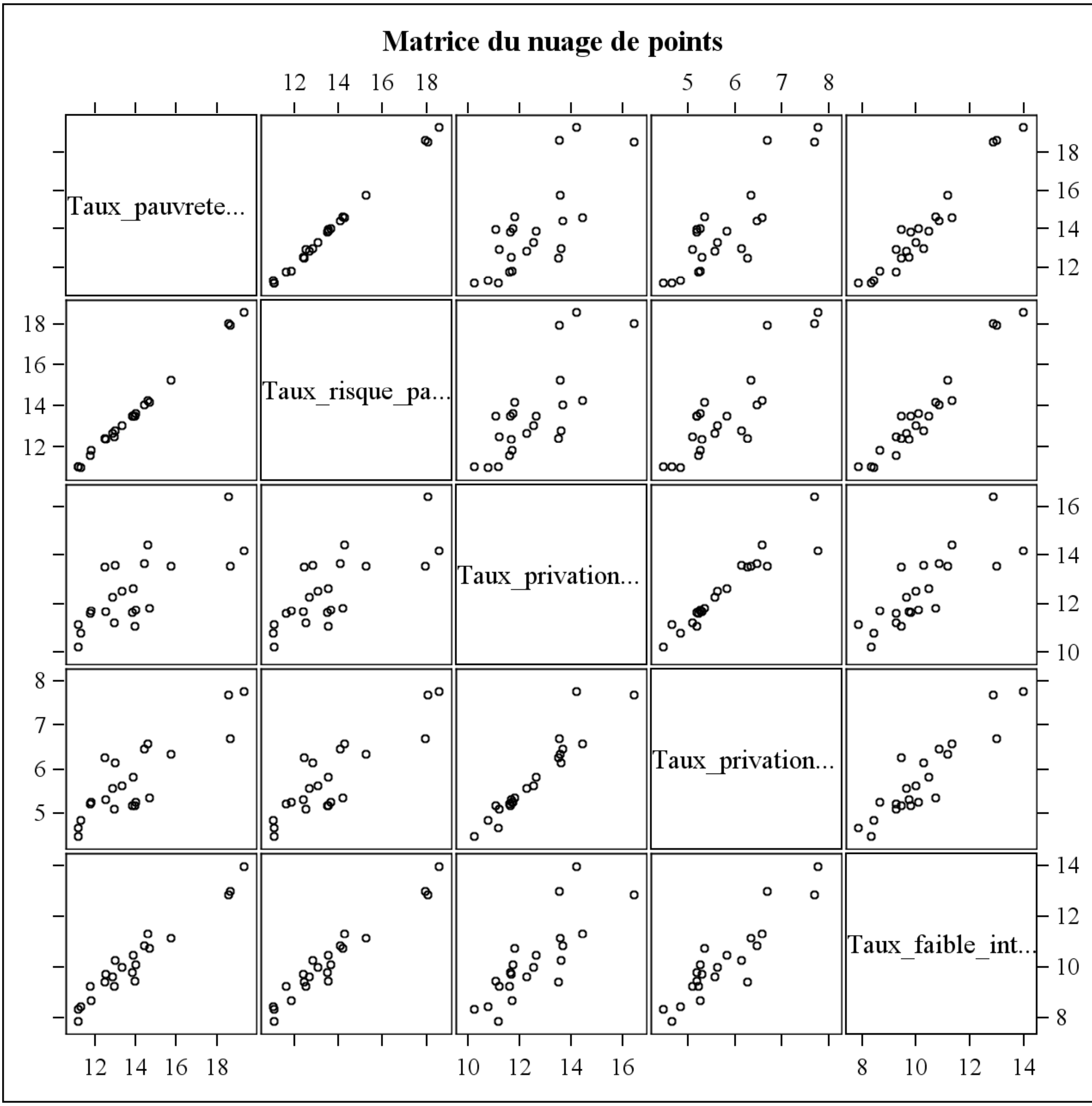
Simple descriptive statistics about
the regional poverty indicators - Year ***2009***

Variable	N	Mean	Standard-error	Min	Max
Poverty indicator from RDL	22	13.41	2.5	10.60	20.00
$\hat{\theta}_1$	22	13.23	2.4	10.55	19.47
$\hat{\theta}_2$	22	13.16	1.7	10.55	17.09
$\hat{\theta}_3$	22	5.44	0.87	4.22	7.76
$\hat{\theta}_4$	22	8.54	1.5	6.42	12.73
$\hat{\theta}_5$	22	18.97	2.7	15.87	26.51
$\hat{\theta}_6$	22	1.09	0.44	0.45	2.05

Simple descriptive statistics about the regional poverty indicators - Year ***2010***

Variable	N	Mean	Standard-error	Min	Max
Poverty indicator from RDL	22	13.91	2.3	11.17	19.32
$\hat{\theta}_1$	22	13.58	2.2	11.00	18.57
$\hat{\theta}_2$	22	12.49	1.5	10.23	16.42
$\hat{\theta}_3$	22	5.76	0.89	4.47	7.77
$\hat{\theta}_4$	22	10.20	1.5	7.85	13.98
$\hat{\theta}_5$	22	19.65	2.5	16.45	25.03
$\hat{\theta}_6$	22	1.50	0.53	0.75	2.91

We cross 5 concepts of poverty indicators : the poverty indicator from RDL and respectively $\hat{\theta}_1$ to $\hat{\theta}_4$.



Correlations between the different poverty indicators
Year 2010

	Indicator RDL	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$
Indicator RDL	1	0.999	0.742	0.857	0.969	0.990	0.843
$\hat{\theta}_1$	0.999	1	0.761	0.870	0.970	0.991	0.854
$\hat{\theta}_2$	0.742	0.761	1	0.948	0.803	0.788	0.881
$\hat{\theta}_3$	0.857	0.870	0.948	1	0.911	0.874	0.982
$\hat{\theta}_4$	0.969	0.970	0.803	0.911	1	0.977	0.900
$\hat{\theta}_5$	0.990	0.991	0.788	0.874	0.977	1	0.844
$\hat{\theta}_6$	0.843	0.854	0.881	0.982	0.900	0.844	1

We can also compare the regional situations according to the different poverty criteria : the richest region is ranked 1 and the poorest one is ranked 22 :

$\Downarrow$

Ranks of the regions according to the poverty concept
Year ***2010***

REG	Rank RDL	Rank $\hat{\theta}_1$	Rank $\hat{\theta}_2$	Rank $\hat{\theta}_3$	Rank $\hat{\theta}_4$	Rank $\hat{\theta}_5$	Rank $\hat{\theta}_6$
11	6	7	15	16	7	6	17
21	17	18	21	19	19	18	18
22	16	16	19	18	17	17	19
23	10	10	18	15	14	10	15
24	4	4	6	7	6	5	8
25	11	11	13	13	12	12	12
26	7	6	8	10	10	8	10
31	20	21	22	21	20	22	21
41	13	13	14	14	15	15	14
42	3	1	2	3	3	1	5
43	8	9	12	12	9	9	13
52	1	2	4	2	1	3	1
53	2	3	1	1	2	2	2
54	12	12	7	6	11	13	4
72	9	8	5	4	5	7	3
73	14	14	3	5	8	11	11
74	18	17	11	11	16	16	7
82	5	5	9	8	4	4	9
83	15	15	10	9	13	14	6
91	21	20	16	20	21	20	20
93	19	19	17	17	18	19	16
94	22	22	20	22	22	21	22

Direct estimations (out of curiosity…)
Year 2010

REG	θ_1		θ_5	
	Direct methodology	**"Small area" methodology**	**Direct methodology**	**"Small area" methodology**
11	10.73	12.43	16.51	17.70
21	14.88	14.26	21.13	20.97
22	20.09	14.05	25.95	20.52
23	13.64	12.79	17.99	19.16
24	11.15	11.58	17.12	17.50
25	9.40	13.06	14.82	19.41
26	13.86	12.37	15.55	18.47
31	18.65	18.04	25.31	25.03
41	16.99	13.52	24.12	19.82
42	12.80	11.00	19.72	16.45
43	13.94	12.66	18.04	18.71
52	9.12	11.04	13.26	16.84
53	13.53	11.05	18.31	16.82
54	14.53	13.49	20.12	19.63
72	12.86	12.50	19.17	18.39
73	14.78	13.52	20.90	19.19
74	18.40	14.16	25.93	20.41
82	9.54	11.84	14.62	17.33
83	14.05	13.62	18.50	19.77
91	18.15	17.93	24.86	24.15
93	14.56	15.25	20.82	21.20
94	28.35	18.57	41.95	24.91

$\Rightarrow$ Small area method = **SHRANKING phenomenon.**

As a conclusion, remember :

- A model approach - but there is no way to escape !
- A classical synthetic estimator;
- A simple to implement methodology;
- A very easy way to get any regional estimator, for any external user;
- An estimator with a (very) small variance;
- Some bias of course, but it remains satisfactory if one gets explanatory enough variables of the poverty through complete data sources (and we know some easy ways to assess - and reduce - the bias).

Every NSI should be able to use "at least" this methodology - which of course can be afterwards improved by more sophisticated models.