ISSN 2315-0807

eurostat
Statistical working papers

Regional estimates of poverty indicators based on a calibration technique

2015 edition

eurostat

eurostat
Statistical working papers

Regional estimates of poverty indicators based on a calibration technique

2015 edition

eurostat

Europe Direct is a service to help you find answers to your questions about the European Union.

Freephone number (*)**:**

00 800 6 7 8 9 10 11

(*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

More information on the European Union is available on the Internet (http://europa.eu).

Luxembourg: Publications Office of the European Union, 2015

ISBN 978-92-79-47461-3
ISSN 2315-0807
doi: 10.2785/879307
Cat. No: KS-TC-15-001-EN-N

Author

Pascal Ardilly

Insee, Département des méthodes statistiques ([1])

Summary

To answer to DG Regio's request, Eurostat wishes to carry out regional estimations of the AROPE indicator (At-Risk-Of-Poverty-or-Exclusion) and its components, derived from the EU-SILC survey. For that purpose, we suggest to use estimators which are named ‘synthetic estimators’ by the small area estimation theory. Their calculation would be made with a calibration method on regional margins supplying 22 sets of weights implying all the households of the national SILC sample (one set for each region). This method is adapted to the production of regional indicators built from variables of interest which appear well correlated with the calibration variables. We thus use external sources in order to produce regional margins from variables well correlated with the variables making up the AROPE indicator, in particular at the moment the source ‘Revenus disponibles localisés’.

The method is based essentially on: 1) the existence of ‘enough explanatory’ variables of the phenomena of interest and for which we can have regional margins; 2) the hypothesis that conditionally in these explanatory variables, the geography does not have impact anymore on the phenomenon of interest. Subject to these points, the production of the regional indicators using the SILC source seems satisfactory.

([1]) Institut national de la statistique et des études économiques, France.
e-mail : pascal.ardilly@insee.fr

Content

1. The context

In France, the EU-SILC survey is a yearly survey with a rotational sample stemming from a complex design, drawn in a master-sample and which produces, among others statistics, some indicators of poverty at the national level. From SILC 2010, sub-samples were drawn in the census database. The national weights, taken as they are and applied for any regional estimation, would certainly allow to obtain unbiased estimates of the regional indicators of poverty, but at the price of a huge sampling variance, which goes clearly beyond what we can reasonably accept. Indeed, if the national survey uses a sample size from 10 500 to 11 000 responding households every year, the 'Ile-de-France' region is the only region to obtain more than 1000 responding households: in an 'average' region, only some hundreds of households are successfully interviewed. That is why, to satisfy the Eurostat request to calculate some indicators at the regional level, it was decided to apply — to the annual cross sectional sample only — a method relying on a model of behaviour (the well know 'small area estimation' techniques).

This document specifies the chosen method in France and gives the essential results. It takes place in an exploratory context and proposes a processing which can evolve in time according to the possible availability of new datasets. All the following outcomes apply to the metropolitan France only and concern the only population living in an ordinary household. The following table gives, for every region (22 metropolitan regions — see appendix 1), the total number of responding households to the SILC survey.

Table 1: Total number of responding households to the SILC survey

REG	Sample of responding households 2009	Sample of responding households 2010
11	1 682	1 729
21	271	288
22	398	409
23	259	286
24	426	416
25	286	271
26	315	321
31	723	788
41	467	483
42	278	297
43	265	261
52	725	775
53	583	628
54	352	361
72	637	679
73	480	512
74	157	166
82	850	907
83	239	249
91	420	444
93	756	739
94	33	35
Total	10 602	11 044

2. The parameters of interest

The estimation process concerns six parameters in all.

i) The first indicator θ_1 gives the proportion of people whose equivalised disposal income (defined as the disposable income of the household to which the person belongs divided by the equivalised household size) is lower than 60 % of the standard of living median (at-risk-of-poverty rate). There is, in the SILC file, a dummy variable (code HX080) which spots the households / people concerned by this state of poverty. This variable is calculated at the household level, but we just have to process the variable of household's size to go to the estimation of a number of individuals.

Let's note w_i the weight of the household i, N_i its size (without any filter — all the individuals in the household are concerned, whatever their ages are), and 1_{i_pauvre} the dummy variable indicating the poor households / individuals according to θ_1, we use the estimate

$$\hat{\theta}_1 = \frac{\sum_{i \in s} w_i \cdot N_i \cdot 1_{i\_pauvre}}{\sum_{i \in s} w_i \cdot N_i}$$

where s is the responding household sample for the year. The weight w_i includes a non-response treatment and a national calibration.

ii) The second indicator θ_2 concerns the physical persons subjected to a 'moderate' material deprivation, that is not being able to afford at least 3 items among a list of 9 items proposed by the survey. A dummy variable (code DEPRIVED), in the SILC file at the individual level, identifies the concerned people. If we note w_k the weight of the physical individual k (which integrates the correction of non-response and the national calibration) and Y_k the value of the dummy variable DEPRIVED, we estimate

$$\hat{\theta}_2 = \frac{\sum_{k \in s} w_k \cdot Y_k}{\sum_{k \in s} w_k}$$

Where s is the sample of responding people of the year.

iii) The third indicator θ_3 concerns the physical persons subjected to a 'severe' material deprivation, that is not being able to afford at least 4 items among the list of 9 items evoked in ii). A dummy variable (code SEV_DEP) identifies the concerned people. We obtain the estimator $\hat{\theta}_3$ exactly on the model of $\hat{\theta}_2$.

iv) The fourth indicator θ_4 gives the proportion of the individuals living in households with a very low

level of work intensity. The variable LWI allows to identify the concerned people, for whom it takes the value 1. This indicator θ_4 is different from the preceding ones because it is defined on a specific scope of the population: indeed, it concerns the individuals living in households which are not constituted exclusively by people older than 60 years old, or only by students from 18 to 24 years old, or only by people under 18 years old. Numerically, on 61 millions of individuals covered by SILC, approximately 42 millions are in the scope of this indicator — which represents so 70 % of the individuals.

We note D the domain of the individuals concerned by the indicator, who are indicated by LWI equal to 0 or to 1 — the value 2 being reserved for the individuals out-of-scope. The dummy variable associated to the domain is $1_{k\in D}$. We also define the variable Z_k equal to 1 if the individual k is in the domain D and, at the same time, lives in a household with a very low work intensity. Finally

$$\hat{\theta}_4 = \frac{\sum_{k\in s} w_k \cdot Z_k}{\sum_{k\in s} w_k \cdot 1_{k\in D}}$$

Note that we find some children with a modality LWI equal to 2, thus out-of-scope: indeed, they live in households in which all the adults are 60 years old or more.

v) The fifth indicator θ_5 quantifies the proportion of the individuals who are affected by one at least of the following three states of poverty or social exclusion: (monetary) poverty according to θ_1, severe material deprivation (according to θ_3), low work intensity (according to θ_4). This indicator is considered as the 'key' indicator of poverty by Eurostat as it is part of the strategy EU2020. In particular, allocations of European funds at the regional level will be defined by taking into account this indicator, what gives a quite particular importance for its estimation. The useful individual variable to define this indicator is the maximum of the three concerned components of poverty: this definition is natural because every component is a dummy variable. So we define

$$Y_k = MAX(HX080_k, SEV\_DEP_k, Z_k)$$

The estimator becomes

$$\hat{\theta}_5 = \frac{\sum_{k\in s} w_k \cdot Y_k}{\sum_{k\in s} w_k}$$

vi) The sixth and last indicator θ_6 quantifies the proportion of the individuals who are affected by three states of poverty simultaneously. The involved variable is

$$Y_k = MIN(HX080_k, SEV\_DEP_k, Z_k)$$

and the estimator $\hat{\theta}_6$ has the same functional shape as $\hat{\theta}_5$.

3. The proposed methodology for a regional estimation

A quick overview of the calibration method

We start with a given sample s and the associated unbiased weights d_k. Let's suppose that we get an auxiliary information $X_k \in R^p$, known for every unit in the frame U, so that $\sum_{k \in U} X_k$ is a known margin. We seek new weights w_k, as close as possible to the d_k, so that the following calibration equation is assured:

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

For that, we use a distance function $\sum_{k \in s} D(w_k, d_k)$ and minimize it under the above constraint. The solution is

$$w_k = d_k \cdot F\left(X_k^t \cdot \lambda\right)$$

with F a (known) function connected to the distance D and λ an unknown vector at this stage. To get it, we solve the constraint equation as a function of λ (it is a complex system of p equations with p unknown values).

The calibration stage as two fundamental properties:

- the resulting calibrated estimator $\sum_{k \in s} w_k \cdot Y_k$ has no significative bias if the sample size is large;
- the sampling variance of the calibrated estimator closely depends on the linear correlation between X_k and Y_k: if Y_k is close to any linear combinaison of the X_k components, we can expect a large decrease in the variance compared to the variance of the initial unbiased estimator $\sum_{k \in s} d_k \cdot Y_k$.

There is nothing special to do when we manage a qualitative variable Y_k : in this situation, the residuals ε_k can be defined exactly in the same way as for a quantitative variable. If an auxiliary component of X_k is a qualitative variable, then we turn it into an operational variable through a dummy variable. For instance if X_k is the variable 'sex', we define a dummy variable equal to 1 if k is a woman and 0 if k is a man. The total margin is then equal to the total number of women in the whole population U. The

use of a set of qualitative auxiliary variables X_k means that Y_k is essentially explained by those factors through an additive model. If we do not include margins which count the number of people verifying simultaneously the modalities of different factors, it means that we trust an additive model without cross-effects. If we want to introduce cross-effects, then the margins must be defined in a similar way, at the same level of information. Different softwares exist to implement the calibration technique (Macro %Calmar in SAS, g-calib in SPSS, Sampling package in R, …).

The proposed estimator

The regional estimation of the six previous indicators can be envisaged in several ways — because the package of 'small area estimation' techniques is very vast — but in every case it is necessary to use a model, that is a hypothesis which connects the regional behaviour with a supra-regional behaviour (considered as the reference behaviour). To facilitate the task, the idea currently exploited takes the national behaviour as a reference (it will be possible, as a study and later, to build groups of regions which can constitute a possibly more relevant reference). The basic postulate is the following one: the relationship between a given variable of poverty and a set of explanatory variables does not depend on the region. Technically, it is translated in the following way:

Y_k is the value of the variable of interest relative to the unit k (in our context, it will be a household). This value is collected by the national EU-SILC survey (we suppose without measurement error, what is certainly excessive).

X_k is the value of the explanatory auxiliary variable relative to the unit k, the true regional (and thus national) total of which we know thanks to (pseudo) complete external data sources. This variable is vectorial, with a dimension p.

We re-write the value Y_k according to a linear combination of the components of X_k. According to the theory of the multivariate linear regression, considering that the constant variable is a part of the vector (what it is possible to assure as soon as we know the size of the population), there is always a unique vectorial coefficient B verifying

$$Y_k = B^t.X_k + \varepsilon_k$$

with $\sum_{k \in U} \varepsilon_k = 0$ and $\sum_{k \in U} \varepsilon_k^2$ minimum, where U is the national population of households.

This relationship applies to the national level, thus in the same way at the regional level, for every region. It means that the coefficient B has not to be defined specifically at the regional level. We shall notice that this mathematical equation is absolutely not a model (there is no hypothesis behind it !) and adapts

very well with variables Y_k equal to 0 or to 1, that is with qualitative variables of interest (caution, we are not facing a context where Y_k are random variables — as in econometrics — context which actually would forbid such a writing when Y_k is qualitative and would require to be rather interested in the probability than Y_k equals one).

We fall over very naturally towards the synthetic estimators by considering that we are facing a situation in which the properties $\sum_{k\in U}\varepsilon_k = 0$ and $\sum_{k\in U}\varepsilon_k^2$ minimum lead to build at an individual level the residuals ε_k which counterbalance more or less when we consider infra-national domains: considering a domain $D \subset U$, we would have then in reality $\sum_{k\in D}\varepsilon_k$ close to zero, what would urge to write $\sum_{k\in D}\varepsilon_k = 0$ and it is this equation which precisely constitutes our model. As another way of presenting this model, we can say that the connection between Y and X does not depend on the geography: it means that the national coefficient B is the same that the regional coefficient B.

Yet, set as hypothesis (model) $\sum_{k\in D}\varepsilon_k = 0$ is to write $\sum_{k\in D}Y_k = B^t \cdot \sum_{k\in D}X_k$, what allows to estimate $\sum_{k\in D}Y_k$ by using the member of the right side of the equation because the X_k are auxiliary variables. In this particular case, D is a region and the poverty indicators are ratios of total numbers. It is then a question of estimating in a first step some regional totals like

$$Y_{REG} = \sum_{k\in REG} Y_k$$

where REG indicates the entire regional population (living in an ordinary household). An estimator of this total, in the spirit of the kept model, is

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

The vector B has a dimension p and its estimator $\hat{B}$ uses the whole set of national data X_k and Y_k for k in the sample s: it is the main asset of the model since the national sample has a large size, and as a consequence the estimator $\hat{B}$ has a (very) low sampling variance.

The estimator $\hat{Y}_{REG}$ is known as a 'synthetic' estimator. As it is based on a model of behavior, it is naturally biased, but in return its variance remains very modest. The appreciation of the bias is obviously delicate, because we are never aware of the 'true value', but we have a graphic and visual tool of validation (see appendix 5) and a simple technique consisting in comparing the sum of the regional estimations with the direct national estimation coming from the national SILC sample. An argument of common sense is added, because if the explanatory variables are enough diversified and correlated with

the state of poverty, we can suppose that the specific role of the geography is not significant any more. It is rather obvious for the at-risk-of-poverty θ_1 since the equivalised disposal income is an explanatory variable (but it is less clear for the other aspects of poverty). It is thus necessary to make the exercise consisting in considering the whole set of the explanatory variables given in part 4, in imagining two individuals living, the first one in a region A and the second one in a region B, and who would take exactly the same values for each of the considered auxiliary variables X_k : can we reasonably think that the 'poverty' of the first one will be different of the 'poverty' of the other one ? The residual risk at this level is constituted by a possible missing in the model of a major explanatory variable of poverty, the correlation of which with all the other explanatory variables is not very strong ([2]) and for which the structure would besides differ significantly from a region to the other one ([3]). In our particular case, the auxiliary variables retained (see part 4) are numerous and seem potentially well correlated with the situation of poverty. Nevertheless, it is not difficult to imagine explanatory factors not taken into account (at least in this study) but nevertheless influential, as for example the prices of the real-estate market, or the prices of the goods and services subjected to a local effect.

Finally, the production of regional estimations concerning two consecutive years (incomes 2008 and 2009) allows to spot possible incoherences which could question the model. It is clear that the methodology exposed above is not unique and that there are more sophisticated methods of estimation. Nevertheless, on one hand complication is not synonymic of gain of efficiency, on the other hand it seemed careful to use an easy to understand and to implement technique, which is not dependent on a particular expertise on the very technical topic of small area estimation. It is the reason why the synthetic estimator was proposed as a basis of the French methodology of estimation. This argument about simplicity was taken to the extreme at the level of the implementation. Indeed, the basic and most natural approach consists in processing explicitly the vector of coefficients $\hat{B}$. On the purely technical plan, it is not really difficult to do with a software like SAS for instance, but there are then two practical difficulties to overcome. On one hand it is necessary to do the calculation for every variable of interest Y (in fact, the list can grow if new needs are expressed !), on the other hand and especially, the users of the SILC database will not have the opportunity to find by themselves, in a fast way and without any risk of error, the regional estimations of the poverty indicators disseminated by the national statistical institute (those users will have to launch again the entire procedure of estimation by fitting their own regression, in particular to get beforehand by themselves the true totals of the auxiliary variables, the margins, what is de facto impossible). That is why we conceived a method of calculation of the synthetic indicators which circumvents these serious practical difficulties.

([2]) This possible 'hidden and forgotten' variable has to bring its own part of explanation - if it is a linear combination of the other auxiliary variables, then it is without any impact.

([3]) Otherwise, the associated component of $B^t.X_{REG}$ will be a constant, and the regions will not be differentiated.

It is possible to verify the following result, which is absolutely essential to justify our approach: if we consider the national SILC file and if we make a calibration of this whole file on the regional margins formed by the true regional totals X_{REG} (considered successively, region by region) by using the method called 'linear', we produce a set of weights $w_k^{calé}$ (appropriate to the considered region) which allows to find immediately the synthetic estimate.

In other words**, whatever is the variable of interest** Y (whether it is quantitative or qualitative):

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

The calibrated weight $w_k^{calé}$, as a new variable of the micro data file, can be used with any variable of interest Y — but obviously its statistical relevance depends intrinsically on the correlation between Y and the vector X .

Before the calibration, it is necessary to modify the weights of the national file with the aim of returning any estimation to an order of magnitude comparable to the regional margins. We thus proceed to an operation of 'normalization' in the following way: if d_k is the unbiased weight included in the national file, then

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

where N_{REG} is the size of the regional population of households and N the size of the national population of households.

The application of this method gave satisfactory results from the numeric point of view (see part 5). It has nevertheless two unpleasant drawbacks — but manifestly without any consequence — which it is advisable to underline here. Firstly, the variable Y which is involved in the definition of the parameters θ is either a variable of counting or a dichotomous variable, which does not seem very adapted to the use of a linear relation between Y and some regressors X — relation which applies more naturally to continuous variables. Now, it turns out that the estimation works well without any restriction to a continuous variable. In fact, the ability of the model to predict Y is the essential in this issue, and not the interpretation of the coefficients $\hat{B}$, which have by themselves no interest ([4]). We shall also note that in any sample survey, we do not hesitate to make a calibration in order to improve the estimates of proportions, what seems so heretical because proportions are never means of qualitative variables and

([4]) In a very nearby context, we justify the calibration of the proportion estimates (qualitative variables of interest) by a linear relation — which only pretends to 'assist' the estimation.

because the linear relationship (linear correlation) between these qualitative variables and the calibration variables is the main justification of calibration. Secondly, the method produces negative weights, sometimes in large number (up to 17 % of the weights, approximately). In a situation of classic estimation, it is unacceptable. But in this particular case, beyond the unpleasant situation due to the existence of those weights, one has to consider that it is only a practical tool to calculate a synthetic estimator and circumvent the problem of the explicit calculation of $\hat{B}$: what is important is only to obtain an estimation at the very end which is numerically the right one !

The appendix 2 develops the theory which justifies the use of the calibration. In this particular case, the SAS macro '%Calmar' was used. This appendix shows that:

- the use of the **linear method** works without any problem (option M=1 of %Calmar).
- it is necessary to make beforehand an operation of normalization of the weights to make them compatible with the order of magnitude of the regional totals (see appendix 4).
- it is necessary to include the total size of the population of the statistical units concerned in the margins used for the calibration. In our case, the population concerned is the population of households. It means we have to include the constant variable in the list of regressors (so we get an intercept in the model). The rest of the auxiliary variables is totally free.
- the use of a **non linear** method of calibration no longer produces the synthetic estimator and that when the variable of interest is qualitative, the theoretical justification of the calibration with such a class of methods is more difficult to assess. These considerations do not undermine at all the alternative approaches to the linear method and do not prevent from having the intuition that the calibration by a non linear method gives correct outcomes to estimate the numbers of people we are interested in. Maybe we have *in fine* statistical properties as satisfactory as with the linear approach. But since the linear approach satisfies our expectations by allowing to find a well-known 'small area' estimator, and since we manage an operation of regional estimation with important consequences, it is preferable to reduce the risk and limit to the only use of a linear re weighting method — the interest of the other methods recovering rather from the scientific curiosity.

Note that the proportion of negative weights reflects the amount of the difference between the regional structure and the national structure from the point of view of the calibration variables. So, considering a given region, if the structure X_{REG} (obtained by exploitation of the external sources) is close to the national structure X_{NAT} obtained by a direct exploitation of the national sample SILC, there will be only relatively few negative weights.

4. The constitution of the regional margins

The mobilized auxiliary information is supposed to explain 'in best' the poverty. It results from 3 data sources: the general census of the population, the file 'Revenus Disponibles Localisés' (RDL: it means 'Localized available incomes') and local numbers of beneficiaries of the 'Allocation de Solidarité aux Personnes Agées' (ASPA: it means 'Solidarity transfer for the elderly'). These sources were used for the occasion, because they offer an information *a priori* well correlated with the situation of poverty.

The current operations of calibration of the national EU-SILC survey take benefit of a part of this information, but with a considerable difference: the used source is the Labour force survey. Yet, this source allows actually the production of national margins, but not the production of regional margins. The calibration on regional margins — thus the production of local indicators of poverty — requires a source 'entitled' to produce regional estimations, and in sociodemographic field, at the moment only the census offers this possibility. It means that in prospect of a regular annual production of regional poverty indicators, it will be necessary for France to accept the deadlines constraints of the census, it means for a production of data concerning year **n**, to wait approximately till the middle of the year **n+3**. An alternative can be to build regional margins by piling all the households surveys of a given year — including the Labour force survey and *a priori* without calibration after the phase of non-response treatment (if the sum of the sample sizes is considered as enough !). The auxiliary informations partially listed in part 4 concern sociodemographic concepts which seem relatively simple and we can hope that there is not too much heterogeneousness between the surveys, but nevertheless this operation was never tried and it remains very audacious.

The census

The selected variables are the following ones:

- sex
- age (6 modalities)
- diploma (4 modalities)
- nationality (5 modalities)
- social category : CS (11 modalities)

– live or not in a ZUS (= 'sensitive' urban area ([5]))

– urban unit category (3 modalities)

– household type (5 modalities)

– rent / or not the dwelling in a HLM ([6]) building.

The appendix 3 gives the modalities of those variables.

The data source 'Revenus disponibles Localisés' (RDL)

It is about a complete and annually constituted file, resuming information coming from the tax files and adding to it the amounts of social security transfers. At the moment, those transfers are imputed. This file allowed to produce, for every region and whole France, 5%-fractiles of equivalised disposal income (quantiles from 5% to 5%). The equivalised disposal income is a variable in euro at the household level, which is transferred in the identical on every individual in the household, and which is defined as the ratio of the total household income divided by the equivalised household size.

For every region, we thus get 19 values of 5%-fractiles. Every fractile is defined with regard to the distribution of the equivalised disposal income in the population of individuals (and not in the population of households). If we consider two successive 5%-fractiles, by definition the margin X_{REG} is equal to the twentieth of the regional number of physical individuals (all the ages being together). The calibrated weights are such that, considering the national file SILC, and having previously spotted the individuals whose equivalised disposal income lies between these two fractiles, the sum of the weights of these individuals is equal to the regional margin X_{REG} (the weight of an individual is equal to the weight of its household because the calibration is made at the household level and the variables of interest are 'household' variables allocated uniformly to every individual in the household).

The number of beneficiaries of the 'Allocation de Solidarité aux Personnes Agées' (ASPA)

It was possible to obtain, for every region, the number of beneficiaries of the ASPA living in a common household in metropolitan France. This number, at the national level, is equal to 485 000 persons in 2009 and 489 000 persons in 2010. The current weights of SILC underestimate very strongly this numbers: at the national level, we estimate at 220 000 the number of beneficiaries 2009 and at 297 000 the number

([5]) At the moment, there are about 750 ZUS through the French territory. Those areas are characterized by a significant proportion of deprived people concerning their social and living conditions.

([6]) HLM: dwelling with a rather low rent, reserved for people who have an income below a given threshold.

2010. It is not really surprising if we consider the number of responding people perceiving the transfer ASPA in the national French SILC sample: 69 persons in 2009 and 103 persons in 2010, too small numbers which show that there is a recurring problem of measurement error.

It was very attractive to try to introduce in addition into the margins the regional numbers of beneficiaries of two famous social transfers in France, clearly very well correlated with the poverty: the 'Allocation pour Adulte Handicapé' (AAH — 'Transfer for disable adult') and especially the 'Revenu de Solidarité Active' (RSA — a transfer for people with a small income). Unfortunately, it has not been possible to get those data till now. In particular, it seems *a priori* difficult to separate the transfers paid to persons living in a common household and the transfers paid to persons living in a community. Furthermore, the administration which manages those data publishes merely regional data relative to beneficiaries on December 31st, what is different from the SILC statistics which count the beneficiaries 'during the year', whatever is the concerned period. Naturally, the situation can evolve and if in the future new margins are available, it will always be possible to add them. However, the future output of the Insee project named 'Filosofi' should supply such margins.

We can notice that some information used to calculate the margins of calibration are conceived at the household level, others at the individual level. The calibration concerned the household unit k, what means that the individual variables were systematically transformed into 'household variables' (by a simple sum of the individual values in the household). Finally, all the variables X_k represent numbers of individuals in the household k, verifying such or such modality. The margins thus represent a total number of individuals at the regional level.

5. The outcomes

The results were obtained for two consecutive years: the year SILC 2009, concerning income 2008, and year SILC 2010 concerning income 2009. At the beginning of processing, it was decided to keep some households / individuals who have a negative disposable income (variable HY020). In a very surprising way, to keep or to remove these few units has a small numerical impact on the indicator θ_1 (0,08 points of percentage in 2010 — what can make all the same change the first decimal !). To keep these individuals has nothing in itself of suspect or questionable (we can find actually people bearing strong taxes one given year, but with little income) but it is necessary to know that RDL eliminates the households in question. In any rigor, there is thus a field gap because the margin RDL constituted by the fractiles of income excludes these households while processings made from the file SILC take them into account. We shall consider that this methodological anomaly has modest enough numerical consequences to be accepted.

At the national level

The margins used for the calibration (see part 4) were naturally produced for every region, but also at the national level. It was thus possible to calibrate the national sample on these new margins. We can notice that the latter include — apart from subtleties due to the clustering of the population in modalities — the margins used, in a standard way, for the current calibration of the national survey SILC ([7]), what makes that we do not destroy the calibration made initially. The interesting estimations concern the numbers of poor people according to the various notions of poverty defined in part 2, then obviously the corresponding indicators. The estimated indicator $\hat{\theta}_1$ constitutes a particular case because it can be compared with the indicator of same 'nature' produced by RDL and presented on the official website of Insee. The other regional indicators cannot be estimated — and thus nor validated — by any external source, and it is precisely what justifies that this operation of regional estimation uses the survey SILC ([8]).

The disposable income used to constitute the individual values of the equivalised disposal income, associated with the margins X_{REG}, comes from RDL (see part 4). Concerning SILC, the equivalised disposal income were previously transformed in accordance to the 'RDL concept' by the following operations: tax-exempt incomes was deleted by the income SILC (it includes some allowances and pensions, grants, tax-exempt family transfers, local social benefits, increases of retirement pension for having brought up 3 children or more) and the ISF ([9]) (special tax on high patrimony). The income of the

([7]) At the household level: urban unit stratum, type of household + age, social category and diploma of the reference person ; at the individual level : sex and age.

([8]) If the request of Eurostat had concerned only the at-risk-of-poverty indicator at the regional level, the RDL data source would have been enough for answering the request.

([9]) Impôt de Solidarité sur la Fortune.

independent workers and the year of tax calculation were made consistent with the concept used in RDL. However, income measured in both sources, even if we get identical scopes, is not exactly the same because the social-security benefits in SILC are measured by matching with the social data sources (for the major part of the beneficiaries), whereas they are imputed in RDL thanks to a scale. Looking forward to new sources from Filosofi, this heterogeneousness, which we can consider as acceptable although it is penalizing, must be accepted.

In 2009, the national calibration produced 239 negative weights (the initial file includes 10 602 households) and in 2010 it produced 72 negative weights (for 11 044 observations). The following table gives, for SILC 2010, the distribution of the ratios of weights produced by the national calibration on the new margins (it is indeed the ratio obtained by putting in the denominator the weights already calibrated according to the initial margins). We obtain the same kind of outcome for 2009.

Table 2: Weights ratios produced by the national calibration on the new margins, 2010

Fractile		Ratio of weights 2010
Max	100%	3.96
	99%	1.86
	95%	1.48
	90%	1.36
Q3	75%	1.16
Median	50%	0.99
Q1	25%	0.85
	10%	0.68
	5%	0.52
	1%	0.12
Min	0%	-1.66

The table below gives the national estimations relative to the sizes of population, households as individuals, respectively before the new calibration (= current calibration) and after the new calibration. When we count physical individuals, it is possible to use either the file ‘households’ or the file ‘individuals’.

Table 3: National estimations of population sizes before and after the new calibration

Year	Number of households		Number of individuals			
	Current calibration	New calibration (= census)	Current calibration/ Source household	Current calibration/ Source individuals	New calibration/ Source household	New calibration/ Source individuals
2009([1])	26 997 136	26 866 278	60 665 878	60 673 238	60 997 867	61 021 387
2010	27 293 225	27 106 998	60 997 389	60 997 389	61 298 104	61 298 104

([1]) The small gaps about the estimation of the total number of individuals between what comes from the processing of the file 'individuals' (sum of the individual weights from the table ' individuals' — the weight of an individual is equal to the weight of its household) and what comes from the processing of the file 'households' (sum of the products of the households weights by the households sizes) are probably due to a (small) problem of unit identification in a file (an identifier not allocated in the processing chain). Actually, for some households, the size of the household in the file 'households' differs from the number of individuals in this household distinguished in the file 'individuals'). Anyway, those gaps are of very modest scale and affect merely the year 2009.

It can be surprising to notice a gap about these numbers — household as individuals — between the situation before the new calibration and the situation after the new calibration. Indeed, it is about numbers which play an essential role in the calculation of the poverty indicators. Things being what they are, the gap seems been naturally understandable by the fact that the initial calibration is made on the margins coming from the Labour force survey, while the new calibration uses the census ([10]).

The two following tables supply the estimations (in percentage) of the 6 indicators defined in part 2, for each ratio respectively before new calibration (column 'Current') and after new calibration (column 'New'). Remember that the estimation given in column 'Current' uses calibrated weights, but they are produced after a calibration on the margins currently used for the official national statistics.

Table 4: Poverty indicators, 2009
(%)

$\hat{\theta}_1$		$\hat{\theta}_2$		$\hat{\theta}_3$		$\hat{\theta}_4$		$\hat{\theta}_5$		$\hat{\theta}_6$	
Current	**New**	**Current**	**New**	**Current**	**New**	**Current**	**New**	**Current**	**New**	**Current**	**New**
12.89	12.77	13.55	13.63	5.56	5.82	8.36	8.04	18.47	18.40	1.19	1.29

Table 5: Poverty indicators, 2010
(%)

$\hat{\theta}_1$		$\hat{\theta}_2$		$\hat{\theta}_3$		$\hat{\theta}_4$		$\hat{\theta}_5$		$\hat{\theta}_6$	
Current	**New**	**Current**	**New**	**Current**	**New**	**Current**	**New**	**Current**	**New**	**Current**	**New**
13.28	13.78	12.62	12.47	5.79	5.8	9.89	9.28	19.17	19.45	1.52	1.44

The two following tables give the estimations of the number of poor people according to the 6 respective definitions of poverty. It is those numbers which are then divided by the estimation of the total number of physical persons to produce the poverty indicators.

([10]) An ultimate 'rule of three' could be practised so that the total number of households and/or individuals are equal before and after the calibration, what would make, in a general way, the comparisons between the estimations a little easier to interpret — but the numerical impact on the estimations of numbers would be minor and, involving here ratios, it would be equal to zero in this particular case.

Table 6: Number of poor people, 2009

According to θ_1		According to θ_2		According to θ_3	
Current	New	Current	New	Current	New
7 820 418	7 788 426	8 219 312	8 319 704	3 372 182	3 549 520
According to θ_4		According to θ_5		According to θ_6	
Current	New	Current	New	Current	New
3 873 144	3 816 054	11 207 632	11 229 969	721 499	788 592

Table 7: Number of poor people, 2010

According to θ_1		According to θ_2		According to θ_3	
Current	New	Current	New	Current	New
8 098 613	8 448 001	7 700 525	7 645 298	3 529 922	3 554 406
According to θ_4		According to θ_5		According to θ_6	
Current	New	Current	New	Current	New
4 584 882	4 393 619	11 692 708	11 920 864	924 246	881 022

Even if the order of magnitude is never seriously disturbed by the change of the calibration variables, we notice that there are rather subtle mechanisms which can create significant gaps between the situation before and after calibration. It is typically the case of the number of poor people according to θ_1 in 2010. But these gaps do not reproduce inevitably in the time — for instance typically the situation 2009 is completely different from the situation 2010 concerning the number of poor people according to θ_1 (the new calibration entailing in 2009, if it is not a decrease, at least a non-increase of the number of poor people). When we change the concept of poverty, there is still no logic in the evolutions of the respective indicators. For example in 2010 the new calibration increases appreciably the number of poor people according to θ_1, but at the same time it decreases appreciably the number of poor people according to θ_4. Also, it increases by 0,5 point the at-risk-of-poverty $\hat{\theta}_1$, what is considerable. The annual evolution is consequently impacted there: before calibration, the progress of the at-risk-of-poverty ratio between 2009 and 2010 is estimated at 0,4 point of percentage, but after calibration it is estimated at 1 point of percentage. It is thus necessary to remember that there is a rather large sensibility of the estimations (here the national ones) in the method of calibration.

At the regional level

The calibrations are made region by region: for every region, we consider the national file on one hand, the regional margins on the other hand, and thirdly the national weights standardized as explained in part 3. We look for new weights as close as possible of the standardized weights, which enable to retrieve exactly the local margins when we process the complete national file. Actually, **there is a set of weights by region — but this set can be used to estimate any total at the regional level, as soon as we treat a variable Y correlated with poverty** (more exactly, a variable Y well explained by the set of the calibration variables X).

No calibration failed ([11]) — even for the regions which have a specific structure according to X — what was expected because we use the linear method, which works in any circumstances. As a specific region, it is necessary to encompass Ile-de-France and Corse. For these two regions, the national estimation SILC is really distant or very distant from the regional margin. Two examples among a multitude of others, from SILC 2010:

– In region Ile-de-France, we count 7 629 farmers while the national file initially weighted (after standardization) gives a regional estimation of 84 079 farmers. We notice here an obvious fact: Ile-de-France is clearly less rural than the rest of France… The final set of weights has thus to 'twist' the national structures so that we find in fine 7 692 farmers (concretely, the weights of the farmers in the national file SILC are going to collapse);

– In region Corse, the complete file ASPA counts 9 934 beneficiaries, but the national file SILC initially weighted (after standardization) estimates at only 927 the number of those beneficiaries. The weights of the beneficiaries ASPA in the national file must be thus considerably increased so that we find the 'true' regional size.

In spite of these major — not to say spectacular — imbalances, the calibration was able to succeed in these regions as in all the others, in 2009 as well as in 2010, because the calibration function is linear. In return, we found a large number of negative weights (see part 3). The following table summarizes the situations 2009 and 2010 (remember that the total number of units concerned by the calibration in each region is the national sample size).

([11]) In some circumstances, the software Calmar may be unable to find a set of weights which satisfy the calibration constraint. It was not the case here, thanks to the choice of linear method.

Table 8: Number of negative weights

Region	Number of negative weights 2009	Number of negative weights 2010
11	1 632	1 849
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	1 926	1 883

It clearly emerges that the number of negative weights reflects the scale of the difference between the regional structure and the national structure. The more a region seems 'different' from whole France (from the point of view of the only variables of calibration), the more there are negative weights. Unsurprisingly, regions Ile-de-France and Corse distinguish themselves.

A method — probably the most convincing one — to assess the bias of the model consists in comparing the national estimated population size with the sum of the estimated regional population sizes. The national estimated population size from a large-size sample (national sample SILC) is a priori considered as a good quality estimate and thus serves as a reference. If the regional population sizes estimated by the 'small area' method have not the required quality, their sum is going to be away from the national target. The following tables give the relative error, in percentage, for 2009 then for 2010.

Table 9: Relative bias due to the model, 2009
(%)

Poverty θ_1	Poverty θ_2	Poverty θ_3	Poverty θ_4	Poverty θ_5	Poverty θ_6
-0.76	1.50	5.68	-1.71	-0.05	9.45

Reading: if we add the 22 regional estimations, the number of poor people according to θ_1 is equal to (1 – 0.0076) = 0,9924 times the total number of poor people directly estimated at the national level.

Table 10: Relative bias due to the model, 2010
(%)

Poverty θ_1	Poverty θ_2	Poverty θ_3	Poverty θ_4	Poverty θ_5	Poverty θ_6
4.91	-0.58	0.99	-4.23	2.32	-4.16

It is always delicate to objectively assess the scale of a (relative) bias. By experience, but it commits only the author of this document, I would say that there is nothing to worry about below 5% and that the situation remains still acceptable between 5% and 10%. In view of these results, anyway the questioning of the model does not justify itself. In particular, the bias of the main estimator for Eurostat, $\hat{\theta}_5$, remains — at least for the two concerned years — completely modest. The assessment of the bias can be pursued by the examination of the graphs given in the appendix 5: one axis concerns the regional direct estimations, and the other axis the 'small area' estimations. The region Corse was excluded from these graphs, which it disrupted too much. The lack of bias gives rise to a cloud of points which spreads out more or less symmetrically along the first line Y=X (caution, the criterion is the one of a symmetry of the cloud, not a closeness of the points to the first line). Considering the six clouds presented in the appendix 5, there is no reason to believe in a substantial bias and we can thus consider the model in a rather serene way.

It is nice that the sum of the regional estimations, for every concept of poverty, restores the national 'direct' estimation. That is why we applied a 'rule of three' — generally called benchmarking — which, from the initial 'small area' estimations, allows to assure this property of coherence with the national disseminated statistics. It was decided by the Insee to make the benchmarking associated with every estimation by using as a target the national estimation obtained from the new calibration. The strategy on this point is not obvious. For the production of indicators to be disseminated, there are also communication and coherence issues to be managed. Indeed, an alternative would have been to act in order to find again the number of poor people given by the official statistics (thus with the current calibration) — but it is true that this objective was much less defensible on the strictly technical aspect and the whole approach would not appeared as coherent. As regards the only at-risk-poverty-rate θ_1, it would even have been possible to start with the national rate of poverty from RDL and to reconstitute a target by multiplying it by the size of the population of individuals given by the census — what would mean a calibration on the national at-risk-of-poverty indicator θ_1 from RDL.

After the benchmarking and by construction, the relative bias due to the model, such as it was calculated above, becomes equal to zero ([12]). Both tables below give respectively for 2009 (SILC 2009, income 2008) and for 2010 (SILC 2010, income 2009) the regional estimations after benchmarking, that is the final regional indicators, ready for the dissemination.

Table 11: Estimation of the regional poverty indicators, 2009

Region	RDL 2008	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$
11	12.1	12.05	15.65	6.56	8.25	17.31	1.72
21	13.9	13.82	14.93	6.00	9.28	19.76	1.36
22	13.7	13.68	14.26	6.00	8.82	19.58	1.35
23	12.4	12.14	14.31	5.59	8.45	17.95	1.22
24	11.2	10.98	12.27	4.93	7.50	16.65	0.80
25	12.6	12.57	12.83	5.02	8.09	18.46	0.80
26	12.0	11.75	12.05	4.98	7.88	17.63	0.75
31	17.7	17.40	17.09	6.82	10.74	23.37	1.93
41	13.2	12.92	13.27	5.54	8.58	18.70	1.13
42	10.6	10.55	12.22	5.14	6.96	15.87	0.92
43	12.0	11.87	12.79	5.36	7.62	17.59	0.93
52	10.7	10.69	11.48	4.31	6.42	16.00	0.59
53	10.8	10.69	10.55	4.22	6.80	16.27	0.45
54	13.2	12.92	11.56	4.56	8.02	18.51	0.68
72	12.7	12.47	11.89	4.86	8.16	18.11	0.89
73	13.6	13.58	11.63	4.92	8.07	19.07	0.87
74	14.2	13.90	12.11	4.83	8.91	19.74	0.77
82	11.3	11.20	12.64	5.15	7.21	16.54	0.98
83	13.6	13.55	11.99	4.89	8.32	19.36	0.74
91	18.1	17.69	13.98	6.20	11.01	23.68	1.47
93	15.4	15.19	14.58	5.98	10.05	20.77	1.56
94	20.0	19.47	15.45	7.76	12.73	26.51	2.05

([12]) Such a display would be totally artificial, obviously the assessement of the bias due to the model has to be made **before** the benchmarking.

Table 12: Estimation of the regional poverty indicators, 2010

Region	RDL 2008	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$
11	12.5	12.43	13.50	6.26	9.43	17.70	1.88
21	14.6	14.26	14.43	6.58	11.33	20.97	1.88
22	14.4	14.05	13.65	6.46	10.83	20.52	1.94
23	13.0	12.79	13.58	6.14	10.26	19.16	1.64
24	11.8	11.58	11.60	5.22	9.25	17.50	1.22
25	13.3	13.06	12.52	5.62	9.99	19.41	1.36
26	12.5	12.37	11.66	5.30	9.73	18.47	1.25
31	18.6	18.04	16.42	7.68	12.85	25.03	2.43
41	13.9	13.52	12.62	5.82	10.47	19.82	1.56
42	11.3	11.00	10.78	4.84	8.44	16.45	1.14
43	12.9	12.66	12.26	5.56	9.63	18.71	1.39
52	11.2	11.04	11.15	4.66	7.85	16.84	0.75
53	11.2	11.05	10.23	4.47	8.34	16.82	0.79
54	13.8	13.49	11.63	5.18	9.79	19.63	1.09
72	12.9	12.50	11.20	5.10	9.25	18.39	1.07
73	14.0	13.52	11.08	5.18	9.44	19.19	1.25
74	14.7	14.16	11.80	5.34	10.73	20.41	1.20
82	11.8	11.84	11.71	5.25	8.66	17.33	1.24
83	14.0	13.62	11.73	5.26	10.08	19.77	1.17
91	18.6	17.93	13.54	6.69	12.97	24.15	2.13
93	15.8	15.25	13.56	6.34	11.16	21.20	1.71
94	19.3	18.57	14.18	7.77	13.98	24.91	2.91

We notice that the official poverty indicators from RDL disseminated on the website www.insee.fr (columns RDL 2008 and RDL 2009) are close to the indicators $\hat{\theta}_1$, what constitutes a type of validation of the regional weighting. We can think that these weigths react correctly when the other estimated indicators $\hat{\theta}_i$ are concerned.

The two following tables summarize the distribution of the 7 poverty indicators we are interested in, concerning respectively 2009 and 2010. It emerges from them that the methodology we use produces in fine a significant disparity of the regional situations.

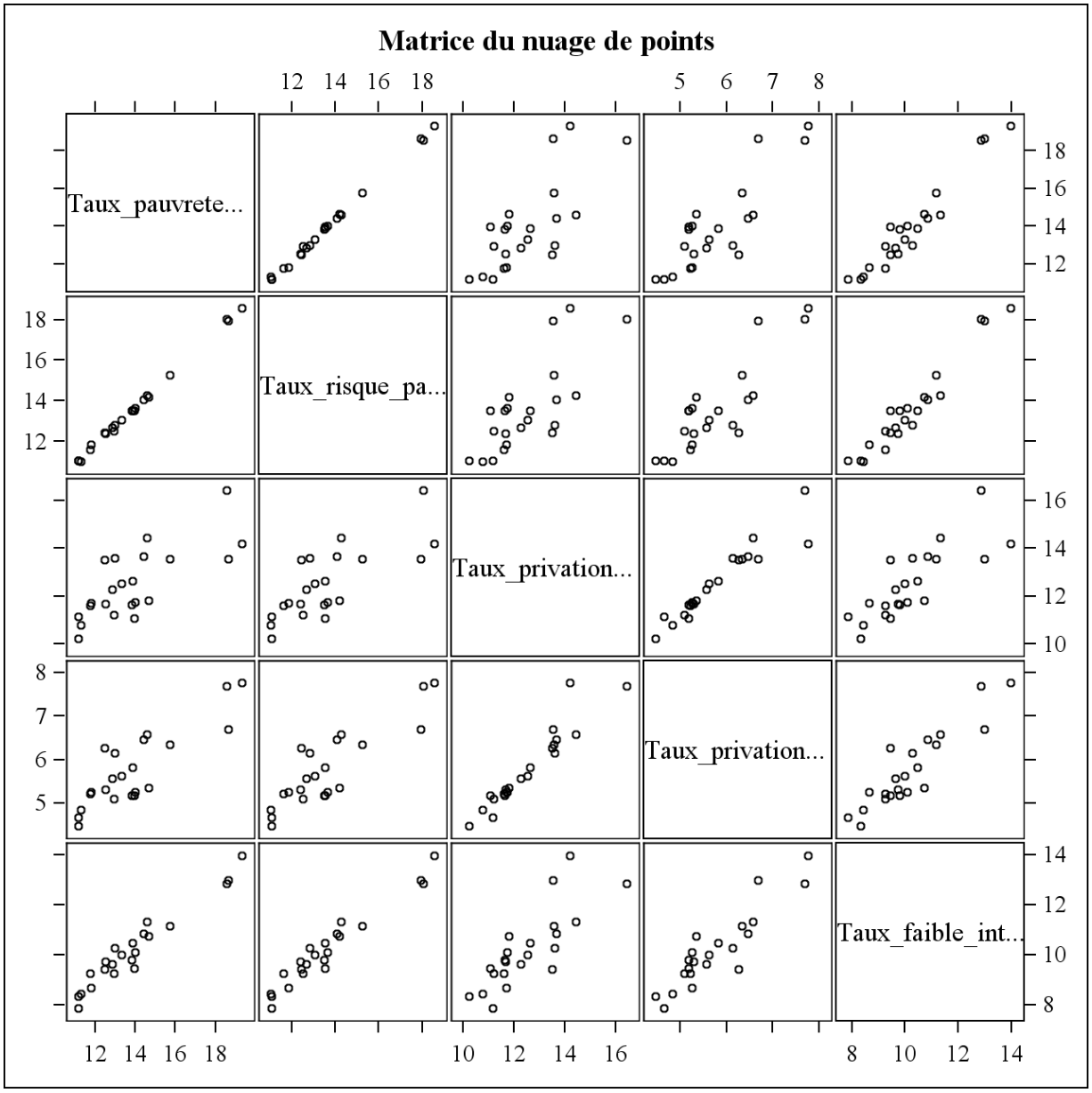
Table 13: Basic descriptive statistics about the regional poverty indicators, 2009

Variable	N	Mean	Standard-error	Minimum	Maximum
Poverty indicator from RDL 2008	22	13.41	2.50	10.60	20.00
$\hat{\theta}_1$	22	13.23	2.40	10.55	19.47
$\hat{\theta}_2$	22	13.16	1.70	10.55	17.09
$\hat{\theta}_3$	22	5.44	0.87	4.22	7.76
$\hat{\theta}_4$	22	8.54	1.50	6.42	12.73
$\hat{\theta}_5$	22	18.97	2.70	15.87	26.51
$\hat{\theta}_6$	22	1.09	0.44	0.45	2.05

Table 14: Basic descriptive statistics about the regional poverty indicators, 2010

Variable	N	Mean	Standard-error	Minimum	Maximum
Poverty indicator from RDL 2008	22	13.91	2.30	11.17	19.32
$\hat{\theta}_1$	22	13.58	2.20	11.00	18.57
$\hat{\theta}_2$	22	12.49	1.50	10.23	16.42
$\hat{\theta}_3$	22	5.76	0.89	4.47	7.77
$\hat{\theta}_4$	22	10.20	1.50	7.85	13.98
$\hat{\theta}_5$	22	19.65	2.50	16.45	25.03
$\hat{\theta}_6$	22	1.50	0.53	0.75	2.91

The following figure crosses 5 poverty indicators: the poverty indicator according to RDL and the indicators ranging from $\hat{\theta}_1$ to $\hat{\theta}_4$. The diagonal, from the left top corner to the right low corner has to be read in this order: rate of poverty RDL $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ and finally $\hat{\theta}_4$. The interest of this graph is to display all the clouds of points we can envisage by crossing 2 by 2 the 5 poverty indicators in question. Every point represents a region. We notice the excellent relationship between the poverty indicator RDL and the estimated indicator $\hat{\theta}_1$ (the cloud is almost a right line), and we can see that obviously the other concepts of poverty are quite correlated between them.



Just above, we give the matrix of the linear correlations when we cross the seven poverty indicators, for year 2010.

Table 15: Correlations between the different poverty indicators, 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	0.99	0.843
$\hat{\theta}_1$	0.999	1	0.761	0.87	0.97	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	0.87	0.948	1	0.911	0.874	0.982
$\hat{\theta}_4$	0.969	0.97	0.803	0.911	1	0.977	0.9
$\hat{\theta}_5$	0.99	0.991	0.788	0.874	0.977	1	0.844
$\hat{\theta}_6$	0.843	0.854	0.881	0.982	0.9	0.844	1

The following table (for the year 2010) compares the regional situations according to the various criteria of poverty when we attribute to every region a rank by criterion: we attribute the rank 1 to the richest region and the rank 22 to the poorest region. The matrix of rank correlation (not supplied here) is similar to the previous matrix.

Table 16: Ranks of the regions according to the poverty concept, 2010

Region	Rank						
	RDL	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$
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

We shall find in appendix 6, for information, the regional estimations for 2009 and 2010 concerning the at-risk-of-poverty rate ($\hat{\theta}_1$) and the main poverty indicator $\hat{\theta}_5$, obtained when we use the only part of the sample SILC which cuts across the region. It is about the regional estimations which are named ‘direct estimations’, which are the ones that we would obtain by a classic approach if we did not apply any ‘small area estimation’ method. We can notice that the distribution of the regional estimations is more narrow when we move from the ‘direct’ estimation to the model-based estimation. It is natural, not to say inevitable, and it attests some homogenization phenomenon intrinsically attached to the model — phenomenon known under the technical name of ‘shrinkage’. There is no evidence in the fact that the shrinkage is associated with a significant bias — as we saw above.

6. How to progress, perspectives

This short and final part summarizes some tracks for possible improvements in the future:

- we can hope to enrich the list of the calibration variables by mobilizing the local information about the social transfers AAH and RSA (see part 4), if however we succeed in isolating the beneficiaries living in a community;

- we can also hope, even if it seems a priori difficult, to build new margins around an information relative to the cost of living — in particular what concerns the cost of the dwelling (which could moreover be enough);

- there is probably a progress to hope by homogenizing more the concepts of standard of living, on one hand from SILC and on the other hand from RDL; the future data system named Filosofi should satisfy this expectation.

Some micro data from SILC 2009 have to be improved — we should not thus consider them as definitive data.

Besides, the timetable of dissemination of the French census outcomes constitutes a very penalizing constraint: so, we could investigate a way to constitute the regional margins from an accumulation of surveys during the year (including the Labour force survey — using the weights corrected by the non-response treatment but before calibration). The investment would be made profitable because this problem arises for any ‘local’ production, whatever the ‘small area estimation’ methodology is used.

Appendix 1: Codification of French regions

11 : Ile-de-France

21 : Champagne-Ardennes

22 : Picardie

23 : Haute-Normandie

24 : Centre

25 : Basse-Normandie

26 : Bourgogne

31 : Nord-Pas-de-Calais

41 : Lorraine

42 : Alsace

43 : Franche-Comté

52 : Pays de la Loire

53 : Bretagne

54 : Poitou-Charentes

72 : Aquitaine

73 : Midi-Pyrénées

74 : Limousin

82 : Rhône-Alpes

83 : Auvergne

91 : Languedoc-Roussillon

93 : Provence-Alpes-Côte d'azur

94 : Corse

Appendix 2: Technical development

The calibration program used (%Calmar tool) is the following one

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

$$\text{under the constraint : } \sum_{k \in s} w_k \cdot X_k = X_{REG}$$

The specificity of the context is due to the fact that s is the **national** responding sample (metropolitan France) and that the margins X_{REG} are the **regional** margins. These margins are vectorial, that means

$$X_k \in R^p \text{ and } X_{REG} \in R^p .$$

The mathematical solution of the program does not depend on the nature of the margin — whose interpretation has no consequence on the technical plan — so that we always have, when the distance function D verifies the 'usual' conditions of regularity:

$$\forall k \in s \; : \; w_k = d_k \cdot F(X_k^t \cdot \lambda)$$

where $\lambda \in R^p$ is an unknown vector to be determined by mobilizing the constraint and d_k is the weight of the household k before the calibration. So

$$\sum_{k \in s} d_k \cdot F(X_k^t \cdot \lambda) \cdot X_k = X_{REG}$$

Obviously, it seems that the weighting to be applied is not the national one which produces (in theory) unbiased estimators, otherwise $\sum_{k \in s} d_k \cdot X_k$ would estimate the national total of the X_k which is obviously disproportionate with the regional margin X_{REG}, and it would lead to an absurd solution. It is necessary to start with an estimation whose order of magnitude is similar to the regional margins, what is naturally possible by transforming the national weights as follows (it is a preliminary process of 'normalization'):

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

where N_{REG} is the size of the regional population and N the size of the national population (which are known both in practice). In this particular case, k spots a household, so the populations are populations of households and the values of the coefficients of normalization are given in appendix 4.

Afterward, we distinguish the linear method, which is the simplest one from the technical point of view, and the other methods (non linear methods).

Case 1: linear method, corresponding to $F(\times) = x + 1$

On a preliminary basis, we remind two useful vectorial relations:

i) $\forall u \in R^p, \forall \lambda \in R^p \quad \left(u^t \lambda\right) \cdot u = \left(u u^t\right) \cdot \lambda$

ii) For any symetric matrix M,

$\forall u \in R^p, \forall \lambda \in R^p \quad (Mv)^t \cdot u = (Mu)^t \cdot v$

The 't' exponent means 'transposition'; the vectors are column vectors.

With a linear function F, the constraint becomes simpler

$$\sum_{k \in s} d_k' \cdot X_k + \left(\sum_{k \in s} d_k' X_k X_k^t \right) \cdot \lambda = X_{REG}$$

We get $\sum_{k \in s} d_k' \cdot X_k = \frac{N_{REG}}{N} \cdot \sum_{k \in s} d_k \cdot X_k = \frac{N_{REG}}{N} \cdot \hat{X}_{NAT}$, where $\hat{X}_{NAT}$ estimates the national total of the X_k.

With the weights d_k' as inputs given to %Calmar (which mechanically starts with the set of weights it is fed with …), one get so

$$\lambda = \left(\sum_{k \in s} d_k' X_k X_k^t \right)^{-1} \cdot \left(X_{REG} - \frac{N_{REG}}{N} \cdot \hat{X}_{NAT} \right)$$

$$\lambda = \left(\frac{1}{N} \sum_{k \in s} d_k X_k X_k^t \right)^{-1} \cdot \left(\overline{X}_{REG} - \frac{1}{N} \cdot \hat{X}_{NAT} \right)$$

Then

$$w_k = d_k' \cdot \left(1 + X_k^t \lambda\right) = d_k' + d_k' X_k^t \lambda$$

Considering the very large size of the sample s, considering the hypotheses of convergence which we can reasonably make, we postulate (U is the national population of the ordinary households)

$$\frac{1}{N}\sum_{k\in s} d_k X_k X_k^t \approx \frac{1}{N}\sum_{k\in U} X_k X_k^t \quad \text{and} \quad \frac{\hat{X}_{NAT}}{N} \approx \overline{X}_{NAT}$$

where $\overline{X}_{NAT}$ indicates the true national mean of X_k (k described the national population U) so that the vector λ must be numerically very close to the deterministic vector λ_0:

$$\lambda_0 = \left(\frac{1}{N}\sum_{k\in U} X_k X_k^t\right)^{-1} \cdot \left(\overline{X}_{REG} - \overline{X}_{NAT}\right)$$

where $\overline{X}_{REG}$ indicates the true (known) regional mean of the X_k. Clearly, allowing for exceptions, we have $\lambda_0 \neq 0$.

The final 'small area' estimator $\hat{Y}_{REG}^{SAE}$ of the regional total $\sum_{k\in REG} Y_k$ is built from the national sample s by using the calibrated weights w_k what gives formally

$$\hat{Y}_{REG}^{SAE} = \sum_{k\in s} w_k \cdot Y_k$$

Considering the expression of the calibrated weights, we obtain

$$\hat{Y}_{REG}^{SAE} = \sum_{k\in s} d_k' \cdot Y_k + \sum_{k\in s} d_k' \left(X_k^t \lambda Y_k\right) = \sum_{k\in s} d_k' \cdot Y_k + \sum_{k\in s} d_k' \left(\lambda^t X_k Y_k\right)$$

$$\hat{Y}_{REG}^{SAE} = \sum_{k\in s} d_k' \cdot Y_k + \left(\left(\sum_{k\in s} d_k' X_k X_k^t\right)^{-1} \cdot \left(X_{REG} - \frac{N_{REG}}{N} \cdot \hat{X}_{NAT}\right)\right)^t \left(\sum_{k\in s} d_k' X_k Y_k\right)$$

$$\hat{Y}_{REG}^{SAE} = \sum_{k\in s} d_k' \cdot Y_k + \left(\left(\sum_{k\in s} d_k' X_k X_k^t\right)^{-1} \cdot \left(\sum_{k\in s} d_k' X_k Y_k\right)\right)^t \cdot \left(X_{REG} - \frac{N_{REG}}{N} \cdot \hat{X}_{NAT}\right)$$

We define

$$\hat{B} = \left(\sum_{k\in s} d_k' X_k X_k^t\right)^{-1} \cdot \left(\sum_{k\in s} d_k' X_k Y_k\right)$$

a vector in R^p which can be simplified :

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

It is the usual estimator of the coefficient of regression B from the linear regression of Y_k on the vector X_k, with

$$B = \left(\sum_{k \in U} X_k X_k^t \right)^{-1} \cdot \left(\sum_{k \in U} X_k Y_k \right)$$

We insist on the fact that the estimation $\hat{B}$ of B comes from the national sample considered as a whole. From there, we get:

$$\hat{Y}_{REG}^{SAE} = \frac{N_{REG}}{N} \cdot \sum_{k \in s} d_k \cdot Y_k + \hat{B}^t \cdot \left(X_{REG} - \frac{N_{REG}}{N} \cdot \hat{X}_{NAT} \right)$$

being given the definition of the weights d_k, the term $\hat{Y}_{NAT} = \sum_{k \in s} d_k \cdot Y_k$ estimates the true national total of the Y_k. So

$$\hat{Y}_{REG}^{SAE} = \frac{N_{REG}}{N} \cdot \hat{Y}_{NAT} + \hat{B}^t \cdot \left(X_{REG} - \frac{N_{REG}}{N} \cdot \hat{X}_{NAT} \right)$$

Important :

We expect that the order of magnitude of the regional margin X_{REG} is similar to the order of magnitude of $\frac{N_{REG}}{N} \cdot \hat{X}_{NAT}$: indeed, $\overline{X}_{REG} = \frac{X_{REG}}{N_{REG}}$ is equal to the (true) regional mean of X_k whereas $\frac{\hat{X}_{NAT}}{N}$ is equal to the (estimated) national mean of the X_k. It justifies the equality of the order of magnitude. On the other hand, it is very clear that there is no reason for the difference between both values to be close to zero: to believe that would be to deny the existence of regional specificities. By experience, we notice moreover very clearly that there are significant differences between regions, and thus the term $\left(X_{REG} - \frac{N_{REG}}{N} \cdot \hat{X}_{NAT} \right)$ is absolutely not equal to zero 'on average', contrary to what we are used to see in the traditional calibrations where we have to deal with a corrective coefficient which is very small (and especially when the sample size is large).

Things being what they are, on the purely calculation aspect, we can write:

$$\hat{Y}_{REG}^{SAE} = \hat{B}^t \cdot X_{REG} + \frac{N_{REG}}{N} \cdot \left(\hat{Y}_{NAT} - \hat{B}^t \cdot \hat{X}_{NAT} \right)$$

Yet, we know that as soon as the constant is one of the regressors (more generally as soon as the constant is a linear combination of the regressors) we have the property (consider the regression conceived with the national sample):

$$\hat{Y}_{NAT} = \hat{B}^t \cdot \hat{X}_{NAT}$$

so that *in fine*

$$\boxed{\hat{Y}_{REG}^{SAE} = \hat{B}^t \cdot X_{REG}}$$

When the calibration on the regional margins conceived from the weights previously standardized is finished, and when we use the calibrated weights given by %Calmar**, we build *de facto* and formally the estimator above, which is nothing else than the very classic synthetic estimator** formed from the vectorial auxiliary information X . This estimator is justified when we have a 'more or less linear' relationship on the whole territory — thus, whatever is the region — between Y and X . It is built on the whole national sample s , so that **it has a low sampling variance**, but (obviously) **in counterpart it has some bias**.

Important :

The synthetic estimator can obviously be calculated directly through a classic linear regression with the national sample s , by using as an input the coefficients of regression $\hat{B}$ and by making a simple product with the regional margins — it is not very difficult with a software (Proc Reg in SAS) and we cannot really meet of unpleasant surprise in this process. Nevertheless, it is better to take advantage of the tool %Calmar (or any other calibration software), very easy to use and to access. In particular, if there is an important number of variables of interest Y to be treated, the standard approach would consist in processing a regression on every variable and this can turn out cumbersome in practice, especially if the user is not familiar with the regression theory. In return, the calibration is made only once and one uses the (unique) set of weights w_k with all the variables of interest Y_k because these calibrated weights w_k depend only on the auxiliary information X_k : it is a determining asset!

But there is a counterpart because with the experience, it appears necessary to indicate a very particular and important point: the calibration approach with the linear method is surprising, not to

say unpleasant, because we can easily get a multitude of negative weights w_k. Now, we are in the habit of being afraid by the negative weights because they have no interpretation and because we cannot reasonably leave them in a file used for data dissemination. But in this particular case, you should not worry about it: in itself, that has no importance because in fine we obtain a numerical estimation which, anyway, is actually $\hat{B}^t \cdot X_{REG}$ and it is the only thing that matters: the result is mathematically exact even if the method is lacking aestheticism from this point of view ! This position seems to me in any case strong since we do not put at the disposal of a large number of users the weighted national file (or then, it is necessary to take some precautions regarding communication). Moreover, it comes a little in contradiction, unfortunately, with the very attractive idea to offer a 'universal' system of estimation from a system of weights calculated once and for all. The presence of negative weights is the direct consequence of the diversity of the field: there are negative weights because there are some components of the vector λ which must be strongly negative (necessary condition), that is in fact some components of the vector $\left(X_{REG} - \frac{N_{REG}}{N} \cdot \hat{X}_{NAT} \right)$. It is rather easy to imagine as soon as there are substantial gaps between the real regional mean $\overline{X}_{REG}$ and the real national mean $\overline{X}_{NAT}$ (probably very well estimated by $\frac{\hat{X}_{NAT}}{N}$). It is obviously a very remarkable difference of context with regard to the usual context of calibration, where the gap between X and $\hat{X}$ is conceived to be quite small…

Case 2: non linear methods

We can also calibrate the weights with a non linear method (in %Calmar for example, there are three non linear options presently programmed). Nevertheless, we know that in every case the function F has the following fundamental property: $F(0)=1$ and $F'(0)=1$ — because it is this property which is on the base of the fundamental theorem of calibration which, in the usual context, makes all the weighting systems asymptotically equivalent to the set of weights given by the linear method (which gives formally the well-known estimator called 'regression estimator'). In this case, this theorem is no longer valid. What follows does not correspond to a rigorous demonstration, but rather intuitions (that I consider obviously right!) which would require to be very methodically validated. Everything starts with the constraint

$$\sum_{k \in s} d_k' \cdot F(X_k^t \cdot \lambda(s)) \cdot X_k = X_{REG} .$$

The large size of the sample s allows to apply the asymptotic conditions (we have no reason for being afraid by the variability issue, but rather to fight the biases of the estimators), and we can imagine easily

that the vector $\lambda(s)$— I indexed it by s for the occasion — converges (in probability) towards an unknown vector, which should verify the following equation (system of p equations with p unknown values if there are p auxiliary variables):

$$\frac{N_{REG}}{N}\sum_{k\in U}F(X_k^t\cdot\lambda_0)\cdot X_k = X_{REG}$$

where U is the national population of the ordinary households. It comes also from the fact that, the ratio $\frac{N_{REG}}{N}$ being bounded, the estimators using the weights d_k converge towards the real values which they estimate. Unfortunately, it does not seem possible to go farther and to clarify more λ_0. In other words,

$$\frac{1}{N}\sum_{k\in U}\left(F(X_k^t\cdot\lambda_0)-1\right)\cdot X_k = \bar{X}_{REG}-\bar{X}_{NAT}$$

Except miracle, $\lambda_0\neq 0$ because otherwise we would have $F(X_k^t\cdot\lambda_0)-1=0$ and thus $\bar{X}_{REG}=\bar{X}_{NAT}$, what is undoubtedly wrong ! Because the size of s is large enough so that $\lambda(s)$ is close of λ_0, we have

$$F(X_k^t\cdot\lambda(s)) = F(X_k^t\cdot\lambda_0)+F'(X_k^t\cdot\lambda_0)\cdot X_k^t(\lambda(s)-\lambda_0)+O_p\left(\frac{1}{n}\right)$$

because the order of magnitude of the variance of $\lambda(s)$ is $\frac{1}{n}$. In this asymptotic context

$$\forall k\in s:\ w_k \approx d_k^{'}\cdot F(X_k^t\cdot\lambda_0)+d_k^{'}\cdot F'(X_k^t\cdot\lambda_0)\cdot X_k^t(\lambda(s)-\lambda_0)+O_p\left(\frac{1}{n}\right)$$

The main part of the weight, so $d_k^{'}\cdot F(X_k^t\cdot\lambda_0)$, can be pretty different of $d_k^{'}$ — as in the linear case moreover. It is unpleasant to notice that we do not control the value of the derivative $F'(X_k^t\cdot\lambda_0)$, which has no reason for being close to 1 (even if we can hope that it does not go away from it … probably it largely depends on the gap between the means $\bar{X}_{REG}$ and $\bar{X}_{NAT}$).

The constraint being respected for any sample s, we have

$$\left(\sum_{k\in s}d_k^{'}\cdot F'(X_k^t\cdot\lambda_0)\cdot X_k X_k^t\right)\cdot(\lambda(s)-\lambda_0)=$$

$$X_{REG} - \sum_{k \in s} d_k' \cdot F(X_k^t \cdot \lambda_0) X_k + O_p\left(\frac{N}{n}\right)$$

Obviously, the vector $\lambda(s)$ behavies as

$$\lambda_0 + \left(\frac{N_{REG}}{N} \sum_{k \in s} d_k \cdot F'(X_k^t \cdot \lambda_0) \cdot X_k X_k^t \right)^{-1} \cdot \left(X_{REG} - \frac{N_{REG}}{N} \sum_{k \in s} d_k \cdot F(X_k^t \cdot \lambda_0) \cdot X_k \right)$$

the dropped terms varying like $\frac{1}{n}$. The inverted matrix has not a bad look but it is complicated by the presence of some unusual individual weights $F'(X_k^t \cdot \lambda_0)$, which in this particular case are no longer equal to 1, whom the exact writing (that we could obtain by taking into account the analytical expression of F) will bring no simplification, and which have no interpretation … We also verify that the vector in brackets converges towards 0, considering the definition of λ_0.

At this stage, the conclusion is that the 'small area' estimator built with any other method that the linear method is at the moment without obvious connection with the estimator coming from the linear method ... thus a priori we still have no asymptotic equivalence of the calibration methods. The determining property at the heart of the theory of classic calibration is due to the fact that asymptotically all the calibration functions behave as the linear function. However, it is not true anymore here, and the gap between λ_0 and zero is the essential reason for that. Besides, for the moment, we have no theoretical justification to use any other calibration technique that the one which is associated with the linear method.

The 'small area' estimator is $\hat{Y}_{REG}^{SAE} = \sum_{k \in s} d_k' F(X_k^t \cdot \lambda) Y_k$, so

$$\hat{Y}_{REG}^{SAE} = \sum_{k \in s} d_k' F(X_k^t \cdot \lambda_0) Y_k + (\lambda - \lambda_0)^t \cdot \sum_{k \in s} d_k' F'(X_k^t \lambda_0) X_k Y_k + O_p\left(\frac{N}{n}\right)$$

Considering the right side of the equality, the first term is $O_p(N)$, the second one is $O_p\left(\frac{N}{\sqrt{n}}\right)$, what allows to eliminate definitively the third term by considering an 'asymptotic equivalent' of the calibrated estimator:

$$\hat{Y}_{REG}^{SAE} \approx \frac{N_{REG}}{N} \left(\sum_{k \in s} d_k Y_k + \sum_{k \in s} d_k (F(X_k^t \cdot \lambda_0) - 1) \cdot Y_k \right) +$$

$$\left(X_{REG} - \frac{N_{REG}}{N} \sum_{k \in s} d_k \cdot F(X_k^t \cdot \lambda_0) \cdot X_k \right)^t \cdot \left(\frac{N_{REG}}{N} \sum_{k \in s} d_k \cdot F'(X_k^t \cdot \lambda_0) \cdot X_k X_k^t \right)^{-1} \cdot \left(\frac{N_{REG}}{N} \sum_{k \in s} d_k \cdot F'(X_k^t \cdot \lambda_0) \cdot X_k Y_k \right) \cdot$$

We define

$$\hat{B}_F = \left(\sum_{k \in s} d_k \cdot F'(X_k^t \cdot \lambda_0) \cdot X_k X_k^t \right)^{-1} \cdot \left(\sum_{k \in s} d_k \cdot F'(X_k^t \cdot \lambda_0) \cdot X_k Y_k \right)$$

It is a coefficient of regression, weighted by the derivatives $F'(X_k^t \cdot \lambda_0)$, the interpretation of which is not obvious, but we can hope (nothing more !), if the means $\overline{X}_{REG}$ and $\overline{X}_{NAT}$ are not too different, that $\hat{B}_F$ is numerically 'not too far' from the $\hat{B}$ defined in the first part.

$$\hat{Y}_{REG}^{SAE} \approx \frac{N_{REG}}{N} \left(\sum_{k \in s} d_k Y_k \right) + \frac{N_{REG}}{N} \left(\sum_{k \in s} d_k (F(X_k^t \cdot \lambda_0) - 1) \cdot Y_k \right) +$$
$$\hat{B}_F^t \cdot \left(X_{REG} - \frac{N_{REG}}{N} \sum_{k \in s} d_k X_k - \frac{N_{REG}}{N} \sum_{k \in s} d_k \cdot \left(F(X_k^t \cdot \lambda_0) - 1 \right) \cdot X_k \right).$$
$$\hat{Y}_{REG}^{SAE} \approx \hat{B}_F^t \cdot X_{REG} + \frac{N_{REG}}{N} \left(\sum_{k \in s} d_k Y_k - \hat{B}_F^t \cdot \sum_{k \in s} d_k X_k \right) +$$
$$+ \frac{N_{REG}}{N} \sum_{k \in s} d_k \left(F(X_k^t \cdot \lambda_0) - 1 \right) \cdot \left(Y_k - \hat{B}_F^t \cdot X_k \right).$$

So finally

$$\boxed{\hat{Y}_{REG}^{SAE} \approx \hat{B}_F^t \cdot X_{REG} + \frac{N_{REG}}{N} \sum_{k \in s} d_k F(X_k^t \cdot \lambda_0) \cdot \left(Y_k - \hat{B}_F^t \cdot X_k \right)}$$

We thus have the analytical expression of an asymptotic equivalent of the calibrated estimator. It seems difficult to clarify it more. It would be advisable to assess the limit of this estimator. We can believe in the convergence of the estimtors weighted by the d_k, so that

$$\hat{B}_F \to B_F = \left(\sum_{k \in U} F'(X_k^t \cdot \lambda_0) \cdot X_k X_k^t \right)^{-1} \cdot \left(\sum_{k \in U} F'(X_k^t \cdot \lambda_0) \cdot X_k Y_k \right)$$

So

$$\hat{Y}_{REG}^{SAE} \to Y_{REG,F}^{lim} = B_F^t \cdot X_{REG} + \frac{N_{REG}}{N} \sum_{k \in U} F(X_k^t \cdot \lambda_0) \cdot \left(Y_k - B_F^t \cdot X_k\right)$$

Considering the definition of λ_0 given at the beginning of this part, we have

$$\boxed{Y_{REG,F}^{lim} = \frac{N_{REG}}{N} \sum_{k \in U} F(X_k^t \cdot \lambda_0) \cdot Y_k}$$

We note that the term B_F conveniently disappeared from this limit value. If $F(x) = x + 1$, then $B_F = B$ and we find the conclusions of the linear method, in particular $\hat{Y}_{REG}^{SAE}$ is the synthetic estimator $\hat{B}^t \cdot X_{REG}$ and $Y_{REG,F}^{lim}$ is equal to $B^t \cdot X_{REG}$. If the opposite occurs, we obtain an original estimator, maybe numerically rather close to the synthetic estimator, maybe it is preferable to the synthetic estimator in certain circumstances, but in any case the better Y_k is explained by the vector X_k, the better this estimator is justified. Things being what they are, we can say that **we consider to be right away and by definition in this configuration,** otherwise we could not justify the 'small area' synthetic estimator. If we agree to use $\hat{B}^t \cdot X_{REG}$, it is precisely because we believe in the following relation:

$$\exists B \in R^p \text{ so that } \forall k \in U : Y_k = B^t . X_k + U_k$$

with, either U_k 'small' (in a context of finite population), or $(U_k)_{k \in U}$ i.i.d. in an infinite population, otherwise one could not accept the resultant bias. It means that in any circumstance, since we agree to apply a 'small area' calibration method by using a linear function, the use of another calibration function is going to give an estimation close to

$$Y_{REG,F}^{lim} = \frac{N_{REG}}{N} \sum_{k \in U} F(X_k^t \cdot \lambda_0) \cdot \left(B^t \cdot X_k + U_k\right) =$$
$$B^t \cdot X_{REG} + \frac{N_{REG}}{N} \sum_{k \in U} F(X_k^t \cdot \lambda_0) \cdot U_k$$

Compared with the linear method, which gives a limit $Y_{REG,lin}^{lim} = B^t \cdot X_{REG}$, the **relative** algebric gap in term of bias becomes:

$$\Delta_F = \frac{Y_{REG,F}^{lim} - Y_{REG,lin}^{lim}}{Y_{REG,lin}^{lim}} = \frac{1}{B^t \cdot X_{REG}} \cdot \frac{N_{REG}}{N} \sum_{k \in U} F(X_k^t \cdot \lambda_0) \cdot U_k$$

So

$$\Delta_F = \frac{1}{B^t \cdot \bar{X}_{REG}} \cdot \frac{1}{N} \sum_{k \in U} F(X_k^t \cdot \lambda_0) \cdot U_k$$

When the population sizes become very big, the order of magnitude of Δ_F is that of the term $\frac{1}{N} \sum_{k \in U} F(X_k^t \cdot \lambda_0) \cdot U_k$. When we use a modelisation approach in a finite population and when the variable Y_k is quantitative, we consider that the mathematical expectation of the residual with regard to the model distribution is zero, so $\varepsilon U_k = 0$ (once again, it is the hypothesis which we have to make in order to justify the synthetic estimator, it is thus basic and implicit at once — the case of a qualitative variable Y_k raises nevertheless a problem to solve this issue). If we consider that the residuals U_k have a bounded variance $VU_k = \sigma_k^2$ (with regard to the model distibution), the 'law of large numbers' says that $\frac{1}{N} \sum_{k \in U} F(X_k^t \cdot \lambda_0) \cdot U_k$ converges towards $\frac{1}{N} \sum_{k \in U} F(X_k^t \cdot \lambda_0) \cdot \varepsilon U_k = 0$ because naturally N is very large. The conclusion is the following one

$$\Delta_F = \frac{Y_{REG,F}^{lim} - Y_{REG,lin}^{lim}}{Y_{REG,lin}^{lim}} \approx 0$$

When we use a model of behavior and when this model is right, this result justifies the use of another function F than the linear function. We thus find, in this new context, the fundamental property of calibration: **when the sample sizes (and the population sizes) are very large, subject to an exact model, the choice of the function F has no importance in term of bias (which is the essential component of the error). Contrary to the classic theory of calibration, we need there explicitly to use a model of behavior which constitutes a real hypothesis. If this model is false, the calibration with a non linear function loses its justification.**

However, if the model is false, that is if the variable of interest is 'badly explained' by the linear combination of the calibration variables, the linear method still produces a synthetic estimator but this last one loses in its turn its relevance. From the operational point of view, it results that it's better to use the linear method, which is the only one to preserve a simple interpretation to the calibrated estimator — the other methods giving a priori nothing more in term of efficiency.

These considerations apply naturally to the quantitative variables of interest, on the other hand the treatment of a qualitative variable of interest raises a theoretical problem — even if I still have the intuition that in practice the method remains as valid as in the quantitative case.

Appendix 3: Auxiliary variables used for margins

(source : census)

Individual level variables

Social category

(Codes: see 'PCS 2003: nomenclature des professions et catégories socioprofessionnelles' — *insee website*)

Modality 1 : ('10','11','12','13')

Modality 2: ('21','22','23')

Modality 3: ('31','33','34','35','37','38')

Modality 4: ('42','43','44','45','46','47','48')

Modality 5: ('52','53','54','55','56')

Modality 6: ('62','63','64','65')

Modality 7: ('67','68')

Modality 8: ('69')

Modality 9: ('71','77','78')

Modality 10: ('72','74','75')

Modality 11: ('81','83','84','85','86')

Age

Modality 1: 14 years old or less — at Dec 31th for the year of survey

Modality 2: between 15 years old (included) and 29 years old (included) — at Dec 31th

Modality 3: from 30 years old (included) to 39 years old (included) — at Dec 31th

Modality 4: from 40 years old (included) to 49 years old (included) — at Dec 31th

Modality 5: from 50 years old (included) to 59 years old (included) — at Dec 31th

Modality 6: 60 years old or more — at Dec 31th

Diploma

Modality 1: people 20 years old and less — at 31/12, for the year of survey

Modality 2: diploma <= BEPC, and 21 years old or more — at 31/12

Modality 3: diploma > BEPC and <= BAC (or BP or BT), and 21 years old or more — at 31/12

Modality 4: diploma > BAC (or BP or BT), and 21 years old or more — at 31/12

Nationality

Modality 1: people aged 15 years old or less, January 1st of the year of survey

Modality 2: French people, aged 16 years old or more, January 1st of the year of survey

Modality 3: European (except French) people, aged 16 years old or more, January 1st of the year of survey

Modality 4: African people, aged 16 years old or more, january 1st of the year of survey

Modality 5: People from Asia, Americs and Oceany, aged 16 years old or more, January 1st of the year of survey

Household level variables

Urban area size

Modality 1: UA <= 10 000 inhabitants

Modality 2: UA with 10 000 to 100 000 inhabitants

Modality 3: UA with 100 000 inhabitants and more (included the urban area of Paris)

Household type

Modality 1: person living alone

Modality 2: monoparental (man or woman, alone with a child or several children)

Modality 3: couple without child (two persons in the same household)

Modality 4: couple with child(ren) — but nobody else in the household

Modality 5: complex household

Rent (or not) a HLM dwelling

Modality 1: STOCD = 22 (census code / corresponds to an effective rent in the requested conditions)

Modality 2: other cases

Appendix 4: Regional normalizing ratios used to calibrate the weights

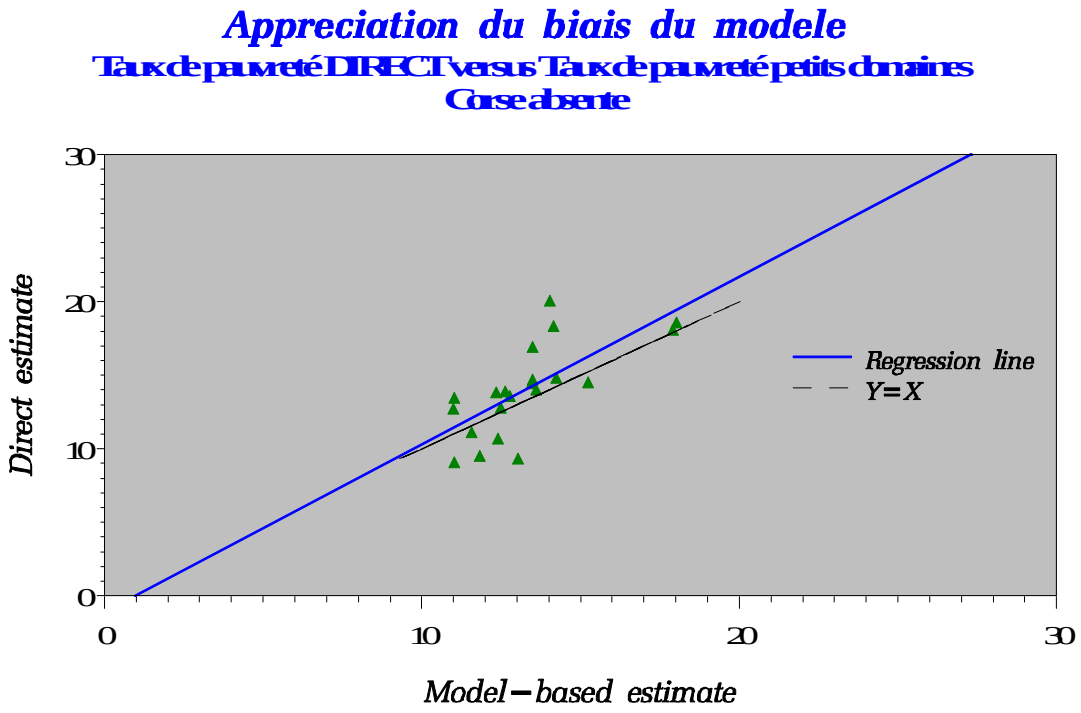
The estimation of the total number of households, in metropolitan France, obtained from SILC for the year 2010 is equal to 27 293 224.

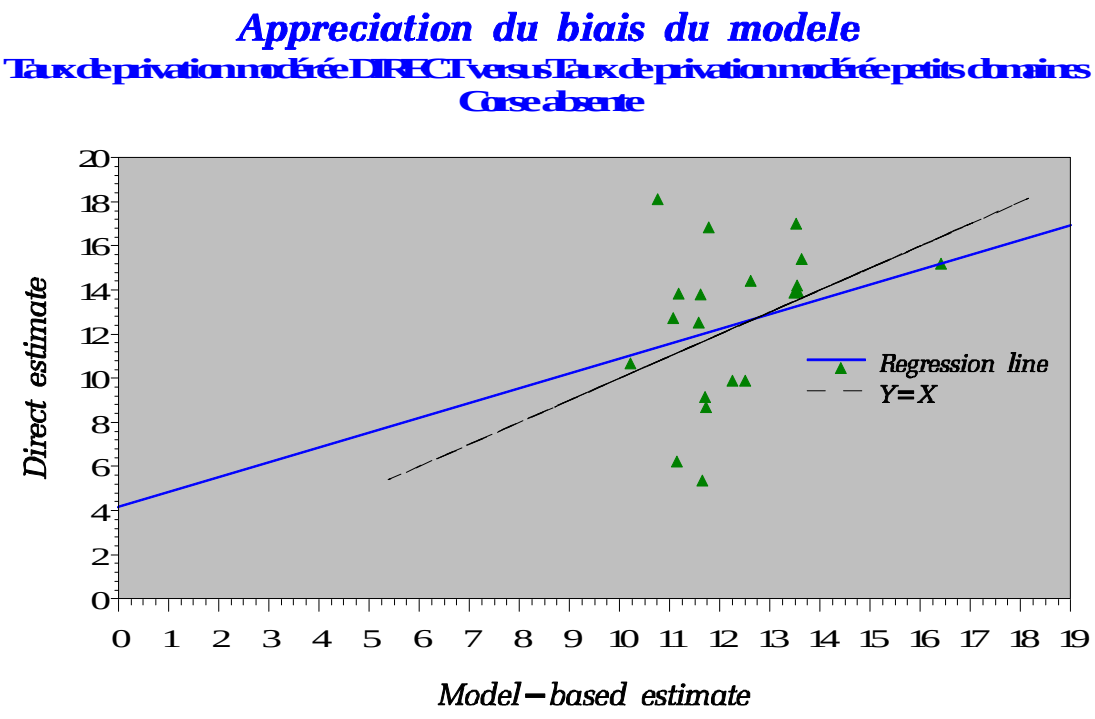
The first column (*somme_poids_region*) gives the estimations of the total number of households by region in 2010, according to the census.

The coefficient of normalization of the weights, named '**ratio**' in the table, is equal to the ratio of the first column on 27 293 224, this number estimating the total number of common households in the metropolitan France (year 2010).

Region	Somme_poids_region	Ratio
11	4 789 230	0.17547
21	737 445	0.02702
22	1 017 567	0.03728
23	691 174	0.02532
24	970 493	0.03556
25	635 470	0.02328
26	736 965	0.02700
31	2 008 733	0.07360
41	1 178 787	0.04319
42	683 975	0.02506
43	639 559	0.02343
52	1 808 780	0.06627
53	1 457 935	0.05342
54	852 799	0.03125
72	1 622 059	0.05943
73	1 243 916	0.04558
74	390 169	0.01430
82	2 157 268	0.07904
83	617 829	0.02264
91	1 113 283	0.04079
93	1 854 661	0.06795
94	85 117	0.00312

Appendix 5: Graphics to assess the bias

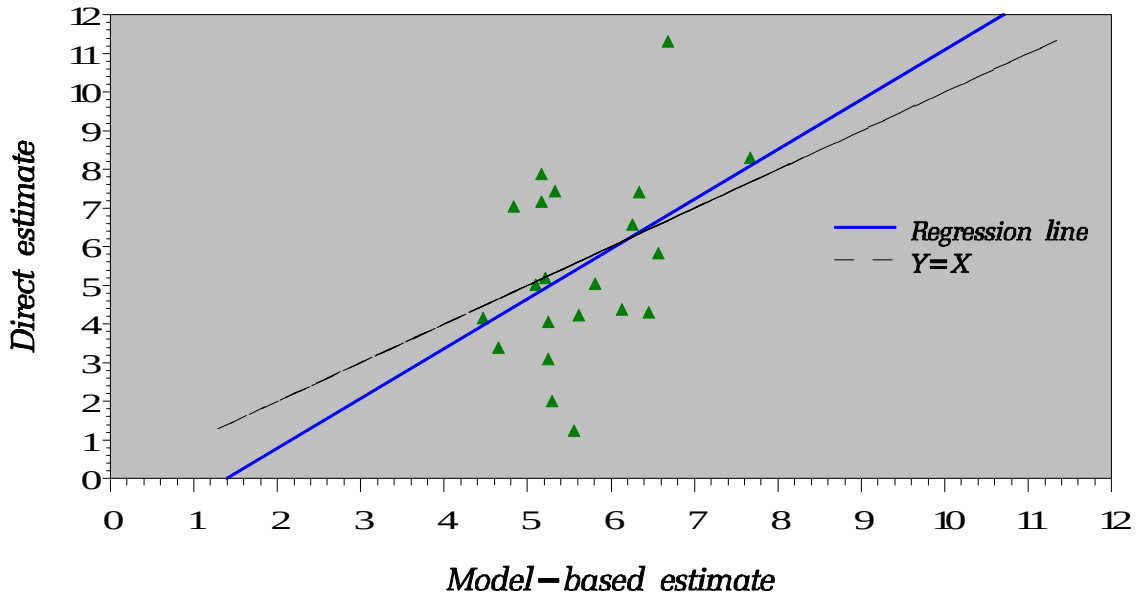




Appreciation du biais du modele

Taux de privation forte DIRECT versus Taux de privation forte petits domaines

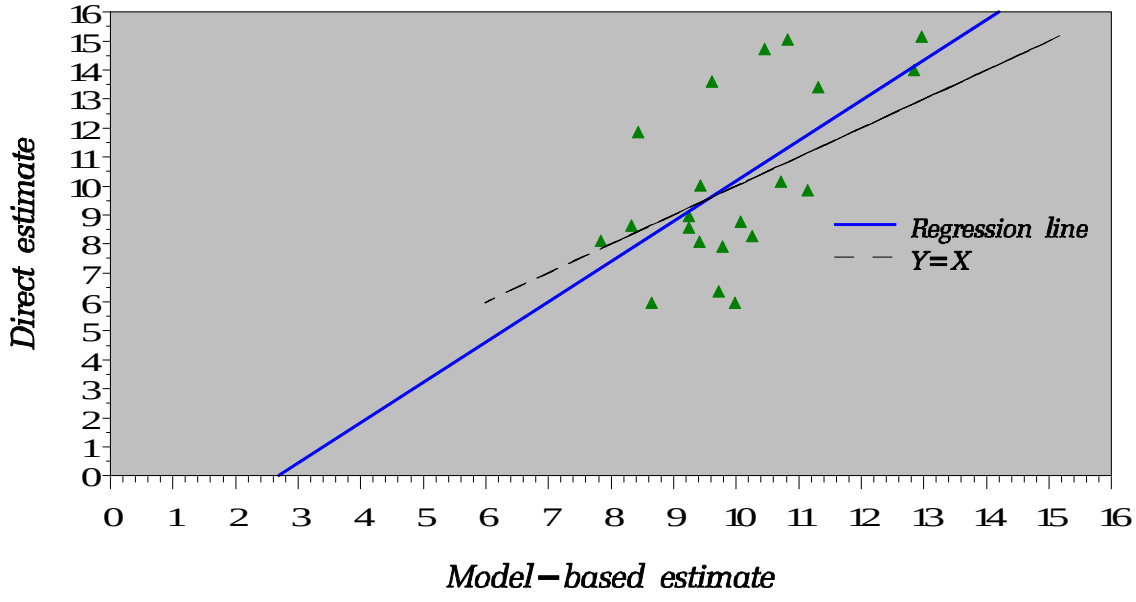
Corse absente



Appreciation du biais du modele

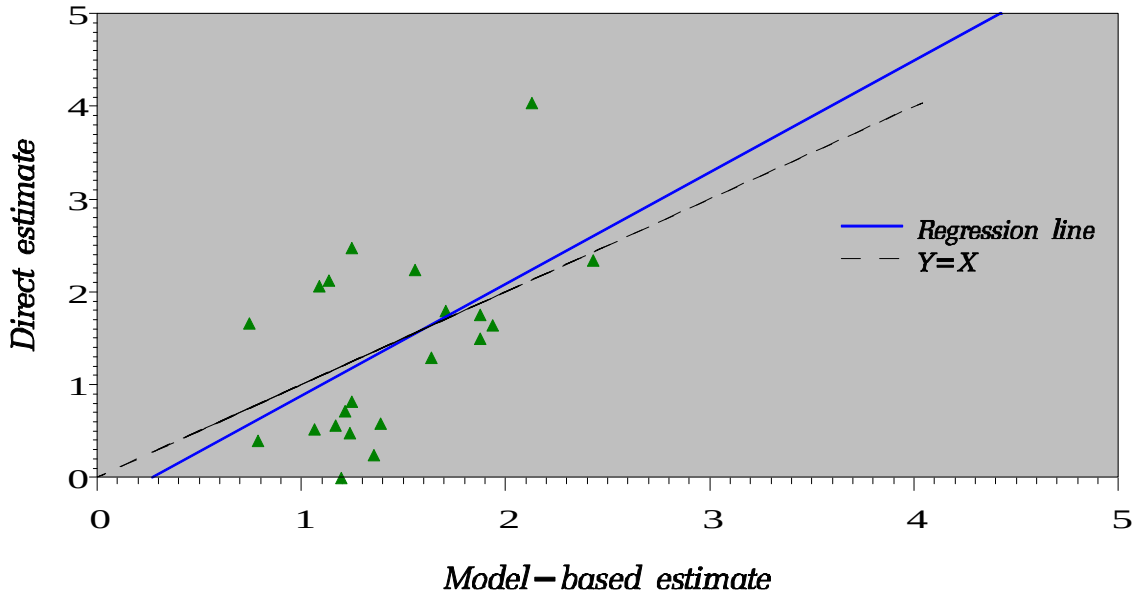
Taux de faible intensité DIRECT versus Taux de faible intensité petits domaines

Corse absente



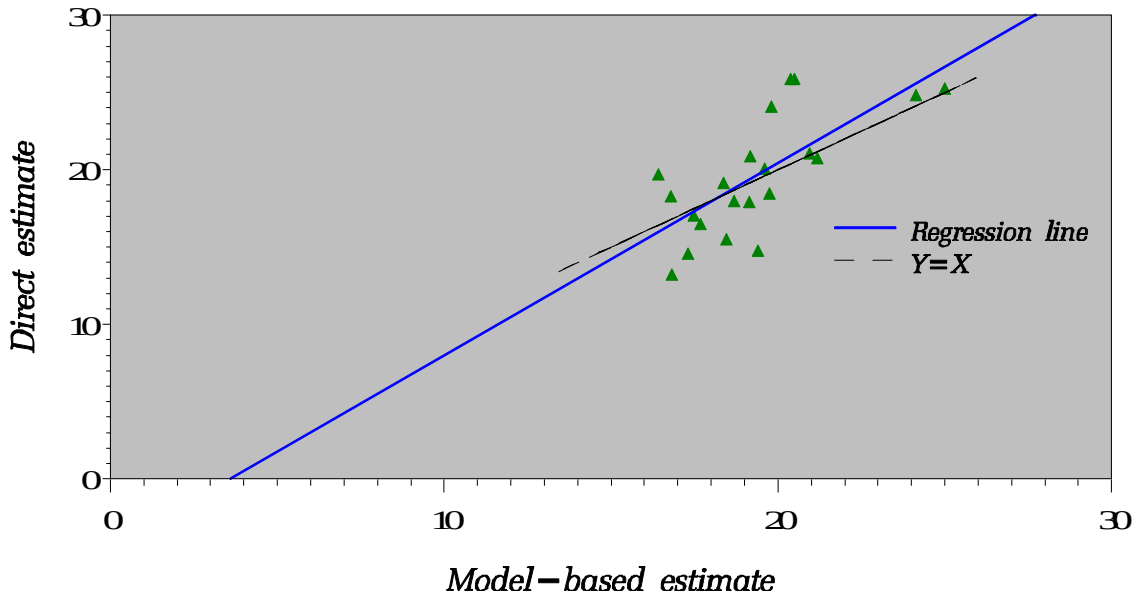
Appreciation du biais du modele

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



Appendix 6: Direct regional estimations versus 'small area' estimations

Year 2009

Region	At-risk-of-poverty indicator	At-risk-of-poverty ($\hat{\theta}_1$)	AROPE indicator	AROPE indicator ($\hat{\theta}_5$)
	Direct method	'Small area' method	Direct method	Small area' method
11	11.19	12.05	16.88	17.31
21	16.16	13.82	20.45	19.76
22	17.50	13.68	24.49	19.58
23	12.10	12.14	17.18	17.95
24	9.79	10.98	16.59	16.65
25	8.81	12.57	13.53	18.46
26	8.83	11.75	16.12	17.63
31	16.03	17.40	22.14	23.37
41	16.59	12.92	21.31	18.70
42	10.30	10.55	16.28	15.87
43	11.53	11.87	18.23	17.59
52	9.99	10.69	14.19	16.00
53	10.27	10.69	14.69	16.27
54	9.21	12.92	15.82	18.51
72	12.27	12.47	18.27	18.11
73	17.07	13.58	22.15	19.07
74	17.41	13.90	24.93	19.74
82	8.31	11.20	13.67	16.54
83	14.94	13.55	18.24	19.36
91	18.59	17.69	25.75	23.68
93	18.25	15.19	23.19	20.77
94	24.18	19.47	42.64	26.51

Year 2010

Region	At-risk-of-poverty indicator	At-risk-of-poverty ($\hat{\theta}_1$)	AROPE indicator	AROPE indicator ($\hat{\theta}_5$)
	Direct method	'Small area' method	Direct method	'Small area' method
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

HOW TO OBTAIN EU PUBLICATIONS

Free publications:

- one copy:
  via EU Bookshop (http://bookshop.europa.eu);
- more than one copy or posters/maps:
  from the European Union's representations (http://ec.europa.eu/represent_en.htm);
  from the delegations in non-EU countries (http://eeas.europa.eu/delegations/index_en.htm);
  by contacting the Europe Direct service (http://europa.eu/europedirect/index_en.htm) or calling 00 800 6 7 8 9 10 11 (freephone number from anywhere in the EU) (*).

  (*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Priced publications:

- via EU Bookshop (http://bookshop.europa.eu).

KS-TC-15-001-EN-N

Publications Office

doi:10.2785/879307