UNIVERSITÀ DEGLI STUDI DI FIRENZE
DIPARTIMENTO DI STATISTICA "G. PARENTI"

Dottorato di Ricerca in Statistica Applicata
XXII ciclo – SECS-S/01

Multidimensional poverty measures: issues in small area estimation

Caterina Ferretti

Tutor: Prof. Achille Lemmi

Co-Tutor: Prof. Gianni Betti

Coordinatore: Prof. Fabio Corradi

Contents

Introduction

The Lisbon European Council in March 2000, agreed to put in place a European strategy aiming at making a decisive impact on the eradication of poverty in the European Union Countries by the year 2010 (the year of struggle against poverty and social exclusion) declaring that the EU should become by 2010 "the most competitive and dynamic knowledge-based economy in the world capable of sustainable economic growth with more and better jobs and greater social cohesion". For this purpose, the adoption of common indicators to monitor living conditions and to guide the implementation of policies of the EU Member States are requested not only at national level but also at regional and at lower geographical levels. Construction of indicators depend on what we want measure, nevertheless some rules have to been respected. Atkinson *et al.* (2002) summarize principles of indicator construction as follows:

i. Identify essence of problem and have clear normative interpretation.
ii. Be robust and statistically validated.
iii. Responsive to effective policy intervention but not subject to manipulation.
iv. Measurable in a comparable way across Member States.
v. Timely.
vi. Measurement not impose too heavy a burden.
vii. Balanced across different dimensions.
viii. Mutually consistent and proportionate weight.
ix. As transparent and accessible as possible to citizens.

In recent years, the interest in poverty, not only as monetary phenomenon but in its multidimensional nature, took a large importance, in the international scientific community, but also in many official statistical agencies and in international institutions. Poverty is a complex phenomenon that cannot be reduced solely to monetary dimension but it has to be also explained by other variables whose impact on poverty is not captured by income.

In 2008 the French President Nicolas Sarkozy established a Commission on the "Measurement of Economic Performance and Social Progress", led by Professor J. Stiglitz and participated by four other Nobel Laureates and well-known expert from over the world. The principal target of the Commission is to identify main causes of the growing divergence between current measures of economic performance and people's perceptions about the quality of their life. It aims to provide meaningful measures of social well-being in the short and long time and to develop research work to overcome limitations of GDP as an indicator of economic performance and social progress. The first distinction is between an assessment of *current well-being*, due to both economic resources and non-economic aspects of people's life and an assessment of *sustainability* that depend on whether stocks of capital that matter for our lives are passed on to future generations. To reach its targets, the Commission followed three main direction of study:

1. Classical GDP issues.
2. Quality of life.
3. Sustainable development and environment.

The main messages of the report are that “time has come to adopt our system of measurement of economic activity to better reflect the structural changes which have characterized the evolution of modern economies” and that “time is ripe for our measurement system to shift emphasis from measuring economic production to measuring people’s well being that should be put in a context of sustainability”.

First of all, it needs to look at income and consumption rather than production. GDP is the most widely-used measure of economic activity, thanks to internationals standards for its calculation, but it has often been treated as a measure of economic well-being, even if it is mainly a measure of market production. On the other hand, material living standards are more closely associated with measures of real income and consumption. These can only be gauged in conjunction with information on wealth and better considered at household level.

In addition to economic production and living standards, the broader concept of quality of life, that overcomes the material side of life, can be evaluated. Measures of quality of life don’t replace conventional economic indicators but provide an opportunity to enrich the view of the condition of a community. Well-being has a multidimensional nature and eight dimensions can be taken in account in defining it:

i. Material living standard (income, consumption and wealth).
ii. Health.
iii. Education.
iv. Personal activities including work.
v. Political voice and governance.
vi. Social connections and relationships.
vii. Environment
viii. Insecurity, of an economic as well as a physical nature.

Both objective and subjective measures of well-being, as people’s self-report, perception happiness and satisfaction, provide key information about quality of life. Of course, people’s assessment about their conditions has no obvious objective counterpart, but a rich literature supports that they help to understand people’s behaviour. For this reason, questions about various aspects of subjective well-being have to be included in standard survey and the links between the most salient features of quality of life across everyone have to be assessed. Finally, the plurality of indicators of quality of life have to been aggregated in a single scalar measure.

The third concern of the Commission is measuring and assessing sustainability. This poses the “challenge of determining if at least the current level of well-being can be maintained for future generations”. This issue is more complex than the other two because it involve the future and many assumptions and normative choices. Moreover, there are interactions between the socio-economic and environmental models followed by the different nations. A well-identified set of indicators is required for sustainability assessment and the components of this set are interpretable as variations of some underlying stocks. These indicators describe the sign of the change in the quantities of the different factors that matter for future well-being. The approach proposed by the Commission is focusing the monetary aggregation on items for which reasonable valuation technique exist, such as physical capital, human capital and certain natural resources. On the other hand, the environmental aspects of sustainability, because of the difficulty to capture them in monetary terms, deserve a separate follow-up based on a set of physical indicators.

This thesis is part of a European Project inside the Seventh Framework Programme: *Small Area Methods for Poverty and Living Condition Estimates* (S.A.M.P.L.E.), born by the purpose to create a European strategy aimed at eradicating poverty in European Country as established by the Lisbon European Council. The aim of this project is to identify and develop new indicators and models that will help the understanding of inequality and poverty with special attention to social exclusion and deprivation. Furthermore, to develop models and implement procedures for estimating these indicators and their corresponding accuracy measures at the level of small area (NUTS3 and NUTS4 level). This project, coordinated by University of Pisa, involves the participation of different European University and Local Administrations. It is structured in six parts, divided in a group of tasks, corresponding to six main areas of research or development. In particular, the activity of CRIDIRE (Siena University) unit, in which this work takes one's place, concerns the *Work package 1*, i.e. to analysis of the mechanisms and the determinants of poverty and inequality and the consequent translation into effective indicators.

According to these remarks, the main objectives of this work are two. First of all, introducing new multidimensional measures of poverty and then, applying the proposed monetary and non-monetary indicators at local level. In our approach, deprivation is treated as a fuzzy state, i.e. as a vague predicate that manifests itself in different shades and degrees. This definition overcomes any limitation of the unidimensional traditional approach characterized by a simple dichotomization of the population into poor and non poor defined in relation to some chosen poverty line that represents a percentage of the media or the median of the equivalent income distribution. The introduction of fuzzy measures implies to additional aspect respect to traditional approach:

i. The choice of membership functions, i.e. quantitative specification of individuals' or households' degrees of poverty and deprivation;
ii. The choice of rules for the manipulation of the resulting fuzzy sets, as complements, intersections, union and aggregation.

Concerning the first point, we introduce a new approach to the analysis of poverty and deprivation, the so called *Integrated Fuzzy and Relative* (IFR) approach. This measure, proposed by Betti *et al*. (2006), combines the TFR approach of Cheli and Lemmi (1995) and the approach of Betti and Verma (1999). In order to describe the multidimensional nature of deprivation, we choose a set of non-monetary indicators and identify different dimensions of deprivation they represent. We include a majority of the so-called 'objective' indicators on non-monetary deprivation, such as the possession of material goods and facilities and physical conditions of life, at the expense of what may be called 'subjective' indicators such as self-assessment or satisfaction of general life conditions. This choice is due to the lacked or mistaken use of subjective variables at European survey level. Moreover, subjective indicators tend to be more culture-specific and hence less comparable across countries and regions. Seven dimensions are identified using factor analysis, as described in Whelan *et al.* (2001). Then, for each dimension, quantitative measures of deprivation are constructed. Weights for the aggregation of individual items included in the dimension are determined within each dimension separately and the set of weights are taken to be item-specific. They take into account dispersion of each item and correlation of this with each other. Moreover, as membership function indicating the individuals' degrees of deprivation in different dimensions we choose exactly the same as that used in the case of income poverty. In order to define a methodology of fuzzy set aggregation over dimensions of deprivation, we choose a 'composite' fuzzy set operator which takes into account whether the sets being aggregated are of a 'similar' or a 'dissimilar' type. Using this composite set operator we can define 'latent' and 'manifest' deprivation as, respectively, the union and intersection of deprivation in multiple dimensions. Empirical results of fuzzy measures of monetary poverty and non-monetary deprivation and their combination are obtained using data of the EU Member States from EU-SILC survey.

Concerning the second objective of this work, we propose two methods: the first one deals with pooling of data or estimates, whereas the other one involves small area estimation techniques. Pooling of data means statistical analysis using multiple data sources relating to multiple populations. It encompasses averaging, comparisons and common interpretations of the information. Different scenarios and issues also arise depending on whether the data sources and populations involved are same/similar or different. Examples of each scenarios are provided, nevertheless we focus primarily on cumulation over space and time from repeated multicountry surveys, taking illustrations from European social surveys. Simple model are developed to illustrate the effect on variance of pooling over correlated samples, such as over waves in a rotational panel design.

On the other hand, we propose a methodology for obtaining empirical best predictors of general, non-linear, domain parameters using unit level linear regression models. It is based on a modified version of Empirical Best (EB) prediction proposed by Molina and Rao (2009), and it can resolve computational problems for big populations or more complex poverty measures, as fuzzy indicators. Head count ratio (HCR), fuzzy monetary indicator (FM) and fuzzy supplementary index (FS) are used as non-linear domain parameters. The method is applied to the estimation of these poverty measures in Tuscany provinces.

This work is organized in five chapters. Chapter 1, after a brief review of traditional approach to poverty, introduces several multidimensional approaches proposed in literature focusing on the proposed fuzzy method. Chapter 2 presents some numerical results at national level for illustration of the methodology described in the previous chapter. Chapter 3 deals with the problem of pooled estimates of indicators, describing different scenarios with some examples. Chapter 4 introduces the basic theory of small area estimation, in particular of mixed effects models and M-quantile models. Finally, Chapter 5 describe the proposed small area estimation procedure of poverty indicators and gives some numerical results for both traditional and fuzzy measures at local level.

Chapter 1

Multidimensional poverty indicators in fuzzy and non fuzzy approach

1.1. Traditional Poverty Approach

The traditional poverty approach is characterized by a simple dichotomization of the population into poor and non poor defined in relation to some chosen poverty line that represents a percentage (generally 50%, 60% or 70%) of the media or the median of the equivalent income[1] distribution.

This approach is unidimensional, that is, it refers to only one proxy of poverty, namely low income or consumption expenditure.

The traditional poverty method takes place in two different and successive stages: the first aims to identify who is poor and who is not according to whether a person's income is below a critical threshold, the poverty line; the second stage consists of summarising the amount of poverty in aggregate indices that are defined in relation to the income of the poor and the poverty line.

We can distinguish between poverty measures and inequality measures as discussed below.

1.1.1. Poverty measures

Poverty measures are used first and foremost for monitoring social and economic conditions and for providing benchmarks of progress or failure. They are indicators by which policy results are judged and by which the impact of events can be weighed, then they need to be trusted and require rigorous underpinning. They depend on the average level of consumption or income in a country and the distribution of income or consumption, then they focus on the situation of those individuals or households at the bottom of the distribution.

[1] The equivalent income of a household is obtained by dividing its total disposable income by the household's equivalised size computed by using an equivalent scale which takes into account the actual size and composition of the household.

The measures will function well as long as everyone agrees that when poverty numbers rise, conditions have indeed worsened and conversely, when poverty measures fall, that progress has been made.
Poverty measures must satisfy a given set of axioms or must have certain characteristics:

- *Scale invariance*: poverty measures should be unchanged if, for example, a population doubles in size while everything else is maintained in the same proportions;
- *Focus axiom*: changes among better-off people below the poverty line do not affect measured poverty;
- *Monotonicity axiom*: holding all else constant, when a poor person's consumption or income falls, poverty measures must rise or at least should not fall;
- *Transfer axiom (Pigou-Dalton principle)*: holding all else constant, taking money from a poor person and giving it to a less poor person must increase the poverty measure and conversely, poverty falls when the very poor gain through a transfer from those less poor;
- *Transfer – Sensitivity axiom*: the reduction of poverty in the case in which a very poor person is made better off in relation to her neighbour should be greater than the reduction in the case in which the recipient is less poor;
- *Decomposability axiom*: poverty measures should be decomposed by sub-population.

The most widely used measure is the *headcount index*, which simply measures the proportion of the population that is counted as poor. Formally:

$$H = \frac{q}{n} \tag{1.1.1}$$

where n is the total population and q is the total number of poor.

The headcount index is simple to construct and easy to understand, but it presents some weaknesses also. For example, it violates the transfer principle of Pigou-Dalton that states that transfers from a richer to a poorer person should improve the measure of welfare. The headcount index does not indicate how poor the poor are, and hence, does not change if people below the poverty line become poorer. Moreover, it calculates the percentage of individuals and not households, as the poverty estimates should be calculated, making a not always true assumption that all household members enjoy the same level of well-being.

A moderately popular measures of poverty is the *poverty gap index*, which adds up the extent to which individuals fall below the poverty line and expresses it as a percentage of the poverty line. Formally:

$$I = \frac{1}{n}\sum_{i=1}^{q}\left(\frac{z - y_i}{z}\right) \tag{1.1.2}$$

where z is the poverty line and y_i the actual expenditure/income for poor people.

The poverty gap is defined as the difference between z and y_i for poor people and zero for everyone else.

Equation (1.1.2) is the mean proportionate poverty gap in the population and shows how much would have to be transferred to the poor bring their incomes or expenditures up to the poverty line. This measure has the virtue that it does not imply that there is a discontinuity at the poverty line but its serious shortcoming is that it may not convincingly capture differences in the severity of poverty among the poor.

The poverty gap index is, then, the average over all people, of the gaps between poor people's standard of living and the poverty line expresses as a ratio to the poverty line. The aggregate poverty gaps shows the cost of eliminating poverty by making perfectly targeted transfer to the poor, in the absence of transactions costs and disincentive affects.

Another poverty measure is the *squared poverty gap index* or *severity poverty index* used to solve the problem of inequality among the poor but not easily interpretable. This is simply a weighted sum of poverty gaps where the weights are the proportionate poverty gaps themselves giving more weight on observations that fall well below the poverty line. Formally:

$$P_2 = \frac{1}{n}\sum_{i=1}^{q}\left(\frac{z-y_i}{z}\right)^2 \qquad (1.1.3)$$

It belongs to a family of measures proposed by Foster, Greer and Thorbecke (1984), which may be written as:

$$FGT = \frac{1}{n}\sum_{i=1}^{q}\left(\frac{z-y_i}{z}\right)^{\alpha} \qquad (1.1.4)$$

where α is a measure of the sensitivity of the index to poverty. For $\alpha = 0$, *FGT*(0) coincides with the headcount index, when $\alpha = 1$ *FGT*(1) is the poverty gap index and for $\alpha = 2$, *FGT*(2) is the poverty severity index. For $\alpha > 0$ this measure is strictly decreasing in the living standard of the poor. Furthermore, for $\alpha > 1$ it is strictly convex in income, that is, the increase in measured poverty due to a fall in one's standard of living will be deemed grater the poorer one is.

FGT class of poverty can be disaggregated for population sub-groups and the contribution of each sub-group to national poverty can be calculated.

Sen (1976) proposed an index that sought to combine the effects of the number of poor, the depth of their poverty and the distribution of poverty within the group. Formally:

$$S = \frac{2}{(q+1)nz}\sum_{i=1}^{q}(z-y_i)(q+1-i) \qquad (1.1.5)$$

This measure can also be written as the average of the headcount and poverty gap indices weighted by the Gini coefficient of the poor (G_P) that ranges from 0 (perfect equality) to 1 (perfect inequality), that is:

$$S = H\left(I + (1-I)G_p\right) \qquad (1.1.6)$$

The Sen index has the virtue of taking into account the income distribution among poor but it lacks intuitive appeal and cannot be decomposed satisfactorily into this constituent components. For these shortcomings it is rarely used in practice.

1.1.2. Inequality measures

Inequality measures are most general than poverty ones because they are defined over the entire population, not only for the population below a certain poverty line. They are concerned with the distribution and a virtue of these is the mean independence, that is, most inequality measures do not depend on the mean of the distribution.

Inequality indicators can be harder to develop than consumption/income poverty indicators because they essentially summarize one dimension of a two-dimensional variable, but they can be calculated for any distribution not just for monetary variables.

The commonest way to measures inequality is by dividing the population into fifths (quintiles) from poorest to richest and reporting the levels or proportions of income or expenditure that accrue to each level.

The Gini (1912) coefficient is the most widely used measure of inequality. It is based on the Lorenz (1905) curve, a cumulative frequency curve that compares the distribution of a specific variable with the uniform distribution that represent equality. The Gini coefficient is constructed by plotting the cumulative percentage of households, from poor to rich, on the horizontal axis and the cumulative percentage of expenditure or income on the vertical axis. It range between 0 (perfect equality) and 1 (complete inequality). Formally:

$$Gini = G = \frac{2}{n^2 \bar{y}} \sum_{i=1}^{n} \left(y_i - \bar{y} \right) \qquad (1.1.7)$$

where the y_i are ordered from the lowest to the highest.

The Gini coefficient satisfies mean independence, population size independence, symmetry and Pigou-Dalton transfer sensitivity axioms, but decomposability and statistical testability properties don't hold for this index.

Otherwise, the Theil (1967) indices and the mean log deviation measure, that belong to family of generalized entropy inequality measures, satisfy all six criteria cited above. The general formula is given by:

$$GE(\alpha) = \frac{1}{\alpha^2 - \alpha} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i}{\bar{y}} \right)^{\alpha} - 1 \right\} \qquad (1.1.8)$$

where $\bar{y}$ is the mean expenditure/income. The values of GE measures vary between 0, equal distribution, and ∞, high inequality. The parameter α in the GE class represents the weight given to distances between incomes at different parts of the income distribution, and can take any real value. For lower values of α, GE is more sensitive to changes in the lower tail of the distribution, and for higher values GE is more sensitive to changes that affect the upper tail. The commonest values of α used are 0,1 and 2.

GE(0), also known as Theil's L, is called mean log deviation measure because it gives the standard deviation of log(y):

$$GE(0) = \frac{1}{n}\sum_{i=1}^{n} -\log\left(\frac{y_i}{\bar{y}}\right) \tag{1.1.9}$$

GE(1) is Theil's T index, which may be written as:

$$GE(1) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\bar{y}}\right)\log\left(\frac{y_i}{\bar{y}}\right) \tag{1.1.10}$$

Atkinson (1970) proposed another class of inequality measures with theoretical properties similar to those of the extended Gini index. Formally:

$$A = 1 - \left\{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\bar{y}}\right)^{1-\varepsilon}\right\}^{1/(1-\varepsilon)} \tag{1.1.11}$$

Finally, another inequality measure, called L-measure, was proposed by Kakwani (1980). This measure bases on Lorenz curve and formally may be defined as follows:

$$L = \frac{l - \sqrt{2}}{2 - \sqrt{2}} \tag{1.1.12}$$

where l is the length of Lorenz curve.

The values of the Lorenz curve length vary from $\sqrt{2}$, equal distribution, to 2, the highest inequality. The L-measure takes values in [0,1].

After some transformations the L-measure may be written as:

$$L = \frac{1}{\left(2 - \sqrt{2}\right)}\left[\frac{1}{\mu}\int_0^{\infty}\sqrt{\mu^2 + y^2}\, f(y)dy - \sqrt{2}\right] \tag{1.1.13}$$

where μ is the mean income.

The L-measure satisfies all axioms which Gini coefficient satisfies and additionally additive decomposability axiom. It is also more sensitive to changes in the lower tail of income distribution, as opposed to Gini coefficient, than to changes in the upper tail.

1.2. Multidimensional Approach

The traditional poverty approach presents two limitations: i) it is unidimensional, i.e. it refers to only one proxy of poverty, namely low income or consumption expenditure; ii) it needs to dichotomise the population into the poor and the non-poor by means of the so called poverty line.

Nowadays there is a widespread agreement about the multidimensional nature of poverty: poverty is a complex phenomenon that cannot be reduced solely to monetary dimension but it has to be also explained by other variables whose impact on poverty is not captured by income. This leads to the need for a multidimensional approach that consists in extending the analysis to a variety of non-monetary indicators of living conditions and at the same time adopts mathematical tools that can represent the complexity of the phenomenon.

Eight different approaches are described in the following sections: social welfare approach, counting approach, Sen's capability approach, distance function approach, information theory approach, axiomatic approach, supervaluationist approach and fuzzy set approach.

1.2.1. Social Welfare Approach

The social welfare approach, relating to income inequality measures, assumes a social evaluation function for a vector of incomes from which an inequality index is derived. This function ranks different distributions of attributes among a set of individuals.

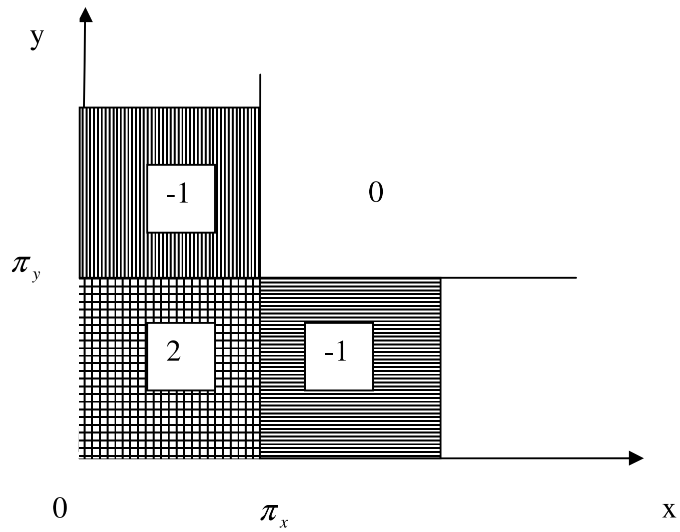
Dalton (1920) was the first to argue that economist were interested in the effects of inequality on economic welfare and that inequality in a distribution should be measured by the loss in welfare that it causes.

Social welfare is measured by a function *S* which represents society's notion of how fair or desirable a particular distribution is. *S* may be a function of individual welfare, the part of individual welfare due to income alone or the incomes that individuals receive and it increases as income increases. One the most common forms of the social welfare function is the additive one, in which the social welfare is the sum of individuals welfares, assuming that the welfare of an individual is independent of the welfare of other individuals.

This approach is based on dominance conditions that allow us to state that "multidimensional deprivation in country A is lower than in country B" for all deprivation measures satisfying certain general properties.

Suppose that x and y are the arguments in a social welfare function representing the position of an individual. In the case of two dimensions, deprivation is represented in the graph 1.2.1. where π_x and π_y are the deprivation thresholds respectively in dimension x and y. If $F(x, y)$ denotes the cumulative distribution, $f(x, y)$ is the density function, $F(x)$ and $F(y)$ are the marginal distributions and $F(\pi_x)$ and $F(\pi_y)$ are respectively the proportions of deprived people on the dimensions x and y, the union is given by $F(\pi_x) + F(\pi_y) - F(\pi_x, \pi_y)$ where $F(\pi_x, \pi_y)$ is the proportion of individuals deprived on both dimensions.

Figure 1.2.1. Deprivation in two dimensions



Let D be a class of deprivation measures formed by integrating over the distribution a function $p(x, y)$, where this is zero when x and y are both above the poverty thresholds:

$$D = \int_0^{\pi_x}\int_0^{\pi_y} p(x, y) f(x, y) dy dx \tag{1.2.1}$$

Following the social welfare approach this quantity has to be minimised. As show by Bourguignon and Chakravarty (2003), a deprivation measures is increased or remain the same, as a result of a correlation increasing perturbation if the cross-derivate of p with respect to x and y is positive, that is the attributes are substitutes. Conversely, when the derivate is negative, they are complements.

The first-degree dominance conditions allow us to rank two distributions: for poverty measures that are substitutes $F(x, y)$ must be lower in country A than in country B given x and y, conversely, for poverty measures that are complements $[F(x) + F(y) - F(x, y)]$ must be lower in country A than in country B given x and y.

Bourguignon and Chakravarty (2003), for example, defined a deprivation index as:

$$p(x, y) = \left[g_x^{\beta} + b g_y^{\beta}\right]^{\alpha/\beta} \tag{1.2.2}$$

where $g_x = \max[0, (1 - x/\pi_x)]$ and $g_y = \max[0, (1 - y/\pi_y)]$ are the relative shortfalls. In the expression (1.2.2) the parameter α is a measure of concavity of the function – $p(x, y)$, β governs the shape of the contours in (x, y) space and b represents the weight of single attributes.

The cross-derivate of p is positive where $\alpha > \beta$, then x and y are substitutes, whereas for $\beta > 1$ they are complements.

1.2.2. Counting Approach

The counting approach consists on counting the number of dimensions in which people suffer deprivation, not distinguishing the extent of the shortfalls. Given a set of key dimensions and a poverty line, the number of dimensions in which a person is poor is counted and becomes the poverty score. Formally:

$$\rho(x_i; z) = 1 \text{ if } \exists \ \ j \in \{1,2,...,m\} : x_{ij} < z_j \qquad (1.2.3)$$

$$\rho(x_i; z) = 0 \text{ otherwise} \qquad (1.2.4)$$

where i = 1, 2, …, n are individuals, j = 1, 2, …, m are attributes and z represents the poverty threshold for each attribute.

The number of poor in the dimensional framework is given by:

$$n_p(X) = \sum_{i=1}^{n} \rho(x_i; z) \qquad (1.2.5)$$

Alternatively, one can count a person poor if she is poor in any dimension or only if she is poor in all dimensions. Atkinson (2003) showed as this approach can be related to the welfare social approach described in the previous section.

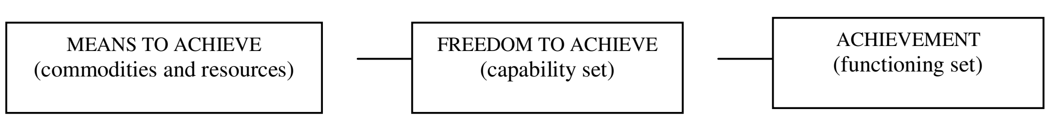
1.2.3. Sen's Capability Approach

Sen's capability approach, on the contrary to other multidimensional approaches of poverty, is not simply a way to enlarge the evaluative well-being to variables other than income, but it gives a different meaning of well-being.

The main characteristic of this theory is the interpretation of well-being: it is not only associated to affluence but to each one's abilities. Moreover, Sen emphasises the importance of the freedom to choose. Himself affirms: "Acting freely and being able to choose are, in this view, directly conducive to well-being" (Sen, 1992).

This approach characterizes individual well-being in terms of what a person is actually able to do or to be. Its main components are the *commodities or resources*, the *functionings* and the *capabilities*.

Figure 1.2.2. A diagrammatic representation of the capability approach



The commodities are all goods and services, not just merchandise. They make possible the functionings that represent achievements of people and reflects life-style; “the various things a person may value doing or being” (Sen, 1992). Capabilities are various combinations of functionings that the person can achieve. “Capabilities is, thus, a set of vectors of functioning, reflecting the person’s freedom to lead one type of life or another (…) to choose from possible livings” (Sen, 1992).

Capability and functionings are influenced by the intrinsic characteristics of the people, like age and gender, as well as by environmental circumstances.

Formally, (Sen, 1985; Kuklys, 2005), the individual capability set Q_i, i.e. the space of potential functionings, can be expressed as:

$$Q_i(X_i) = \{\mathbf{b}_i \mid \mathbf{b}_i = f_i(x_i) \mid (\mathbf{h}_i, \mathbf{e}_i)\} \quad (1.2.6)$$

for some $f_i(\cdot) \in F_i$ and some $x_i \in X_i$. $\mathbf{b}$ is a vector of functionings, f_i is a conversion function, and $\mathbf{h}_i$ and $\mathbf{e}_i$ are respectively vectors of personal factors and environmental factors which influence the rate of conversion of individual resources (x_i) to a given functioning (b_i).

Capability approach, as every multidimensional method of poverty analysis, is characterized by threes different stages: the *description* of human poverty and individual well-being in all its multifaceted and gradual aspects; the *aggregation* of indicators and dimensions into an overall measure of individual well-being; the *inference* to derive logical conclusions from premises that are know or from factual knowledge or evidence. These phases can be resolved using fuzzy set theory and fuzzy logic that have been proved to be powerful tools.

1.2.4. Distance Function Approach

The distance function approach was first applied to the analysis of households behaviour by Lovell *et al.* (1994).

The input distance function $D_{in}(x, y)$ involves the scaling of the input vector and is defined as:

$$D_{in}(x, y) = Max\{\rho : (x/\rho) \in L(y)\} \quad (1.2.7)$$

where

$$L(y) = \{x : x \text{ can produce } y\} \quad (1.2.8)$$

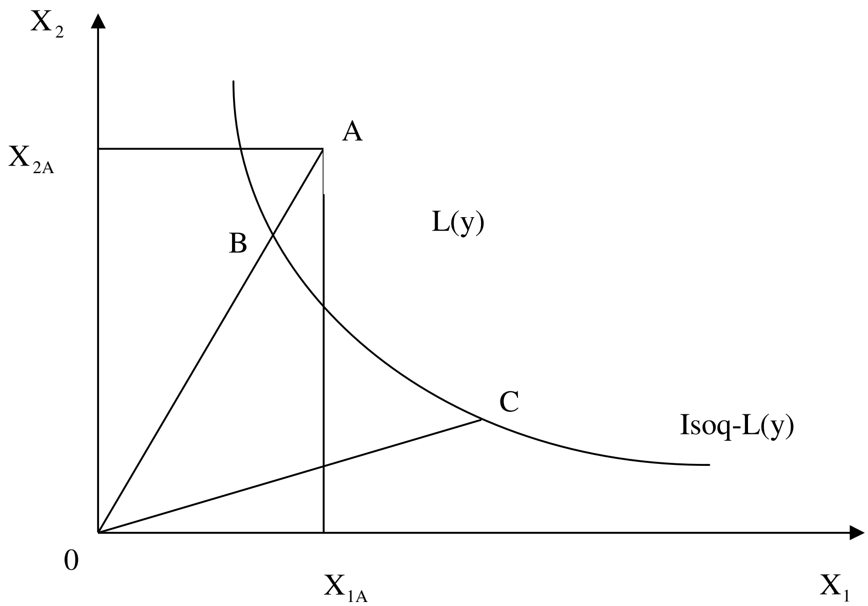
is the input set of all input vectors x which can produce the output vector y.
It holds (Coelli *et al*., 1998) that:

i. The input distance function is increasing in x and decreasing in y;
ii. It is linearly homogeneous in x;
iii. If x belongs to $L(y)$ then $D_{in}(x, y) \geq 1$;
iv. $D_{in}(x, y) = 1$ if x belong to the frontier of the input set (isoquant of y).

Graph 1.2.2 shows the concept of distance function. Here *q* and *q'* are respectively the input vectors corresponding to OA and OB. ρ is equal to the ratio OB/OA. p_0 is the vector of the prices of the inputs. Nothing guaranties that the input contraction defined by the distance function ρ will yield the cheapest cost, at input prices p_0, of producing the output level y_0 defined by the isoquant BC. There exists however at least one vector price *p* for which this distance function $\rho = OB / OA$ will yield the cheapest cost of producing this output level y_0. Then, there is a link between the cost function that seeks out the optimal input quantities given y_0 and p_0 and the distance function that finds the prices that will lead the consumer to reach the output level y_0 by acquiring a vector of quantities proportional to *q*.

The concept of distance function can be applied to measures poverty and life conditions.

Figure 1.2.3. The concept of distance function



1.2.5. Information Theory Approach

The informational theory approach, originally developed in the field of communication, was first utilized in economics by Theil (1967). It is based on the concept of the logarithm of a probability.

Let *E* be an experience whose result is x_i with *i* = 1 to *n*. Let $p_i = \Pr(x = x_i)$, $0 \le p_i \le 1$, be the probability that the result of the experience will be x_i. The information that a given event x_i occurred is not very important if the a priori probability that such an event would occur was high. Conversely, it becomes significant if the a priori probability that an event x_i will occur is very low, knowing that this event did indeed occur.

We can define this information as a function of the probability a priori *p* that a result will occur. One the most common forms is:

$$h(p) = \log(1/p) = -\log(p) \quad (1.2.9)$$

From this, we can derive the expected information, called also entropy:

$$H(p) = \sum_{i=1}^{n} p_i h(p_i) \quad (1.2.10)$$

Combining (1.2.9) and (1.2.10) we obtained the Shannon entropy that can be interpreted as the uncertainty, the disorder or the volatility associated with a given distribution:

$$H(p) = -\sum_{i=1}^{n} p_i \log(p_i) \quad (1.2.11)$$

Shannon entropy is minimal and equal to 0 when a given result x_i is known to occur with certainty and then the information is not important. Conversely, it is maximal when all events have the same probability ($p_i = 1/n$) and we have no idea a priori as to which event will occur.

Maasoumi (1986) applied the information theory to measures of inequality proceeding in two steps: 1) definition of a procedure to aggregate the various indicators of welfare; 2) selection of an inequality index to estimate the degree of multidimensional inequality.

Let x_{ij} be the value taken by indicator *j* for individual *i*, with *i* = 1 to *n* and *j* = 1 to *m*.

Maasoumi proposed to replace the *m* pieces of information on the value of the different indicators for the various individuals by a composite index x_c which will be a vector of *n* components, one for each individual. Then, the vector $x_{i1},...,x_{im}$ corresponding to individual *i* will be replace by the scalar x_{ci} that represents the utility that individual *i* derives from the various indicators or an estimate of the welfare of such a individual. As composite indicator x_c Maasoumi chose a weighted average of the different indicators.

Miceli (1997) proposed to use the distribution of the composite index x_c suggested by Maasoumi to derive multidimensional poverty measures, applying to each indicators a weight proportional to its mean (the more diffused the durable good is the higher its weight is) or an equal weight (1/*m*) to all the indicators. To identify the poor Miceli adopted a relative approach defining the poverty line as some percentage of the median value of the composite indicator x_c.

1.2.6. Axiomatic Approach

The axiomatic approach has been developed by Tsui (1995, 2002) and Chakravarty *et al.* (1998). It is based on the idea that a multidimensional index of poverty is an aggregation of shortfalls of all the individuals where the shortfall with respect to a given need reflects the fact that the individual does not have even the minimum level of the basic need.

Already, Sen (1976) suggested two basic postulates for an income poverty index: i) the monotonicity axiom, i.e. poverty should increase if the income of a poor person decreases; ii) the transfer axiom, i.e. poverty should increase if there is a transfer of income from a poor person to anyone who is richer. Later on, several other axioms have been suggested in literature.

Let $z = (z_1,...,z_k)$ be the *k*-vector of the minimum levels of the *k* basic needs and $x_i = (x_{i1},...,x_{ik})$ the vector of the *k* basic needs of the *i*-th person. Let *X* be the matrix of the quantities x_{ij} which denote the amount of the *j*-th attribute accruing to individual *i*.

A multidimensional poverty measure has to satisfy several properties (Chakravarty *et al*., 1998):

i. *Symmetry*: This property assumes that the multidimensional poverty index depends only on the various attributes *j* that the individuals have and not on their identity.
ii. *Focus*: If for any individual *i* an attribute *j* is such that $x_{ij} > z_j$, *P*(*X*; *z*) does not change if there is an increase in x_{ij}.
iii. *Monotonicity*: If for any individual *i* an attribute *j* is such that $x_{ij} \leq z_j$, *P*(*X*; *z*) does not increase if there is an increase in x_{ij}.
iv. *Principle of Population*: An *m*-fold replication of *X* will not affect the value of the poverty index.
v. *Continuity*: An index of multidimensional poverty *M*(*X*) should be a continuous function, that is, it should be only marginally affected by small variations in x_{ij}.
vi. *Non-Poverty Growth*: If the matrix *Y* is obtained by adding a rich person to the population defined by *X*, then *P*(*Y*; *z*) ≤ *P*(*X*; *z*).
vii. *Non-decreasingness in Subsistence Levels of Basic Needs*: If z_j increases for any *j*, *P*(*X*; *z*) does not decrease.
viii. *Scale Invariance*: This implies that the ranking of any two matrices of attributes is preserved if the attributes are rescaled according to their respective ratio scales.
ix. *Normalization*: *P*(*X*; *z*) = 1 whenever $x_{ij} = 0$ for all *i* and *j*.
x. *Subgroup Decomposability*: Assume n_i is the population size of subgroup *i* (*i* = 1 to *m*) with $n = \sum_{i=1}^{m} n_i$ representing the total size of the population. Then the poverty index for the whole population (where the data on each subpopulation is represented by a matrix X_i) may be expressed as:

$$P(X_1,...,X_m) = \sum_{i=1}^{m} \frac{n_i}{n} P(X_i; z) \qquad (1.2.12)$$

xi. *Factor Decomposability*:

$$P(X;z)=\sum_{j=1}^{k}a_j P(x_j;z_j) \qquad (1.2.13)$$

where x_j; is the *j*-th column of *X*, a_j is the weight attached to attribute *j* such that $\sum_{j=1}^{k}a_j=1$.

xii. *Transfer Axiom*: Let X_p be the submatrix of *X* corresponding to the poor. If *Y* is derived from *X* by multiplying X_p by a bistochastic matrix (not a permutation matrix), then $P(Y;z)\leq P(X;z)$ given that the bundles of attributes of the rich remain unaltered.

xiii. *Nondecreasing Poverty under Correlation Increasing Arrangement*: This property refers to switches of some attributes between individuals that increase the correlation of the attributes.

Chakravarty *et al*. (1998) derive the following two propositions.
Proposition 1: The only non constant focused poverty index that satisfies the properties of subgroup decomposability, factor decomposability, scale invariance, monotonicity, transfer axiom, continuity and normalization is defined as:

$$P(X;z)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}a_j f(x_{ij}/z_j) \qquad (1.2.14)$$

where *f* is continuous, non-increasing and convex with $f(0) = 1$ and $f(t) = c$ for all $t \geq 1$ and $c<1$ c is a constant. The parameters a_j are positive and constant with $\sum_{j=1}^{k}a_j=1$.

Proposition 2: The poverty measure $P(X;z)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}a_j g(x_{ij}/Z_j)$ satisfies the properties of Symmetry, Population Replication, Non-Poverty Growth and Non-Decreasingness in Subsistence Levels of Basic Needs. If g $(g(t) = (f(t) - c)/(1 - c))$ is twice differentiable on $(0,1)$ P, the poverty index, satisfies also the property of Nondecreasing Poverty under Correlation Increasing Arrangement.

The following multidimensional poverty index may be considered

$$P(X;z)=\frac{1}{n}\sum_{j=1}^{k}\sum_{i\in S_j}a_j\left[1-(x_{ij}/z_j)^e\right] \qquad (1.2.15)$$

where s_j is the set of poor people with respect to attribute *j*.

1.2.7. Supervaluationist approach

This approach implies the concept of vagueness in measuring poverty because "poor" is a vague predicate, i.e. it allows for borderline cases where it is not clear whether the predicate applies or not, there is not sharp borderline between cases where the predicate does and does not apply and it is susceptible to a Sorites paradox.

Supervaluationism proposed by Fine (1975) suppose that the truth of vague predicates depends on how they are made more precise. A vague statement is called "super-true" if it is true on all plausible ways of making it more precise or, equivalently, in all admissible "precisifications". Fine's account involves, for any vague predicate, a number of admissible ways of making statements involving the predicate more precise. Fine "maps" the various ways of making statements involving a vague predicate more precise in terms of a "specification space" that includes "base points" where the statement is initially specified. These points are extended by making the statement more precise until a partial or complete specification.

Qizilbash (2006) follows the Fine's theory in allowing for a set of admissible specification of 'poor' that can be vague. Each admissible specification involves a set of dimensions of poverty and a range of critical levels relating to each dimensions. Any dimension of poverty which appears on all admissible specifications is called a core dimension. In each dimension, someone who falls at or below the lowest admissible critical level is judged to be definitely poor in that dimension. If this person is definitely poor on a core dimension, she is core poor, that is in Fine's terms, it is super true that she is poor. Analogously, someone who falls at or above the highest critical level is definitely not poor in that dimension and if he is definitely not poor on all admissible dimensions is non-poor. Those who are neither poor or non-poor fall at the margins of poverty.

In this framework, fuzzy poverty measures can be interpreted as measures of vulnerability in each dimension where there will be some who falls between the highest and lowest critical levels, and so are neither definitely poor nor definitely not poor in that dimension. These people can be seen as vulnerable in as much as they are poor in terms of some admissible critical level in the relevant dimension, and would be defined as poor if that critical level was used. Fuzzy poverty measures capture how close these individuals come to being definitely poor in the relevant dimension.

In Qizilbash's account the notion of vulnerability which underlies the interpretation of fuzzy poverty measures is different. Fuzzy measures are conceived as measures on the specification space in a particular dimension and they are so relate to the range of precisifications of poor on which someone is judged to be poor in a particular dimension: as that range increases that person is more vulnerable. Then, anyone who is defined as poor on all but one critical level in some dimension might classify as "extremely vulnerable".

Qizilbash adds to vagueness about the critical level at or below which a person classifies as poor (*vertical vagueness*), already used in the literature on fuzzy poverty measures, vagueness about the dimensions of poverty (*horizontal vagueness*).

This framework can be extended to allow for the vagueness of predicates such as "extreme" and "chronic". One of the characteristics of the supervaluationist approach is that if someone is doing sufficiently badly in some core dimension of poverty, he is core poor, without checking the level of achievement on all dimensions of poverty.

1.2.8. Shapley Decomposition

Deutch and Silber (2006) proposed to use Shapley Decomposition to study the most significant determinants of multidimensional poverty.

Let an index *I* be a function of *n* variables and let I_{TOT} be the value of *I* when all the *n* variables are used to compute *I*. Moreover, Let $I_{/k}^{k}(i)$ be the value of the index *I* when *k* variables have been dropped so that there are only (*n*-*k*) explanatory variables and *k* is also the rank of variable *i* among the *n* possible ranks that variable *i* may have in the *n*! sequences corresponding to the *n*! possible ways of ordering *n* numbers. Thus:

- $I_{/k-1}^{k}(i)$ gives the value of the index I when only (*k*-1) variables have been dropped and k is the rank of the variable i;
- $I_{/1}^{1}(i)$ gives the value of the index *I* when this variable is the first one to be dropped;
- $I_{/0}^{1}(i)$ gives the value of the index *I* when the variable *i* has the first rank and no variable have been dropped (all the variable are included in the computation of *I*);
- $I_{/2}^{2}(i)$ corresponds to the (*n*-1)! cases where the variable *i* is the second one to be dropped and two variables as a whole have been dropped;
- $I_{/1}^{2}(i)$ gives the value of the index *I* when only one variable has been dropped and the variable *i* has the second rank;
- $I_{/n-1}^{n}(i)$ corresponds to the (*n*-1)! cases where the variable *i* is dropped last and is the only one to be take into account;
- $I_{/n}^{n}(i)$ gives the value of the index *I* when variable *i* has rank *n* and *n* variable have been dropped (it is 0 by definition).

Deutch and Silber define the contribution $C_j(i)$ of variable *i* to the index *I*, assuming this variable *I* is dropped when it has rank *j*, in the following way:

$$C_j(i) = \frac{1}{n!}\sum_{h=1}^{(n-1)!}\left[I_{/(j-1)}^{j}(i) - I_{/j}^{j}\right]^h \qquad (1.2.16)$$

where *h* refers to one of the (*n*-1)! cases where the variable *i* has rank *j*.

The overall contribution of variable *i* to the index *I* may then be defined as:

$$C(i) = \sum_{k=1}^{n} C_k(i) \qquad (1.2.17)$$

From this, we derive that:

$$I = \sum_{i=1}^{n} C(i) \qquad (1.2.18)$$

1.2.9. Foster-Greer-Thorbecke index of multidimensional poverty

As pointed out in section 1.2, the multidimensional approach to poverty measurement is characterised by the obvious advantages over unidimensional: it can capture various aspects of deprivation that are not restricted to monetary measure. Moreover, applying the fuzzy sets theory allows overcoming most of limitations of the single poverty line splitting the sample into the poor and the non-poor. On the other hand, multidimensional poverty indices cannot hold all axioms passed by some unidimensional formulas, especially Foster-Greer-Thorbecke index holding large variety of properties. Sub-group decomposability is especially desirable when spatial distribution of poverty and deprivation is the object of interest.

Recently Alkire and Foster (2008) have developed a framework for measuring poverty in the multidimensional environment that is analogous to the FGT family of indices. The resulting formula is characterised by the following properties:

i. can be applied prior to any additive aggregation technique that aggregates first across persons,
ii. satisfies certain basic axiomatic properties for uni- and multi-dimensional poverty measures,
iii. can accommodate ordinal as well as cardinal data, although some properties are available only with ordinal data,
iv. can apply equal weights or general weights (assuming that all dimensions are equally important is not necessary therefore).

Moreover this index is intuitively attractive, hence it may be used in evaluations and discussions on the social policy. The index employs two types of cut-offs: first, within each dimension to identify the deprived in that dimension, and second, across dimensions to count the number of dimensions in which the individual is deprived.

Multidimensional FGT index satisfies the following axioms (see section 1.2.6):

xiv. Symmetry,
xv. Poverty and deprivation Focus,
xvi. Monotonicity,
xvii. Principle of Population,
xviii. Non-Poverty Growth,
xix. Non-decreasingness in Subsistence Levels of Basic Needs,
xx. Scale Invariance,
xxi. Normalization,
xxii. Subgroup Decomposability,
xxiii. Factor Decomposability,
xxiv. Transfer Axiom,
xxv. Nondecreasing Poverty under Correlation Increasing Arrangement.

Some axioms are passed only for particular ranks of FGT index (α values in formula 1.1.4), moreover it does not satisfies the continuity axiom. However using fuzzy approach could overcome the latter disadvantage.

1.2.10. Fuzzy Set Approach

The fuzzy set approach, first proposed by Cerioli and Zani (1990), was born by the necessity of overcome the simple dichotomization of the population into poor and non poor defined in relation to some chosen poverty line. Poverty is not an attribute that characterises an individual in terms of presence or absence, but is rather a vague predicate that manifests itself in different shades and degrees. This approach will be largely explained in the next paragraph.

1.3. The Fuzzy approach

As explained above, fuzzy approach considers poverty as a matter of degree rather than an attribute that is simply present or absent for individuals in the population. In this case, two additional aspects have to be introduced:

iii. The choice of membership functions, i.e. quantitative specification of individuals' or households' degrees of poverty and deprivation;

iv. The choice of rules for the manipulation of the resulting fuzzy sets, as complements, intersections, union and aggregation.

Given a set *X* of elements $x \in X$, any fuzzy subset *A* of *X* will be defined as:

$$A = \{x, \mu_A(x)\} \qquad \forall x \in X \tag{1.3.1}$$

where $\mu_A(x): X \rightarrow [0,1]$ is called the *membership function* (*m.f.*) in the fuzzy subset *A* and its value indicates the degree of membership of *x* in *A*. Then $\mu_A(x) = 0$ means that *x* does not belong to *A*, whereas $\mu_A(x) = 1$ means that *x* belongs to *A* completely. When $0 < \mu_A(x) < 1$ then *x* partially belongs to *A* and its degree of membership of *A* increases in proportion to the proximity of $\mu_A(x)$ to 1.

1.3.1. Fuzzy monetary

In the conventional approach, the *m.f.* may be seen as $\mu(y_i) = 1$ if $y_i < z$, $\mu(y_i) = 0$ if $y_i \geq z$ where y_i is the equivalised income of individual *i* and *z* is the poverty line.

Cerioli and Zani (1990) have been the first authors to incorporate the concept of poverty as a matter of degree at the methodological level following the theory of Fuzzy Sets proposed by Zadeh (1965).

Let *y* be the known total income. The membership function to poor set can be defined by fixing a value z_1 up to which an individual is definitely poor and a value z_2 above which an individual is definitely not poor. For incomes between z_1 and z_2 the membership function takes value in [0, 1] and declines linearly. Formally:

$$\mu_i = 1 \; if \; y_i < z_1; \quad \mu_i = \frac{z_2 - y_i}{z_2 - z_1} \; if \; z_1 \le y_i \le z_2; \quad \mu_i = 0 \; if \; y_i > z_2 \qquad (1.3.2)$$

The traditional approach is a particular case of the fuzzy approach with $z_1 = z_2 = z$.

Cheli and Lemmi (1995) in their *Totally Fuzzy and Relative* approach attempted to overcome the limits of Cerioli and Zani membership function, that is, the arbitrary choice of the two threshold value and the linear form of the function within such values. They defined the *m.f.* as the distribution function $F(y_i)$ of income, normalized (linearly transformed) so as to equal 1 for the poorest and 0 for the richest person in the population. Formally:

$$\mu_i = (1 - F_{(M),i}) \qquad (1.3.3)$$

where F_i is the income distribution function. By definition, the mean of this *m.f.* is always 0.5. In order to make this mean equal to some specified value (such as 0.1) so as to facilitate comparison with the conventional poverty rate, Cheli (1995) takes the *m.f.* as normalized distribution function, raised to some power $\alpha \ge 1$. Formally:

$$\mu_i = FM_i = (1 - F_{(M),i})^{\alpha} = \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma \mid y_\gamma > y_i}{\sum_{\gamma=2}^{n} w_\gamma \mid y_\gamma > y_1} \right)^{\alpha}, \; i = 1, 2, ..., n; \; \mu_n = 0 \qquad (1.3.4)$$

where y_i is the equivalised income of the *i*-th individual, $F_{(M),i}$ is the value of the income distribution function $F(y_i)$ for the *i*-th individual, $(1 - F_{(M),i})$ is the proportion of individuals less poor than the person concerned with mean ½ by definition, w_γ is the sample weight of individual of rank γ ($\gamma = 1,...,n$) in the ascending income distribution and α is a parameter.

The value of α is arbitrary, but Cheli and Betti (1999) have chosen the parameter α so that the mean of the *m.f.* is equal to the head count ratio computed for the official poverty line. Increasing the value of this exponent implies giving more weight to the poorer end of the income distribution.

Betti and Verma (1999) have used a somewhat refined version of the expression (1.3.4) in order to define what they called Fuzzy Monetary indicator (FM):

$$\mu_i = FM_i = (1 - L_{(M)i})^{\alpha} = \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma y_\gamma \mid y_\gamma > y_i}{\sum_{\gamma=2}^{n} w_\gamma y_\gamma \mid y_\gamma > y_1} \right)^{\alpha}, \; i = 1, 2, ..., n; \; \mu_n = 0 \qquad (1.3.5)$$

where y_γ is the equivalised income and $L_{(M),i}$ represent the value of the Lorenz curve of income for individual *i*; then $1-L_{(M),i}$ represents the share of the total equivalised income received by all individuals who are less poor than the person concerned. It varies from 1 for the poorest to 0 for the richest individual. The mean of $1-L_{(M),i}$ values equals (1+G)/2, where G is the Gini coefficient of the distribution.

1.3.2. Fuzzy supplementary

In addition to the level of monetary income, the standard of living of households and individuals can be described by a host of indicators, such as housing conditions, possession of durable goods, perception of hardship, expectations, norms and values.

To quantify and put together diverse indicators of deprivation several steps are necessary. Specially, decisions are required to assigning numerical values to the ordered categories, weighting the score to construct composite indicators, choosing their appropriate distributional form and scaling the resulting measures in a meaningful way.

Choice and grouping of indicators

Firstly, from the large set which may be available, a selection has to be made of indicators which are substantively meaningful and useful for a given analysis. Secondly, it is useful to identify the underlying dimensions and to group the indicators accordingly (Whelan *et al.* 2001).

One possibility would be to construct a summary index of deprivation employing all items chosen for the analysis. However, Nolan and Whelan (1996) claim that this might well be unsatisfactory because, aggregating the items into a single index ignores the fact that different items may reflect different dimensions of deprivation, and adding them together may lose valuable information. Whelan *et al.* 2001 suggest, as the first stage in an analysis of life-style deprivation, examine systematically the range of deprivation items to see whether the items cluster into distinct groups. Factor analysis can be used to identify such clusters of interrelated variables. The procedure consists in an exploratory factor analysis to see how many factors are the optimal solution. It is possible then proceed to make use of confirmatory factor analysis to compare goodness of fit for different number of factor solution.

Following Kelloway (1998) we can consider the measures of absolute, relative and parsimonious fit as follows:

- The Root Mean Squared Error of Approximation (RMSEA) is based on the analysis of residuals with smaller values indicating a good fit. Values below 0.1, 0.05 and 0.01 indicate a good, very good and outstanding fit respectively.
- The Adjusted Goodness of Fit Index (AGFI) is based on the ratio of the sum of the squared discrepancies to the observed variances, but adjusts for degrees of freedom. The AGFI ranges from 0 to 1 with values above 0.9 indicating a good fit.
- The Normal Fit Index (NFI) indicates the percentage improvement in fit over the baseline independence model.

- The Comparative Fit Index (CFI) is based on the non-central X2, and is given by 1-[(X2 model –df model)/(X2 independence –df independence)]. The CFI ranges between 0 and 1, with values exceeding 0.90 indicating a good fit.
- The Parsimonious Goodness of Fit Index (PGFI) adjusts GFI for the number of estimated parameters in the model and the number of data points. The values of the PGFI range from 0 to 1 but it is unlikely to reach the 0.09 cut-off used for other indices and is best used to compare two competing models.

Deprivation indices can be constructed after the selection of indicators and dimensions following three different possible methods:

i. A straightforward additive procedure where the number of items on which the individual/household is deprived is simply summed (ESRI);
ii. An additive measure that considers factor scores from the homonym analysis (Lelli, 2001);
iii. A mean of the deprivation scores for each individual and dimension, weighting each item by the extent to which deprivation of that kind is experienced in the population in question. This last is of course a fuzzy measure.

<u>*Assigning numerical value to ordered categories*</u>

Moreover, it is necessary to assign numerical values to the ordered categories and to weight and scale measures. Individual items indicating non-monetary deprivation often take the form of simple "yes/no" dichotomies or sometimes ordered polytomies. The simplest scheme for assigning numerical values to categories is by assigning that the ranking of the categories represents an equally-spaced metric variable. Cerioli and Zani (1990) defined the membership function of an individual as follows.

If a vector of *k* categorical variables $X_1,...,X_k$ is observed on the *n* individuals of the population, the membership function of the fuzzy set of the poor can be defined as:

$$\mu_A(i) = \frac{\sum_{j=1}^{k} g(x_{ij}) w_j}{\sum_{j=1}^{k} w_j}, \qquad i = 1,...,n \qquad (1.3.6)$$

where $g(x_{ij}) = 1$ if the corresponding x_{ij} denotes deprivation and $g(x_{ij}) = 0$ otherwise. w_j denotes the weight of the variable X_j (*j* = 1, …, *k*).

If variable X_j is of ordinal scale, it is possible to identify a modality $x_j^{'}$ of X_j denoting lack of resources and a modality $x_j^{''}$ that excludes poverty. These modality are put in decreasing order beginning with the one that denotes the greatest deprivation. If $\psi_j^{'}$, $\psi_j^{''}$, ψ_{ij} represent the score of categories $x_j^{'}$, $x_j^{''}$, x_{ij} respectively, then:

$$g(x_{ij}) = \begin{cases} 1 & \text{if } \psi_{ij} \le \psi'_j \\ \dfrac{\psi'_j - \psi_{ij}}{\psi''_j - \psi'_j} & \text{if } \psi'_j \le \psi_{ij} \le \psi''_j \\ 0 & \text{if } \psi_{ij} \ge \psi''_j \end{cases} \tag{1.3.7}$$

For the weights w_j, Cerioli and Zani proposed the following specifications:

$$w_j = \ln \frac{1}{p_j} \tag{1.3.8}$$

where p_j is the proportion of individuals with deprivation in variable X_j. Substituting (1.3.8) in (1.3.6) we obtain:

$$\mu_A(i) = \frac{\sum_{j=1}^{k} g(x_{ij}) \ln \frac{1}{p_j}}{\sum_{j=1}^{k} \ln \frac{1}{p_j}} \tag{1.3.9}$$

A collective index of poverty is simply obtained by Cerioli and Zani using the relative cardinality (Dubois and Prade, 1980) of the fuzzy set of the poor: $|A| = \sum_{i=1}^{n} \mu_A(i)$. Such an index, included between 0 and 1, represents the proportion of individuals that belong to the fuzzy subset of the poor and it is given by:

$$P = \frac{|A|}{n} \tag{1.3.10}$$

Cheli and Lemmi (1995) proposed an improvement by replacing the simple ranking of the categories with their distribution function in the population. Formally:

$$g(x_{ij}) = H(x_j) \tag{1.3.11}$$

where $H(x_j)$ is the sampling distribution function of the variable X_j. The normalised form is given by:

$$g(x_{ij}) = g(x_j^{(k)}) = \begin{cases} 0 & \text{if } x_{ij} = x_j^{(1)}; k = 1 \\ g(x_j^{(k-1)}) + \dfrac{H(x_j^{(k)}) - H(x_j^{(k-1)})}{1 - H(x_j^{(1)})} & \text{if } x_{ij} = x_j^{(k)}; k > 1 \end{cases} \tag{1.3.12}$$

where $x_j^{(1)},...,x_j^{(m)}$ represent the categories of the variable X_j arranged in increasing order with respect to the risk poverty and $H(x_j^{(k)})$ is the distribution function of the variable X_j once its categories have been arranged as described above.

In this way, a 0 *m.f.* value is always associated with the modality corresponding to the lowest risk of poverty, whereas value 1 is associated with the modality corresponding to the highest risk. Cheli and Lemmi proposed the following weights:

$$w_j = \ln(1/\overline{g(x_j)}) \tag{1.3.13}$$

where $\overline{g(x_j)} = \frac{1}{n}\sum_{i=1}^{n} g(x_{ij})$ represents the fuzzy proportion of the poor with respect to X_j and if X_j is dichotomic it coincides with the crisp proportion p_j.

<u>Weighting for constructing composite measures</u>

An early attempt to choose an appropriate weighting system of several indicators at macro level data was made by Ram (1982), using principal components analysis, which was also adopted by Maasoumi and Nickelsburg (1988). For the construction of fuzzy measures, however, it is necessary to weight and aggregate items at the micro level. At the micro level, Nolan and Whelan (1996) adopted factor analysis. In order also to give more weight to more widespread items, Cerioli and Zani (1990) specified the weights of any item as a function of the proportion deprived of the item. Betti and Verma (1999) proposed to determinate the weights to be given to items in the aggregation within each dimension separately. Also, the set of weights are taken to be item-specific; for a given item they are common to all individuals in the population. Such weights comprise two factor: The first factor is determined by the variable's dispersion and it may be taken as proportional to the coefficient of variation of deprivation score for the variable concerned $w_k^a \propto cv_k$. This means that when an item of deprivation affects only a small proportion, the weight given to it varies inversely to the square-root of the proportion. Thus deprivations which affect only a small proportion of the population, and hence are likely to be considered more critical, get larger weights at the micro level; while those affecting large proportions, hence likely to be regarded as less critical, get smaller weights. However, the contribution of the deprived individuals to the average value of deprivation in the population resulting from the item concerned turns out to be directly proportional to the square-root of *d*. In other words, deprivation affecting a small proportion of the population is treated as more intense at the individual person's level but, of course, its contribution to the average level of deprivation in the population as a whole is correspondingly smaller. The second factor is determined in order to control redundancy. To do this, it is necessary to limit the influence of characteristics that are highly correlated with others included in the analysis. Even for the overall index, it is reasonable to consider this correlation separately within each of the five dimensions of deprivation identified earlier, i.e. to take the weight of item *k* in deprivation dimension *h* as the inverse of an average measure of its correlation with items in that dimension. Thus the results are not affected by arbitrary inclusion or exclusion of items highly correlated with other items in the dimension. An average measure of the correlation can be computed as:

$$w_k^b \propto \left(\frac{1}{1+\sum_{k'}^{K} \rho_{k,k'} \mid \rho_{k,k'} < \rho_H} \right) * \left(\frac{1}{1+\sum_{k'}^{K} \rho_{k,k'} \mid \rho_{k,k'} > \rho_H} \right) \qquad (1.3.14)$$

where $\rho_{k,k'}$ is the correlation between the two indicators corresponding to items *k* and k'. In the first factor of the equation, the sum is taken over all indicators whose correlation with variable *k* is less than a certain threshold ρ_H (determined, for instance, by the point of largest gap between the ordered set of correlation values encountered). The motivation for this model is that (i) w_k^b is not affected by the introduction of variables entirely uncorrelated with *k*; (ii) it is only marginally affected by small correlations; but (iii) is reduced proportionately to the number of highly correlated variables present. To surmise, the weight given to an item is directly proportional to the variability of the item in the population and inversely proportional to its average correlation with items in the deprivation dimension to which it belongs. The final weight is taken as proportional to the product of the two factors: $W_k \propto w_k^a \cdot w_k^b$. The scaling of the weights can be arbitrary, although scaling them to sum to 1.0 within each dimension is convenient.

<u>*Functional form of the distribution*</u>

As in the Fuzzy Monetary approach, the individual's degree of non-monetary deprivation FS_{hi} for each dimension $(h:1,...,m)$ can be defined in two alternative manners:

i. The proportion of individuals who are less deprived than *i*:

$$\mu_i = FS_{hi} = (1-F_{(S),hi})^{\alpha} \qquad (1.3.15)$$

where $F_{(S),hi}$ is the distribution function of *S* evaluated for individual *i* dimension *h*.

ii. The share of the total non-deprivation *S* assigned to all individuals less deprived than *i*:

$$\mu_i = FS_{hi} = (1-L_{(S),hi})^{\alpha} \qquad (1.3.16)$$

where $L_{(S),hi}$ is the value of the Lorenz curve of *s* for individual *i* in dimension *h*, calculated according to the form as follows:

$$1-L_{(S),hi} = \left[\frac{\sum_{\gamma=i+1}^{n} w_{h\gamma} s_{h\gamma} \mid s_{h\gamma} > s_{hi}}{\sum_{\gamma=2}^{n} w_{h\gamma} s_{h\gamma} \mid s_{h\gamma} > s_{h1}} \right] \qquad (1.3.17)$$

The parameter α is determined so as to make the overall non-monetary deprivation rate numerically identical to the monetary poverty rate *H*.

1.3.3. Combination across dimensions of deprivation

In the previous sections, we have defined fuzzy measures of poverty and deprivation in multiple dimensions: monetary poverty on the one hand, and non-monetary deprivation in different aspects of life, on the other. The next step of interest in multidimensional analysis is to identify the extent to which deprivation in different dimensions tends to overlap for individual units, households or persons.

It is often also useful to construct measures of deprivation averaged over different dimensions in someway. The common objective of these procedures is to summarise the diverse dimensions of deprivation in terms of fewer indicators—ultimately perhaps in terms of a single indicator of the 'overall' level of deprivation, permitting an unambiguous ranking of individuals in the population.

For this purpose some aggregation operations on the fuzzy sets have to be defined.

Let $\mu_{i,h}$ be the degree of deprivation in dimension *h* for individual *i*. We seek measures of the type:

$$\mu_{i.} = f(\mu_{i1}, \mu_{i2}, \mu_{i3}, ...) \qquad (1.3.18)$$

which summarise in an appropriate and useful way some common aspects of deprivation in different dimensions to which an individual is subject.

Three types of aggregation operations are useful for this aim:

Intersection. This is relevant when the interest is in the simultaneous presence of deprivation in different dimensions. Mathematically (1.3.18) takes the form of set intersections. The resulting measure $\mu_{i.}$ reflects the extent to which the different forms of deprivation overlap for the individual *i* concerned. Large overlaps may be seen as implying that the different forms do not really reflect distinct dimensions, or alternatively, that the resulting deprivation is more intense for being present in multiple dimensions simultaneously.

Union. This is relevant when the interest is in the presence of deprivation, irrespective of its particular form or dimension. Here (1.3.18) takes the form of union of sets.

Averages. Averaging of the individual's degrees of deprivation in different dimensions can be meaningful when all forms of deprivation are important and, in that sense, are compensatory among themselves. Their combined effect can therefore be summarised by averaging in some appropriate form. Deprivation in some dimensions may be considered more important or intense than in other dimensions. In this case a weighted average may be more appropriate, with the weight determined to reflect relative importance of the dimensions. Apart from ordinary mean (or median), one may use a generalised weighted mean, say the form:

$$\mu_{i.} = \left(\sum_j \overline{w}_j \mu_{i,j}^{\beta} / \sum_j \overline{w}_j \right) \qquad (1.3.19)$$

where $\beta \neq 0$ is a parameter chosen to obtain means of different types. ($\beta = 0$ corresponds to arithmetic mean).

In order to illustrate the set intersection and union operators, let us consider only two dimensions of deprivation, monetary poverty *m*, and non-monetary deprivation *s*. In the conventional, 'crisp' formulation, individuals are categorised as deprived and non-deprived in each of the two dimensions. We can view any individual as belonging to one and only one of the four subpopulations defined by the intersections $m \cap s$ (m, s = 0,1).

Fuzzy set operations are a generalisation of the corresponding 'crisp' set operations in the sense that the former reduce to (exactly reproduce) the latter when the fuzzy membership functions, being in the whole range [0,1], are reduced to a 0,1 dichotomy.

There are, however, more than one ways in which the fuzzy set operations can be formulated, each representing an equally valid generalisation of the corresponding crisp set operations. The choice among alternative formulations has to be made primarily on substantive grounds: some options are more appropriate (meaningful, convenient) than others, depending on the context and objectives of the application. While the rules of fuzzy set operations cannot be discussed fully in this chapter, we need to clarify their application specifically for the study of poverty and deprivation.

Since fuzzy sets are completely specified by their membership functions, any operation with them is defined in terms of the membership functions of the original fuzzy sets involved. For simplicity, let be (a, b) the membership functions of two sets for individual i, where $a = FM_i$ and $b = FS_i$, $s_1 = \min(a,b)$, $s_2 = \max(a,b)$ and $\bar{a} = 1 - a$, $a \cap b$, $a \cup b$ the basic set operations of complementation, intersection and union.

Table 1.3.1 displays the most common ways to specify fuzzy intersection and union that satisfy a set of essential requirements such as 'reduction to the crisp set operation', 'boundary condition', 'monotonicity', 'cummutativity', etc. (for details see Klir and Yuan, 1995).

Table 1.3.1. Basic forms of fuzzy set intersections and unions

	Intersection $a \cap b$	Union $a \cup b$
Standard	$i(a, b) = \min(a, b) = i_{\max}$	$u(a, b) = \max(a, b) = u_{\min}$
Algebraic	$i(a, b) = a*b$	$u(a, b) = a + b - a*b$
Bounded	$i(a, b) = \max(0, a + b - 1)$	$u(a, b) = \min(1, a + b)$

The Standard fuzzy operations provide the largest intersection and by contrast the smallest union among all the permitted forms. They are appropriate for intersection and union of similar fuzzy sets, i.e. sets for which the membership functions are expected to have a substantial positive correlation, but not uniformly throughout in the application to poverty analysis because their sum would exceed 1 and the marginal constraints would not be satisfied. An obvious example is a pair of sets, one defining the degree of income poverty, and the other deprivation of a certain type such as 'basic non monetary deprivation'. The Standard fuzzy intersection and union are the only ones which satisfy the desirable condition of "idempotency", i.e. intersection(a, a) = union(a, a) = a.

The Bounded operator is appropriate for the aggregation of dissimilar sets for which the membership functions are expected to have a substantial negative correlation. This, for example, will be the case with one set defining the degree of presence of poverty, and the other defining the degree of absence of deprivation in a certain dimension.

The Algebraic operator is appropriate for the aggregation of sets in the absence of such correlations. It is the only one that satisfies the marginal constraints but it could give non acceptable results.
Betti and Verma (2004) proposed to use in the analysis of fuzzy sets defining deprivation in different dimensions the so called 'Composite' set operator:

1. For sets representing similar states – such as the presence or absence of both types of deprivation – the Standard operations (which provide larger intersections than Algebraic operations) are used.
2. For sets representing dissimilar states- such as the presence of one type but the absence of the other type of deprivation – the Bounded operations (which provide smaller intersections than Algebraic operations) are used.

A possible, more flexible, but of course more demanding on data and substantive judgement alternative would be to consider a weighted combination the Composite and Algebraic set operators, for instance in the following form, which also meets the consistency requirement:

1. For sets representing similar states $\rightarrow$ (1-*w*)(Standard) + *w*(Algebraic)
2. For sets representing dissimilar states $\rightarrow$ (1-*w*)(Bounded) + *w*(Algebraic)

Parameter *w* can be thought of as a measure of the degree to which different types of states can be distinguished. When $w = 0$ we have the Composite scheme defined above, with its sharp distinction between similar and dissimilar states. With $w = 1$, we have the Algebraic scheme, applicable when the different states are 'neutral' with respect to each other. With $0 < w < 10$, one may represent intermediate types of situations.

Table 1.3.2 shows the application of this Composite set operations and Graph. 1.3.1 illustrates them graphically.

Table 1.3.2. Joint measures of deprivation according to the Betti and Verma Composite operation

		Non-monetary deprivation		
		non-poor (0)	poor (1)	Total
Monetary deprivation	non-poor	$\min(1-FM_i, 1-FS_i) =$ $1-\max(FM_i, FS_i)$	$\max(0, FS_i - FM_i)$	$1-FM_i$
	poor	$\max(0, FM_i - FS_i)$	$\min(FM_i, FS_i)$	FM_i
	Total	$1-FS_i$	FS_i	1

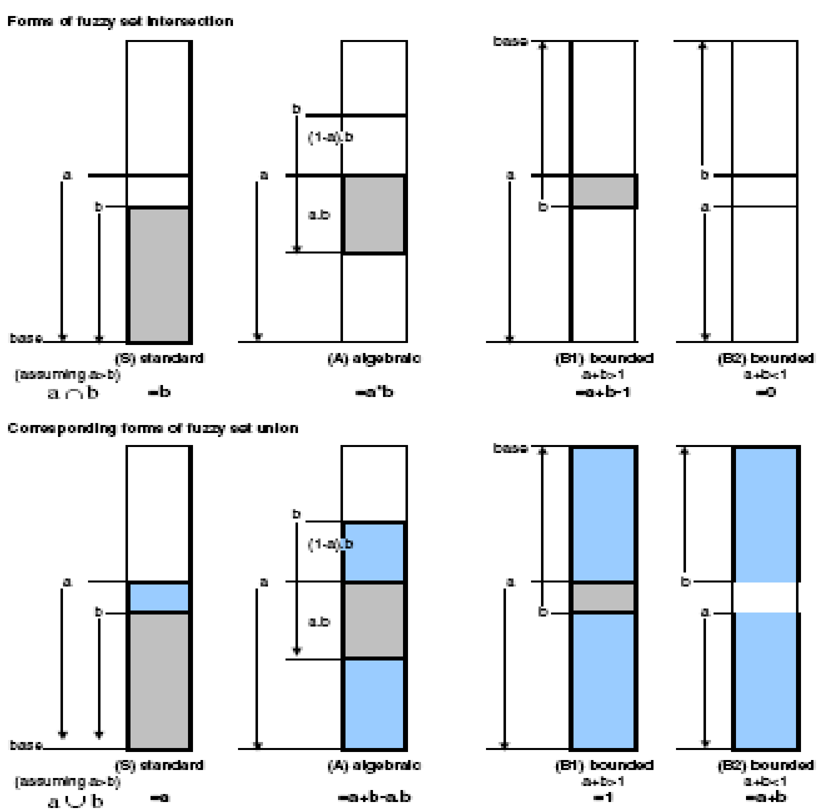
In the Graph 1.3.1, the 'universal set' X (i.e. membership = 1 for any element of the population of interest) is represented by a rectangle of unit length, and within it are placed the units' membership functions ($0 \leq a \leq 1$, $0 \leq b \leq 1$) on the two subsets. Different placements correspond to different types of fuzzy set operations. It can be seen that the Standard intersection and union are obtained by placing the two sets (a, b) on the same base, so that the smaller (say b) lies completely within the larger (say a). Consequently, their intersection is maximised, so as to equal the smaller of the sets. By the same token, their union is minimised, so as to equal the larger of the sets. It is for this reason that the Standard operator is appropriate for the aggregation of similar fuzzy sets. By this we mean sets representing similar states, such as the presence of deprivation in different dimensions, the membership of which can be expected to be positively correlated, often quite strongly.

Similarly, we can see that the Bounded form is obtained by placing the two sets at the opposite ends of X, thus minimising their intersection and maximising their union. The intersection is reduced to (a + b – 1), which is non-zero only if a + b > 1; the union is increased to (a + b), or to 1 if a + b > 1. The Bounded operators are appropriate for the aggregation of dissimilar sets, such as one representing the presence of deprivation in a certain dimension while the other representing the absence of deprivation in another dimension. Membership of such dissimilar sets can be expected to be negatively correlated.

The Algebraic form is obtained by placing one set (say b) symmetrically over the other (a), such that it overlaps (a) and (1 - a) proportionally: the overlaps being b·a and b·(1 - a), respectively. This can be seen to imply a lack of correlation between a unit's membership in the two sets.

Obviously, by placing one of the two sets on the base of X, and shifting the placement of the other (within X), a continuum of different types of fuzzy set operators can be generated.

Figure. 1.3.1. Graphical representation of the fuzzy set operations



The propensity to income poverty, FM_i, and the overall non-monetary deprivation propensity, FS_i, may be combined to construct composite measures which indicate the extent to which the two aspects of income poverty and non-monetary deprivation overlap for the individual concerned. These measures, at the individual level *i*, are:

i. *Manifest deprivation* (MAN_i), representing the propensity to both income poverty and non-monetary deprivation simultaneously:
ii. *Latent deprivation* (LAT_i), representing the individual being subject to at least one of the two, income poverty and/or non-monetary deprivation.

The corresponding combined measures are obtained using the Composite set operations. The Manifest deprivation propensity of individual *i* is the intersection (the smaller) of the two (similar) measures FM_i and FS_i:

$$MAN_i = \min(FM_i, FS_i) \qquad (1.3.20)$$

Similarly, the Latent deprivation propensity of individual *i* is the complement of the intersection indicating the absence of both types of deprivation, i.e. the union (the larger) of the two (similar) measures FM_i and FS_i:

$$LAT_i = 1 - \min(\overline{FM}_i, \overline{FS}_i) = \max(FM_i, FS_i) \qquad (1.3.21)$$

From empirical experience (Betti and Verma 2002; Betti *et al.* 2005), it appears that the degree of overlap between income poverty and non-monetary deprivation at the level of individual persons tend to be higher in poorer areas and lower in richer areas. A useful indicator in this context is the Manifest deprivation index defined as a percentage of Latent deprivation index and included between 0 and 1. When there is no overlap (i.e., when the subpopulation subject to income poverty is entirely different from the subpopulation subject to non-monetary deprivation), Manifest deprivation rate and hence the above mentioned ratio equals 0. When there is complete overlap, i.e., when each individual is subject to exactly the same degree of income poverty and of non-monetary deprivation, the Manifest and Latent deprivation rates are the same and hence the above mentioned ratio equals 1.

1.4. The proposed approach: Integrated Fuzzy and Relative (IFR)

In this section we introduce a new approach to the analysis of poverty and deprivation that may be called *Integrated Fuzzy and Relative* (IFR) approach.

This measure, proposed by Betti *et al*. (2006), combines the TFR approach of Cheli and Lemmi (1995) and the approach of Betti and Verma (1999), seen in the previous paragraphs.

The new fuzzy poverty measure described in the next paragraphs also has an economic meaning, in that it is expressible in terms of the generalised Gini measures.

1.4.1. Income poverty

In this approach both the share of individuals less poor than the person concerned (as in Cheli and Lemmi, 1995) and the share of the total equivalised income received by all individuals less poor than the person concerned (as in Betti and Verma, 1999) are take into account. Specifically, the measure is defined as:

$$\mu_i = FM_i = (1 - F_{(M),i})^{\alpha-1}(1 - L_{(M),i}) = \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma \mid y_\gamma > y_i}{\sum_{\gamma=2}^{n} w_\gamma \mid y_\gamma > y_1} \right)^{\alpha-1} \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma y_\gamma \mid y_\gamma > y_i}{\sum_{\gamma=2}^{n} w_\gamma y_\gamma \mid y_\gamma > y_1} \right) \quad (1.4.1)$$

where, as in section 1.3, y_γ is the equivalised income, $F_{(M),i}$ is the income distribution function, w_γ is the sample weight of individual of rank γ ($\gamma = 1,...,n$) in the ascending income distribution, $L_{(M),i}$ represent the value of the Lorenz curve of income for individual i.

The parameter α, as in the previous approaches, is chosen so that the mean of the *m.f.* is equal to the head count ratio H:

$$E(FM) = \frac{\alpha + G_\alpha}{\alpha(\alpha+1)} = H \quad (1.4.2)$$

The Fuzzy Monetary measure as defined above is expressible in terms of the generalised Gini measure G_α. This family of measures is a generalisation of the standard Gini coefficient with $\alpha = 1$ and it is defined (in the continuous case) as:

$$G_\alpha = \alpha(1-\alpha)\int_0^1 \left\{(1-F)^{\alpha-1}(F - L(F))\right\} dF \quad (1.4.3)$$

This measure weights the distance $(F - F(L))$ between the line of perfect equality and the Lorenz curve by a function of the individual's position in the income distribution, giving more weight to its poorer end.

1.4.2. Non–monetary deprivation

Betti *et al*. (2006) proposed to treat the non-monetary scores in a way entirely analogous to that for monetary poverty measures, described above. On the basis of the *Integrated Fuzzy and Relative* approach, the function corresponding to equation (1.4.1) would be:

$$\mu_i = FS_i = \left(1 - F_{(S),i}\right)^{\alpha-1}\left(1 - L_{(S),i}\right) = \left[\frac{\sum_{\gamma=i+1}^{n} w_\gamma \mid s_\gamma > s_i}{\sum_{\gamma=2}^{n} w_\gamma \mid s_\gamma > s_1}\right]^{\alpha-1}\left[\frac{\sum_{\gamma=i+1}^{n} w_\gamma s_\gamma \mid s_\gamma > s_i}{\sum_{\gamma=2}^{n} w_\gamma s_\gamma \mid s_\gamma > s_1}\right],$$

$$i = 1, 2, ..., n; \mu_{hn} = 0 \qquad (1.4.4)$$

where $F_{(S),i}$ represents the distribution function of the overall supplementary deprivation (*S*) evaluated for individual *i*, and $L_{(S),i}$ the value of the Lorenz curve of *S* for individual *i*. The parameter α is determined so as to take the overall non-monetary deprivation rate numerically identical to the monetary poverty rate *H*.

Chapter 2

Data Analysis at national level

2.1. EU-SILC survey

In the present analysis we used data from the European Survey on Income and Living Conditions (EU-SILC), the major new source of comparative statistics on income and living conditions in Member States of the European Union and some neighbouring countries. EU-SILC is the successor of the ECHP (European Community Household Panel) a data collection used over the 1994-2001 period. It has been developed to overcome, or at least ameliorate the main shortcomings of the previous one, like the problem of sample attrition, loss of representativeness, lack of flexibility in the design and content of the survey, lack of timeliness in the production of the data, lack of sustainability of the survey, etc…

Flexibility is an essential feature of EU-SILC. This means that EU-SILC dataset may comprise different types and combinations of data sources, with different designs.

In fact, it has been developed as a flexible yet comparable instrument for the follow-up and monitoring of poverty and social exclusion at the EU and national levels. It covers data and data sources of various types: cross-sectional and longitudinal; household-level and person-level; economic and social; from registers and interview surveys; from new and existing national sources.

It envisages the creation of one or more micro-data bases in each country, to be used for the follow-up and monitoring of poverty and social exclusion at the EU and national levels.

Depending on the country, micro-data could come from:

i. one existing national source (survey or register);
ii. two or more existing national sources (surveys and/or registers) directly linkable at micro-level;
iii. one or more existing national sources combined with a new survey — all of them directly linkable at micro-level;
iv. a new harmonised survey (or survey system) to meet all EU-SILC requirements.

The standard integrated design involves a rotational panel in which a new sample of households and persons is introduced each year to replace a part (normally one quarter) of the existing sample. Persons enumerated in each new sample are followed-up in the survey for four (or more) years. A common rotational sample of this type yields each year a cross-sectional sample as well as longitudinal samples of various durations. In most situations, these sample data have to be weighted to make them more representative of the target population of the survey. The complex structure of the sample means that the corresponding weighting procedures can also be quite complex.

Households form the basic units of sampling, data collection and data analysis. To ensure comparability of key indicators a rigorous and harmonised definition of the household is necessary for all countries. In the standard EU-SILC definition, a private household means a person living alone or a group of people who live together in the same private dwelling and share expenditures, including the joint provision of the essentials of living.

Four types of data are involved: (i) variables measured at the household level; (ii) information on household size and composition and basic characteristics of household members; (iii) income and other more complex variables measured at the personal level, but aggregated to construct household-level variables (which may then be ascribe to each member for analysis); and (iv) more complex non-income or 'social' variables collected and analysed at the personal level.

For set (i)-(iii) variables, a sample of households including all household members is required. Among these, sets (i) and (ii) are normally collected from a single, appropriately designated respondent in each sample household. Alternatively, some or all of these data may be compiled from registers or other administrative sources.

Set (iii) variables - concerning mainly, but not exclusively, the detailed collection of household and personal income - must be collected directly at the personal level, covering all persons in each sample household. In many countries, these income and related variables are collected through personal interviews with all adults aged 16+ in each sample household. This collection is normally combined with that for set (iv) variables, since the latter must also be collected directly at the personal level. These are the so-called survey countries. By contrast, in some countries, set (iii) variables are compiled from registers and other administrative sources, thus avoiding the need to interview all members (adults aged 16+) in each sample household. These are the so-called register countries.

Set (iv) variables, for their complex and personal nature, are normally collected through direct personal interview in all countries. For the survey countries, this collection is normally combined with that for set (iii) variables as noted above, covering all persons in each sample household even if it is not essential since these variables need not to be aggregated to the household level. Register countries take in account a representative sample of persons, since for these countries interviewing all household members for set (iii) is not involved.

Then, different types of units of analysis are involved in EU-SILC for which sample weights have to be defined: (i) private households; (ii) all persons residing in sample households; (iii) all household members aged 16+; and optionally (iv) one selected adult per sample household. One may also be interested in special groups, such as children[2].

Both cross-sectional and longitudinal data are required in EU-SILC. The cross-sectional component covers information pertaining to the current and a recent period such as the preceding calendar year. It aims at providing estimates of cross-sectional levels and of net

[2] Following the EU-SILC terminology, we will refer to these different types of units and to the associated data files as follows: H ('household'), R ('register' covering all members), P ('personal' covering all adults aged 16+), S ('selected respondent'), and Q (children or other special groups)

changes from one period (year) to another. The longitudinal component covers information compiled or collected through repeated enumeration of individual units, and then linked over time at the micro-level. It aims at measuring gross (micro-level) change and elucidating the dynamic processes of social exclusion and poverty. Both cross-sectional and longitudinal micro-data sets are updated on an annual basis. However, the first and clear priority is given to the production of comparable, timely and high quality cross-sectional data. Longitudinal data are limited in content and possibly also in sample size. Furthermore, for any given set of individuals, micro-level change is followed up only for a limited duration. In EU-SILC a period of four years is taken as the minimum duration for longitudinal follow-up at micro level.

Combining the various types of units and the time dimension, the new data sets disseminated each year consist of the following: cross-sectional data pertaining to the most recent reference year for households and persons; data pertaining to three different longitudinal periods, covering 2, 3 and 4 years preceding the survey, only for persons.

Another important aspect of EU-SILC is the choice of sample that depend on substantive requirements, cost constraints and practical considerations more than a number of additional factors due to comparative and multi-country nature of the survey (for details see Verma, 2001, Verma and Betti, 2006). A minimum effective sample size is fixed for both longitudinal and cross-sectional components. In particular, sample size is fixed according to a minimum level of reliability for the national estimate of the risk of poverty rate.

At this moment, data are available at cross-sectional level for years 2004, 2005, 2006 and 2007. In round 2004 only EU 15 countries are present; in rounds 2005, 2006, 26 countries are present and in rounds 2007 27 countries. Table 2.1.1, in the next page, shows the number of households interviewed for each country.

Table 2.1.1. EU-SILC household sample sizes (2004-2007)

Country	**2004**	**2005**	**2006**	**2007**
AT	4,521	5,148	6,028	6,806
BE	5,275	5,137	5,860	6,348
CY		3,746	3,621	3,505
CZ		4,351	7,483	9,675
DE		13,106	13,799	14,153
DK	6,866	5,957	5,711	5,783
EE	3,993	4,169	5,631	5,146
ES	15,355	12,996	12,205	12,329
FI	11,200	11,229	10,868	10,624
FR	10,273	9,754	10,036	10,498
GR	6,252	5,568	5,700	5,643
HU		6,927	7,722	8,737
IE	5,477	6,085	5,836	5,608
IS	2,907	2,928	2,845	2,872
IT	24,270	22,032	21,499	20,982
LT		4,441	4,660	4,975
LU	3,571	3,622	3,836	3,885
LV		3,843	4,315	4,471
MT				3,477
NL		9,356	8,986	10,219
NO	6,046	5,991	5,768	6,013
PL		16,263	14,914	14,286
PT	4,989	4,615	4,367	4,310
SE	5,748	6,133	6,803	7,183
SI		8,287	9,478	8,707
SK		5,147	5,105	4,941
UK		10,826	9,902	9,275
TOT	116,743	197,657	202,978	210,451

2.2. Imputation of data

Missing data problems can arise from diverse sources in a number of forms. We focused on the problem of imputation for item non-response but similar problems can arise when the information is available on some but not all the members of a household.

Imputing missing data aims to minimise the mean squared error of survey estimates, in particular the non-response bias component that arises when the pattern of missing data is not random and, more practically, to reach consistency between the results from different analyses and the convenience of not having to deal with the missing data problem at the analysis stage.

Missing values of variables using in this analysis are been imputed trough *IVEware* (Imputation and Variance Estimation Software) and in particular IMPUTE module. This is a multivariate imputation procedure that can handle relatively complex data structures (hundreds of variables, some continuous, others counts, many dichotomous or polytomous, and semi-continuous or limited dependent variables) when the data are missing at random.

IMPUTE module produces imputed values for each individual in the data set conditional on all the values observed for that individual. The imputations are obtained by fitting a sequence of regression models, depend on the type of variable being imputed, and drawing values from the corresponding predictive distributions specified by the regression model with a flat or non-informative prior distribution for the parameters in the regression model. Covariates include all other variables observed or imputed for that individual. The sequence of imputing missing values can be continued in a cyclical manner, each time overwriting previously drawn values, building interdependence among imputed values and exploiting the correlational structure among covariates. To generate multiple imputations, the same procedure can be applied with different random starting seeds or taking every *p*-th imputed set of values in the cycles mentioned above.

Five types of variables are assumed: (1) continuous; (2) binary; (3) categorical (polytomous with more than two categories); (4) counts; and (5) mixed (a continuous variable with a non-zero probability mass at zero). The types of regression models used are linear, logistic, Poisson, generalized logit or mixed logistic/linear, depending on the type of variable being imputed. IMPUTE take also into account two common features of survey data that add to the complexity of the modelling process: the restriction of imputations to subpopulations, and the bounding of imputed values. For details see Raghunathan *et al*. (2001).

2.3. Fuzzy Monetary Indicators

In order to calculate the Fuzzy Monetary Indicator (FM) we consider the distribution of household equivalised disposal income assigned to each individual. The distribution of the equivalised disposal income is trimmed taking as low bound the 15% of the median of the same distribution. Such indicator for the *i*-th individual is calculated using formula 1.4.1:

$$FM_i = (1 - F_{(M),i})^{\alpha-1}(1 - L_{(M),i}) \qquad (2.3.1)$$

2.4. Fuzzy Supplementary Indicators

Fuzzy Supplementary indicator has been calculated following these steps:

1. Identification of items;
2. Transformation of the items into the [0, 1] interval;
3. Exploratory and confirmatory factor analysis;
4. Calculation of weights within each dimension (each group);
5. Calculation of scores for each dimension;
6. Calculation of an overall score and the parameter α;
7. Construction of the fuzzy deprivation measure in each dimension (and overall).

2.4.1. Calculation of the deprivation score for each dimension and the overall score

Aggregation over a group of items in a particular dimension h (h = 1, 2, …, m) is given by a weighted mean taken over j items: $s_{hi} = \sum w_{hj} \cdot s_{hj,i} / w_{hj}$ where w_{hj} is the weight of the j-th deprivation variable in the h-th dimension (see section 2.4.7). An overall score for the i-th individual is calculated as the unweighted mean:

$$s_i = \frac{\sum_{h=1}^{m} s_{hi}}{m} \tag{2.4.1}$$

2.4.2. Calculation of the parameter α

We calculate the FS indicator for the i-th individual over all dimensions using formula 1.4.4:

$$FS_i = \left(1 - F_{(S),i}\right)^{\alpha - 1}\left(1 - L_{(S),i}\right) \tag{2.4.2}$$

As for FM indicator, the parameter α is determined so as to make the overall non-monetary deprivation rate numerically identical to the head count ratio computed for the official poverty line (60% of the median).

The parameter α estimated is used to calculate the FS indicator for every single dimension.

2.4.3. Construction of the fuzzy deprivation measure in each dimension

The FS indicator for the h-th deprivation dimension for the i-th individual is defined as combination of the $(1 - F_{(S),hi})$ indicator and the $(1 - L_{(S),hi})$ indicator .

$$\mu_i = FS_{hi} = \left(1 - F_{(S),hi}\right)^{\alpha - 1}\left(1 - L_{(S),hi}\right) = \left[\frac{\sum_{\gamma=i+1}^{n} w_{h\gamma} \mid s_{h\gamma} > s_{hi}}{\sum_{\gamma=2}^{n} w_{h\gamma} \mid s_{h\gamma} > s_{h1}}\right]^{\alpha - 1}\left[\frac{\sum_{\gamma=i+1}^{n} w_{h\gamma} s_{h\gamma} \mid s_{h\gamma} > s_{hi}}{\sum_{\gamma=2}^{n} w_{h\gamma} s_{h\gamma} \mid s_{h\gamma} > s_{h1}}\right],$$

$$h = 1, 2, ..., m; i = 1, 2, ..., n; \mu_{hn} = 0 \tag{2.4.3}$$

The $(1 - F_{(S),hi})$ indicator for the i-th individual is the proportion of individuals who are less deprived, in the h-th dimension, than the individual concerned. $F_{(S),hi}$ is the value of the score distribution function evaluated for individual i in dimension h and $w_{h\gamma}$ is the sample weight of the i-th individual of rank γ in the ascending score distribution in the h-th dimension.

The $(1-L_{(S),hi})$ indicator is the share of the total lack of deprivation score assigned to all individuals less deprived than the person concerned. $L_{(S),hi}$ is the value of the Lorenz curve of score in the h-th dimension for the i-th individual. The parameter α is calculated only once as shown in section 2.4.2.

2.4.4. Identification of items

Firstly, from the large set of EU-SILC variables, a selection has been made of indicators which are substantively meaningful and useful for the construction of Fuzzy Supplementary Indicators.

For our purpose, we have identified a set of items which could serve as indicators of concept of life-style deprivation. All these items are considered at household level, even if some of them are taken from the individual dataset and then they have been aggregated at household level.

The first set of items regards the lack of possession of a widely-desired item. These are:

- A telephone including mobile phone;
- A colour TV;
- A computer;
- A washing machine;
- A car.

In all these cases we consider a household to be deprived only if the lack of the item is enforced, because the household cannot afford the item.

A second set of items relates to the lack of ability to afford items that are considered as basic:

- Keeping home adequately warm;
- Paying for one week annual holiday away from home;
- Eating a mean with meat, chicken, fish (or vegetarian equivalent) every second day;
- Facing unexpected financial expenses.

A third set relates to absence of housing facilities so basic one can presume all household would wish to have them:

- A bath or shower in dwelling;
- An indoor flushing toilet for sole use of household.

The fourth set of items relate to problems with accommodation and the environment, with the implicit assumption that the households wish to avoid such difficulties:

- Leaking roof, damp walls/floors/foundation, or rot in window frames or floor;
- Too dark, not enough light in dwelling;

- Noise from neighbours or from the street;
- Pollution, grime or other environmental problems;
- Crime violence or vandalism in the area.

The fifth set relates the arrears that the household has experienced in the last 12 months;

- Arrears on mortgage or rent payments;
- Arrears on utility bills;
- Arrears on hire purchase instalments or other loan payments.

The sixth set is just an item related to the capacity of the household to make ends meet.

The seventh set relates to the health condition of the household. These items are from individual variables that have been aggregated at household level. We consider this dimension because we think that, in dealing with life-style deprivation, also the lack of health should be important. The items considered are:

- General health;
- Suffer from any chronic (long-standing) illness or condition;
- Limitation in activities because of health problems;
- Unmet need for medical examination or treatment;
- Unmet need for dental examination or treatment.

This dimension is not comparable for register countries, for which the unit of analysis is just the selected respondent.

The eighth set relates to the education. For this set we have constructed two composed indicators:

- Households with early school livers not in education or training;
  Households with at least one person aged 18-24 who have only lower secondary education (PE040: ISCED level currently attended: value 2 or less) and are not in education or training leading to a qualification at least to upper secondary (PE010: current education activity: value 2)
- Households with persons with low educational attainment.
  Households with at least one person aged 25-64 who have only lower secondary education or less (PE040).

The least dimension concerns the labour market. Also for this set we have constructed two composed indicators:

- Jobless households;
  This indicator identifies the worklessness of the household, using variable PL030. For details about the construction see next section.
- Intensity or duration of unemployment at household level.
  This indicator is constructed using variables PL070, PL072, PL080, PL085, PL087, PL090. For details about the construction see next section.

The variables used are listed in the Annex.

2.4.5. Transformation of the items into the [0, 1] interval

When the item is constituted by a fix number of categories, then it is transformed using the following procedure. For each item we determine a deprivation score as follows:

$$d_{j,i} = \frac{1 - F(c_{j,i})}{1 - F(1)}, \quad j = 1, 2, ..., k; i = 1, 2, ..., n \tag{2.4.4}$$

where $c_{j,i}$ is the value of the category of the *j*-th item for the *i*-th individual and $F(c_{j,i})$ is the value of the *j*-th item cumulation function for the *i*-th individual.

We transform the deprivation score in a positive score as:

$$s_{j,i} = 1 - \frac{1 - F(c_{j,i})}{1 - F(1)} = \frac{F(c_{j,i}) - F(1)}{1 - F(1)}, \quad j = 1, 2, ..., k; i = 1, 2, ..., n \tag{2.4.5}$$

In the particular, but common case, where the variable is a dichotomy, the deprivation index *d* is 1 for deprivation and 0 otherwise, while the positive score *s* is 0 for deprivation and 1 otherwise. For polychotomous items we assign to each household instead of the real value of the category, a value corresponding to the percentage of households that are "better off" then it. In the few cases in which the indicator is a composite one the score *s* represents the percentage of people in the household that experienced it.

In particular the indicator concerning the worklessness of the household is constructed as follows. First we exclude the households consist of only persons aged 18-24 in full-time education or older then a country specific retirement age. In order to choose an appropriate retirement age we have proceeded as follows. Among the people that have ever worked, we consider the distribution of the ones that are retired (PL030=5) by age and gender. Looking at the ratio of people that at a particular age are retired among all the people in that age, we have been searching for a sudden jump in the distribution. Once this point has been found we have confirmed it looking at the legal age of retirement for a specific country.

Among the remaining households we classify the people as employed or not using variable PL030. We have identified the degree of worklessness of an household constructing a ratio where at the numerator there are all the people in the household for which variable PL030 takes value 1, 2 or 7. The denominator is the sum of the people of the household for which PL030 takes value 1, 2, 3, 6, 7, 9 and 5 and 8 only if the age of the person is less then the retirement age chosen above. So at household level we consider the household worklessness constructing an indicator that ranges from 0 to 1, where 0 identify that all the household is 'jobless'; that is the percentage of jobless people.

To construct the indicator concerning the duration of unemployment, we calculate at household level the ratio:

$$1 - \frac{\sum_{ind=1}^{HH-size} PL080_{ind}}{\sum_{ind=1}^{HH-size} (\mathrm{PL070} + \mathrm{PL072} + \mathrm{PL080} + \mathrm{PL085} + \mathrm{PL087} + \mathrm{PL090})_{ind}} \tag{2.4.6}$$

The variable for general health, PH010, is aggregated as follows. To the categories 1-2-3 is assigned value 1 and to categories 4-5 value 0. Then this variable is aggregated at household level so that an household is considered deprived for that indicator if at least one person in the household is deprived for the item. So the score *s* assumes value 1 if no one in the household is deprived concerning that item, and it assumes value 0 is at least one person is deprived.

The same kind of household aggregation is done for all the personal variables concerning the health and the educational status.

2.4.6. Exploratory and confirmatory factor analysis

Exploratory and confirmatory factor analysis allow us to identify the dimension of deprivation as explained in section 1.3.2. The former gives a preliminary framework of the dimensions. Following exploratory factor analysis nine dimensions are the optimal solution. We then proceeded to rearrange some factors in the dimensions found in order to create more meaningful groups. Finally, we did a confirmatory factor analysis to test the goodness of the model hypothesised. In summary the seven final dimensions are:

1 <u>Basic life-style</u> – these concern the lack of ability to afford most basic requirements:
- Keeping the home (household's principal accommodation) adequately warm.
- Paying for a week's annual holiday away from home.
- Eating meat chicken or fish every second day, if the household wanted to.
- Ability to make ends meet

2 <u>Consumer durables</u> - these concern enforced lack of widely desired possessions ("enforced" means that the lack of possession is because of lack of resources)
- A car or van.
- A colour TV.
- A pc
- A washing machine.
- A telephone.

3 <u>Housing amenities</u> – these concern the absence of basic housing facilities (so basic that one can presume all households would wish to have them):
- A bath or shower.
- An indoor flushing toilet.
- Leaking roof and lamp
- Rooms to dark

4 <u>Financial situation</u> – these concern the lack of ability to pay in time due to financial difficulties:
- Inability to cope with unexpected expenses.
- Arrears on mortgage or rent payments.
- Arrears on utility bills.
- Arrears on hire purchase instalments.

5 Environmental problems – these concern problems with the neighbourhood and the environment:
- Pollution.
- Crime, violence, vandalism.
- Noise.

6 Work and education – these concern the absence of education and job
- Households with early school livers not in education or training.
- Households with persons with low educational attainment.
- Jobless households.
- Intensity or duration of unemployment at household level.

7 Health related – these concern problems with personal health:
- General health.
- Chronic illness.
- Mobility restriction.
- Unmet need for medical examination or treatment.
- Unmet need for dental examination or treatment.

All the indicators of goodness of the model are significant. Below, we report measures of absolute, relative and parsimonious fit as follows:

- The Goodness of Fit Index (GFI) is 0.94. It is based on the ratio of the sum of squared discrepancies to the observed variances; it ranges from 0 to 1 with values above 0.9 indicating a good fit.
- The Adjusted Goodness of Fit Index (AGFI) is 0.93. It is the GFI adjusted for degrees of freedom of the model, that is the number of the fixed parameters. It can be interpreted in the same manner.
- The Parsimonious GFI is 0.86. It adjusts GFI for the number of estimated parameters in the model and the number of data points.
- The Root Mean Square Residual (RMR) is 0.06. The fit is considered really good if RMR is equal or below 0.06. The Root Mean Squared Error of Approximation (RMSEA) is 0.0475. It is based on the analysis of residuals, with small values indicating a good fit. Values below 0.1, 0.05 and 0.01 indicate a good, very good and outstanding fit respectively

2.4.7. Calculations of weights within each dimension

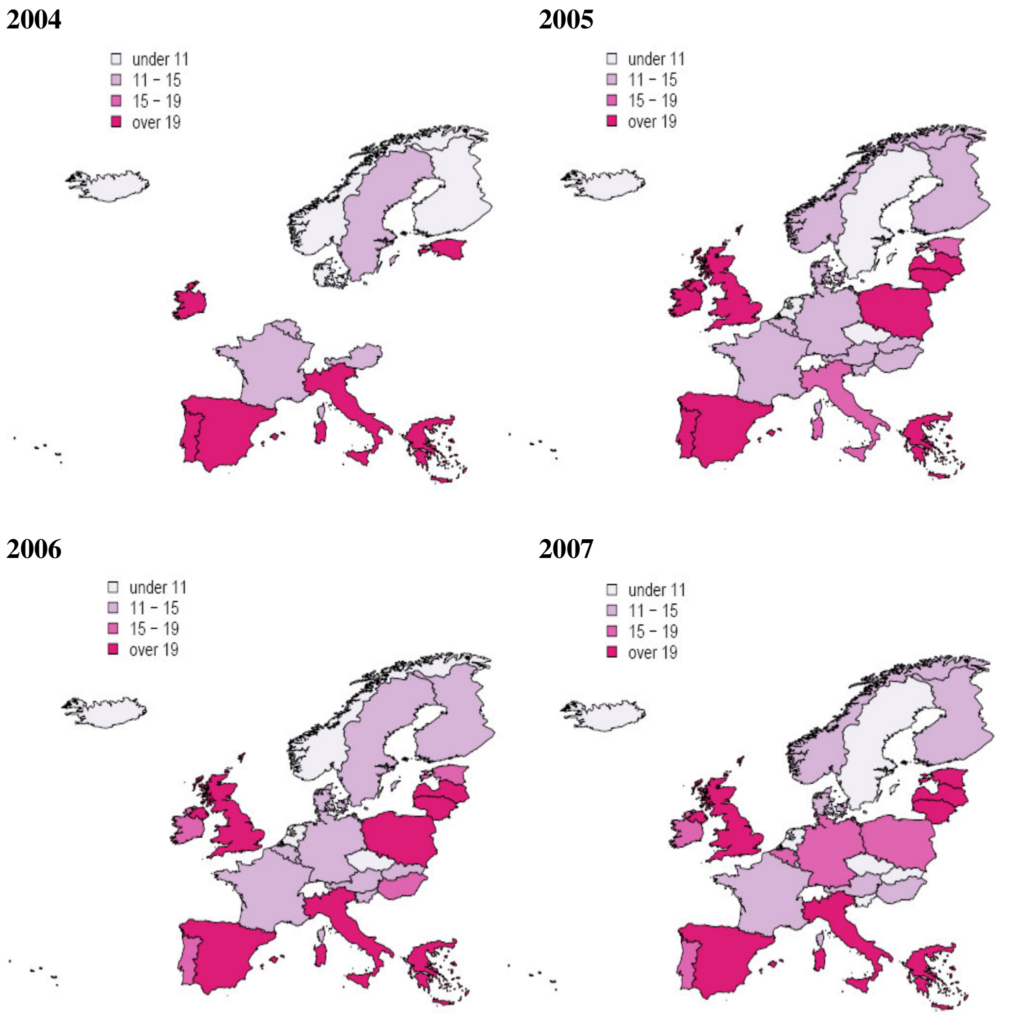
The weights to be given to items are determined within each dimension separately and the set of weights are taken to be item-specific as explained in section 1.3.2. The second factor is given by formula 1.3.14, whereas, since our analysis is carried on using the deprivation scores *s*, instead of the deprivation index *d*, w_{hj}^a should be modified as follows:

$$w_{hj}^a \propto \frac{std_{hj}}{1 - mean_{hj}} \qquad (2.4.7)$$

2.5. Empirical results

Fuzzy measures of monetary poverty and non-monetary deprivation have been constructed, step by step as described in the previous sections, based on EU-SILC survey data. A cross-sectional analysis have been conducted from 2004 to 2007 waves. Figure 2.5.1 shows cartograms of fuzzy monetary indicators (equal to HCR and to the overall non-monetary index) in European Countries. Differences among these years are not so very significant.

Figure 2.5.1. Cartograms of fuzzy monetary indicators in European Countries (2004-2007)



In this section, we described in details only results of year 2007 , but similar conclusions can be done for the other years whose results are reported in appendix A.

Table 2.5.1 shows values of fuzzy monetary indicator and fuzzy supplementary indicator, as overall deprivation index and in term of different dimensions of deprivation for 26 European Countries.

Denmark, Finland, Island, Norway, Sweden, Slovenia and Netherland are register Countries, then as explained in section 2.4.4, they miss health dimension (FS7).

The first column, FS0, is the overall deprivation rate. It is in fact the conventional poverty rate (HCR) for each country. The values of the FM (fuzzy monetary) and FS (fuzzy supplementary) deprivation indices are simply scaled for each country to numerically equal the conventional HCR. Those overall poverty or deprivation rates show large differences among EU countries, from the low value of 9.5% in CZ to the high of 21.2% in LV. In six countries the rate is below 11% (CZ, IS, NL, SK, SE, SI), it exceeds 19% in seven (LV, GR, IT, ES, EE, LT, UK). The average over countries is close to 15%.

We note that there is fairly strong correlation between the ranking of countries according to the overall and dimension-specific indices of deprivation. However, quite large differences in the rankings according to different dimensions are also present. Numerically, deprivation rates for individual dimensions are not scaled in the methodology described above to equal – individually or even in the average over dimensions – the overall poverty or deprivation rate FS0. In fact, over countries, in these data the average of rates for individual dimensions (at 11%) is lower than the average of overall rates (15%).

In certain dimensions, the average over countries is 12-14%, which is quite close to that for the overall index (15%). This group includes:

FS1 – basic life-style
FS5 – environment
FS6 – work and education
FS7 – health related

For the remaining dimensions, the average values obtained are much lower (7-9%). These dimensions are:

FS2 – consumer durables
FS3 – housing amenities
FS4 – financial situation

Table 2.5.1. Fuzzy measures at Country level (2007)

	Rate of deprivation by dimension of deprivation								**mean**
Country	**FS0**	**FS1**	**FS2**	**FS3**	**FS4**	**FS5**	**FS6**	**FS7**	**FS1-FS7**
CZ	0.095	0.092	0.061	0.055	0.045	0.106	0.087	0.085	0.076
IS	0.100	0.087	0.021	0.041	0.084	0.071	0.083		0.065
NL	0.102	0.080	0.040	0.051	0.051	0.097	0.087		0.068
SK	0.105	0.087	0.063	0.059	0.055	0.103	0.094	0.095	0.079
SE	0.107	0.085	0.040	0.058	0.065	0.085	0.089		0.070
SI	0.109	0.094	0.052	0.066	0.075	0.100	0.093		0.080
DK	0.117	0.099	0.057	0.064	0.062	0.100	0.093		0.079
AT	0.120	0.098	0.058	0.070	0.047	0.102	0.105	0.088	0.081
NO	0.123	0.082	0.044	0.058	0.085	0.084	0.100		0.076
HU	0.124	0.127	0.085	0.096	0.083	0.112	0.106	0.140	0.107
FI	0.130	0.097	0.067	0.063	0.075	0.112	0.110		0.087
FR	0.131	0.101	0.058	0.078	0.078	0.126	0.111	0.107	0.094
LU	0.135	0.092	0.028	0.071	0.055	0.119	0.110	0.106	0.083
BE	0.151	0.131	0.071	0.087	0.081	0.141	0.127	0.102	0.105
DE	0.152	0.124	0.058	0.079	0.063	0.145	0.119	0.130	0.103
CY	0.155	0.140	0.058	0.075	0.117	0.146	0.128	0.143	0.115
PL	0.173	0.200	0.105	0.113	0.094	0.135	0.146	0.167	0.137
IE	0.175	0.128	0.083	0.095	0.086	0.133	0.143	0.124	0.113
PT	0.181	0.130	0.115	0.119	0.097	0.158	0.151	0.154	0.132
UK	0.191	0.143	0.060	0.103	0.105	0.162	0.146	0.137	0.122
LT	0.191	0.167	0.124	0.158	0.082	0.143	0.152	0.176	0.143
EE	0.194	0.126	0.114	0.149	0.090	0.183	0.155	0.181	0.143
ES	0.197	0.145	0.073	0.103	0.095	0.172	0.163	0.143	0.128
IT	0.198	0.164	0.064	0.100	0.117	0.192	0.155	0.169	0.137
GR	0.203	0.165	0.109	0.113	0.152	0.169	0.160	0.165	0.148
LV	0.212	0.219	0.136	0.171	0.081	0.224	0.169	0.246	0.178
average	0.149	0.123	0.071	0.088	0.081	0.132	0.122	0.140	0.108

NOTES FS0 stands for "HCR = FM = FS"
FS1 – FS7 refer to the seven dimensions of deprivation defined in section 2.4.6.

We believe that the indices for individual dimensions represent a mixture of relative and absolute levels of deprivation, even if the relative aspect predominates. However, values observed for dimensions 2-4 imply that, compared to overall deprivation and to other dimensions, deprivation in these dimensions may be less severe in the absolute sense in EU countries on the average.

Table 2.5.2 examines the pattern of variation across countries and dimensions more closely, bringing out the relationship in scores across different dimensions in relative terms. The figures shown are ‘normalised’, meaning that we have rescaled them to remove the effect of variations among countries in the overall deprivation (or poverty) rates FS0, and also to remove the effect of differing average values for the various dimensions.

The last column shows the average over the dimensions (FS1-FS7) of the 'normalised' values. This average, by definition, is 1.0 over all countries.

Table 2.5.2. "Normalised" Fuzzy measures at Country level (2007)

	'Normalised rates'								mean
Country	**FS0**	**FS1**	**FS2**	**FS3**	**FS4**	**FS5**	**FS6**	**FS7**	**FS1-FS7**
CZ	1.00	1.17	1.34	0.97	0.86	1.26	1.11	0.95	1.093
IS	1.00	1.05	0.45	0.70	1.53	0.81	1.02		*0.927*
NL	1.00	0.95	0.82	0.84	0.92	1.08	1.04		*0.943*
SK	1.00	1.01	1.26	0.96	0.96	1.11	1.09	0.96	1.050
SE	1.00	0.96	0.79	0.92	1.10	0.90	1.01		*0.946*
SI	1.00	1.04	1.00	1.02	1.26	1.04	1.04		1.068
DK	1.00	1.03	1.02	0.93	0.97	0.97	0.97		*0.980*
AT	1.00	0.98	1.01	0.98	0.72	0.96	1.06	0.78	*0.929*
NO	1.00	0.80	0.75	0.80	1.27	0.78	0.99		*0.900*
HU	1.00	1.24	1.44	1.31	1.22	1.03	1.04	1.20	1.212
FI	1.00	0.90	1.08	0.83	1.05	0.98	1.03		*0.979*
FR	1.00	0.93	0.94	1.00	1.08	1.09	1.02	0.87	*0.988*
LU	1.00	0.82	0.43	0.89	0.74	1.00	0.99	0.84	*0.816*
BE	1.00	1.04	0.98	0.97	0.98	1.05	1.02	0.72	*0.966*
DE	1.00	0.99	0.81	0.88	0.76	1.08	0.95	0.92	*0.912*
CY	1.00	1.09	0.79	0.81	1.37	1.06	1.00	0.98	1.014
PL	1.00	1.39	1.27	1.10	0.99	0.88	1.03	1.03	1.097
IE	1.00	0.89	1.00	0.92	0.90	0.86	0.99	0.76	*0.902*
PT	1.00	0.86	1.33	1.10	0.97	0.99	1.01	0.90	1.024
UK	1.00	0.91	0.66	0.91	1.00	0.96	0.93	0.76	*0.875*
LT	1.00	1.05	1.36	1.39	0.78	0.84	0.97	0.98	1.054
EE	1.00	0.79	1.23	1.30	0.85	1.07	0.97	1.00	1.030
ES	1.00	0.89	0.78	0.88	0.88	0.99	1.00	0.77	*0.884*
IT	1.00	1.00	0.68	0.85	1.08	1.10	0.95	0.91	*0.937*
GR	1.00	0.99	1.13	0.94	1.37	0.94	0.96	0.87	1.028
LV	1.00	1.25	1.35	1.37	0.70	1.20	0.97	1.24	1.154
average	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

NOTES 'Normalised rates' N_{ij} : all values scaled such that:

(1) for each dimension (j), average over countries rescaled to = 1.0; and
(2) for each country (i), FS_j values scaled to correspond to FS0 = 1.0.

$$N_{ij} = \left(\frac{FS_{ij}}{FS_j}\right) \Big/ \left(\frac{FS_{i0}}{FS_0}\right)$$

The overall non-monetary dimension and each of the seven non-monetary dimensions have been combined with the monetary dimension in order to obtain measures of manifest and latent deprivation which correspond respectively to intersection and union of the fuzzy sets. Table 2.5.3 reports values of latent and manifest deprivation for aggregated measures of overall deprivation and the combination of the monetary dimension with each of the seven non-monetary dimensions. The M0/L0 ratio is in general lower in areas with lower levels of deprivation (for example IS and NL), and higher in areas with higher levels (LV and GR). High values of this ratio imply that different types of deprivation overlap and this means that deprivation in the income and non-monetary domains is more likely to afflict the same individuals in the population. On the other hand, low values imply the absence of such overlap at the micro level. Analogously, for each dimension, the overlap between monetary and non-monetary deprivation increases for Countries with higher levels of poverty and deprivation, even if the ranking is not so sharp and there are some exceptions like CK in the second and sixth dimensions and LV in the forth dimension.

Table 2.5.3. Latent and Manifest deprivation at aggregated level and for each dimension of deprivation (2007)

Country	FS	L0	M0	*M0/L0*	L1	M1	*M1/L1*	L2	M2	*M2/L2*	L3	M3	*M3/L3*	L4	M4	*M4/L4*	L5	M5	*M5/L5*	L6	M6	*M6/L6*	L7	M7	*M7/L7*
CZ	0.095	0.154	0.037	0.238	0.153	0.034	*0.224*	0.126	0.029	*0.233*	0.133	0.017	*0.128*	0.124	0.016	*0.132*	0.183	0.018	*0.100*	0.145	0.037	*0.254*	0.166	0.014	*0.084*
IS	0.100	0.171	0.028	0.162	0.162	0.024	*0.151*	0.114	0.007	*0.058*	0.131	0.010	*0.075*	0.163	0.021	*0.127*	0.157	0.014	*0.090*	0.168	0.015	*0.092*			
NL	0.102	0.174	0.030	0.171	0.150	0.031	*0.209*	0.124	0.017	*0.140*	0.141	0.012	*0.085*	0.134	0.020	*0.146*	0.181	0.018	*0.099*	0.162	0.027	*0.169*			
SK	0.105	0.170	0.039	0.231	0.156	0.036	*0.232*	0.140	0.027	*0.195*	0.147	0.017	*0.118*	0.141	0.019	*0.136*	0.190	0.018	*0.094*	0.160	0.039	*0.241*	0.177	0.022	*0.127*
SE	0.107	0.180	0.034	0.188	0.160	0.032	*0.200*	0.131	0.016	*0.122*	0.153	0.013	*0.084*	0.148	0.024	*0.162*	0.176	0.017	*0.095*	0.172	0.024	*0.139*			
SI	0.109	0.178	0.039	0.216	0.167	0.035	*0.212*	0.137	0.023	*0.170*	0.153	0.022	*0.142*	0.161	0.023	*0.143*	0.191	0.018	*0.094*	0.164	0.037	*0.225*			
DK	0.117	0.192	0.041	0.213	0.180	0.036	*0.198*	0.150	0.024	*0.157*	0.165	0.016	*0.097*	0.153	0.025	*0.164*	0.191	0.025	*0.133*	0.185	0.026	*0.139*			
AT	0.120	0.196	0.044	0.227	0.175	0.043	*0.246*	0.153	0.025	*0.165*	0.169	0.021	*0.123*	0.148	0.019	*0.130*	0.200	0.023	*0.113*	0.186	0.039	*0.212*	0.186	0.023	*0.122*
NO	0.123	0.204	0.042	0.204	0.172	0.033	*0.189*	0.146	0.021	*0.140*	0.165	0.016	*0.096*	0.179	0.029	*0.162*	0.188	0.019	*0.102*	0.193	0.031	*0.160*			
HU	0.124	0.196	0.051	0.262	0.205	0.046	*0.223*	0.174	0.035	*0.202*	0.185	0.034	*0.184*	0.174	0.032	*0.183*	0.211	0.025	*0.118*	0.180	0.049	*0.272*	0.233	0.030	*0.131*
FI	0.130	0.212	0.048	0.226	0.182	0.045	*0.246*	0.162	0.034	*0.211*	0.177	0.017	*0.095*	0.173	0.032	*0.184*	0.215	0.027	*0.123*	0.202	0.038	*0.186*			
FR	0.131	0.209	0.054	0.259	0.186	0.046	*0.246*	0.165	0.024	*0.148*	0.184	0.025	*0.134*	0.178	0.031	*0.172*	0.226	0.031	*0.139*	0.194	0.048	*0.246*	0.205	0.033	*0.163*
LU	0.135	0.218	0.053	0.243	0.173	0.054	*0.315*	0.145	0.018	*0.125*	0.183	0.024	*0.131*	0.157	0.033	*0.211*	0.225	0.030	*0.132*	0.198	0.048	*0.243*	0.212	0.029	*0.139*
BE	0.151	0.232	0.071	0.306	0.215	0.067	*0.309*	0.182	0.040	*0.221*	0.208	0.030	*0.145*	0.188	0.044	*0.232*	0.253	0.039	*0.152*	0.220	0.058	*0.262*	0.210	0.042	*0.202*
DE	0.152	0.239	0.064	0.270	0.215	0.061	*0.284*	0.183	0.027	*0.146*	0.203	0.027	*0.134*	0.189	0.026	*0.138*	0.256	0.041	*0.158*	0.219	0.052	*0.238*	0.239	0.043	*0.178*
CY	0.155	0.246	0.065	0.262	0.229	0.066	*0.289*	0.189	0.025	*0.134*	0.203	0.028	*0.136*	0.232	0.040	*0.174*	0.268	0.033	*0.122*	0.232	0.052	*0.224*	0.249	0.049	*0.196*
PL	0.173	0.266	0.081	0.304	0.289	0.084	*0.292*	0.227	0.051	*0.225*	0.238	0.048	*0.202*	0.226	0.041	*0.182*	0.276	0.032	*0.117*	0.254	0.066	*0.259*	0.288	0.052	*0.181*
IE	0.175	0.272	0.078	0.288	0.238	0.065	*0.275*	0.213	0.045	*0.213*	0.236	0.034	*0.144*	0.216	0.045	*0.210*	0.269	0.040	*0.147*	0.250	0.068	*0.273*	0.253	0.046	*0.184*
PT	0.181	0.279	0.084	0.299	0.241	0.070	*0.292*	0.236	0.060	*0.255*	0.250	0.050	*0.202*	0.232	0.046	*0.198*	0.296	0.043	*0.147*	0.272	0.061	*0.223*	0.273	0.062	*0.227*
UK	0.191	0.300	0.082	0.274	0.261	0.074	*0.282*	0.220	0.031	*0.143*	0.256	0.037	*0.146*	0.242	0.054	*0.222*	0.303	0.049	*0.163*	0.267	0.070	*0.261*	0.279	0.049	*0.174*
LT	0.191	0.288	0.095	0.328	0.273	0.085	*0.309*	0.245	0.071	*0.289*	0.277	0.072	*0.260*	0.235	0.039	*0.166*	0.299	0.035	*0.117*	0.269	0.074	*0.275*	0.307	0.060	*0.197*
EE	0.194	0.296	0.091	0.308	0.245	0.075	*0.306*	0.244	0.063	*0.259*	0.276	0.067	*0.244*	0.232	0.052	*0.224*	0.326	0.051	*0.156*	0.283	0.065	*0.231*	0.304	0.071	*0.232*
ES	0.197	0.314	0.081	0.258	0.268	0.074	*0.277*	0.237	0.034	*0.144*	0.262	0.039	*0.149*	0.246	0.046	*0.186*	0.319	0.051	*0.160*	0.288	0.072	*0.249*	0.284	0.056	*0.198*
IT	0.198	0.304	0.093	0.306	0.276	0.086	*0.313*	0.228	0.034	*0.151*	0.257	0.041	*0.161*	0.254	0.062	*0.243*	0.335	0.055	*0.163*	0.273	0.081	*0.296*	0.301	0.066	*0.221*
GR	0.203	0.301	0.104	0.347	0.272	0.096	*0.353*	0.259	0.053	*0.203*	0.269	0.047	*0.175*	0.273	0.081	*0.299*	0.329	0.042	*0.128*	0.285	0.078	*0.273*	0.298	0.070	*0.235*
LV	0.212	0.314	0.110	0.350	0.319	0.112	*0.351*	0.265	0.083	*0.314*	0.308	0.075	*0.245*	0.263	0.030	*0.116*	0.377	0.059	*0.157*	0.299	0.082	*0.274*	0.364	0.094	*0.257*

Chapter 3

Pooled estimates of indicators

The problem of sample size requires a more sophisticated statistical approach than simply using direct estimates for single rounds of sample surveys of moderate size that can be lead to inaccurate estimates.

For this purpose three solutions are possible:

1. making the best use of available sample survey data such as by cumulating and consolidating the data to construct more robust measures which can permit a greater degree of spatial disaggregation;
2. using more sources in combination so as to produce more precise estimates for small domains; using small area estimation (SAE) techniques;
3. exploiting to the maximum "meso" data - such as the highly disaggregated tabulations of New-Cronos - for the purpose of constructing regional indicators

The first aspect is examined in the next sections, whereas chapter 4 deal with the second one.

3.1. Introduction

3.1.1. Pooling and its fundamental objectives

By pooling we mean statistical analysis or the production of estimates on the basis of multiple data sources. The first distinction to be made when we speak about "pooling" is between: (a) the *pooling of data*, i.e. aggregation of micro-level data from the same or different populations, surveys and times, on the one hand, and (b) the *pooling of estimates*, i.e. the production of a common estimate as a function (such as a weighted mean) of estimates produced from individual data sources.

There are three fundamental objectives of pooling of statistical data or estimates.

(1) The first one is cumulation or aggregation in order to obtain *more precise estimates,* albeit normally with some loss of detail. For instance, cumulation and consolidation of data could be one solution which makes the best use of available sample survey data for constructing more robust measures which permit a greater degree of spatial disaggregation.

(2) The second fundamental objective of pooling is to permit *comparisons*, for instance between different populations, between different geographical parts of the given population, or for the "same" population at different times. Comparisons often take the form of estimates of trends or differences in levels across populations or times.
(3) The third fundamental objective is more general and broader. It concerns *common interpretation* of statistical information from different sources and/or for different populations in relation to each other, and possibly also against some common standards.

3.1.2. Prerequisite for pooling: comparability

How different the data sources are from each other is actually a matter of degree: there is no simple dichotomy "same" versus "different". For meaningful pooling whether of micro data or of estimates, it is necessary that the different data sources are "comparable", which is also a matter of degree. The concept of comparability implies the requirement that "data or estimates can be legitimately, i.e. in a statistically valid way, put together (aggregated, pooled), compared (differenced), and interpreted (given meaning) in relation to each other and against some common standard".

It must be emphasises that comparability is absolutely central to the problems and procedures of pooling of data and estimates. In fact, a "sufficient" degree of comparability is a precondition for such pooling to be meaningful (Verma, 2002).

3.1.3. Diverse scenarios

As noted above, different possibilities arise depending on whether the population and the sample involved in the pooling are different or are the same for the different element. At the one extreme, we have the situation where both the population and the sample (or other types of data sources) involved are different: the data or estimates are being pooled across different population, using different sources of data in each. At the other extreme, we have the situation where both the population and the sample are the same or similar.

On this basis, we may distinguish four main types of situations or scenarios. Within each scenario further divisions or subcategories may be identified. Furthermore, as noted above, methodologically we must distinguish between pooling of data and pooling of estimates. The detailed pattern can also differ depending on whether we are dealing with the pooling of microdata or of aggregated estimates. The important point to keep in mind is that the following distinctions are not necessarily sharp or absolute: being the "same" or "different" is a matter of degree.

Table 3.1.1. Different types of situations involved in pooling

	Data source	
Population	Same/Similar (s)	Different /Dissimilar (d)
Same/Similar (S)	S.s	S.d
Different /Dissimilar (D)	D.s	D.d

Scenarios D.s and S.s are the ones most widely encountered in practice, whereas the other two scenarios present more complex technical problems. An example of scenario (D.d) is the Luxembourg Income Study (LIS), which uses different sources in different countries for constructing a data source for comparative research on income distribution. Scenario (S.d) means using different sources to obtain a more complete picture for a given population, such as from income and expenditure surveys. More often, we are dealing with pooling of data from similar sources. Typically scenario (D.s) involves pooling over space (e.g. over countries in a multinational survey), and scenario (S.s) involves pooling over time (e.g. in a periodic survey).

3.2. Scenario D.d. Different population, different data sources

Both the population and data sources differ. This extreme is generally *the most challenging in terms of the requirements of comparability*, as it has already been noted above. Examples of these kind of pooling can be found in Betti *et al.* (2001) and Betti (1998).

3.3. Scenario D.s. Different population, similar or same data source

Data and estimates from similar sources, pooled over different populations. The most common example of such a situation is provided by highly standardised and comparable multi-country surveys, such as the EU-LFS, ECHP and EU-SILC in the European Union.

In practice, it is useful to distinguish between two sub-types within this scenario. This depends on whether the process primarily involves (a) aggregation of different data sets or estimates starting from individual components, or (b) disaggregation or division of a common data set into individual components.

The former typically involves pooling across standardised national sources. The latter presents a much more common – even universal – situation involving partition, for example of a national data set for providing separate estimates for regions, population groups or other sub-national reporting domains. The weighting and estimation procedures involved in "pooling" in the two situations can be quite different.

3.3.1. Examples of category D.s (a): aggregation of data or estimates

3.3.1.1. Pooling of national estimates

Let us consider estimate ϕ_i for a certain statistic for country *i* in EU. In comparisons among countries, obviously, each ϕ_i receives the same weight. However, for estimates aggregated over countries, of the form

$$\phi = \Sigma P_i . \phi_i \qquad (3.3.1)$$

a choice has to be made of the weights P_i. The most common practice by far is to take the

P_i's in proportion to the countries' population size, thus producing statistics for the 'average EU citizen'. However, given the large differences in country sizes, this means that the results are determined mainly by the large countries, and the samples from the smallest ones are mostly wasted. By contrast, it can also be argued that in much policy debate (and in voting for decision making), it is the situation in the 'average EU *country*' that is of interest. This amounts to taking the P_i*'s* as equal. But it can also be argued that both these are rather extreme positions. Countries as well as individual citizens are relevant as units, so that larger countries could be given more weight, but less than proportionate to their population size (Verma, 1999).

Whatever the choice of P_i, (3.3.1) takes the form of pooling macro (country) level estimates. One of the strengths of an inter-country survey such as ECHP is that it provides standardised data sets for all countries. Hence it is possible to take the convenient approach of pooling the national data at the *micro level* for analysis as a single set. This is achieved by appropriately scaling the case weights w_{ij} (for household or person j in country i) as

$$w'_{ij} = w_{ij}.\left(P_i / \Sigma w_{ij}\right) \qquad (3.3.2)$$

For ratios and relationships at the country level, estimates of ϕ_i are not affected by the scaling of the weights, and (3.3.2) gives the same results as obtained using the original weights w_{ij}. For aggregation over countries, (3.3.2) gives results identical to (3.3.1) when ϕ_i is a linear function of unit values v_{ij}. Be more specific, it is necessary to distinguish between different types of estimators involved.

There are four types of estimates to be considered. The first two are:

(1) Aggregates, proportions and means, generally of the form $\phi_i = \Sigma w_{ij}.v_{ij} / \Sigma w_{ij}$;

(2) Ratios and relationships, commonly of the form $\phi_i = \Sigma w_{ij}.v_{ij} / \Sigma w_{ij}.u_{ij}$;

The two forms (3.3.1) and (3.3.2) give identical results as except for the following. The difference between (3.3.2) and (3.3.1) is that between "combined" and "separate" ratio or regression estimates. Normally form (3.3.2), which corresponds to a combined estimate, is preferred because of its smaller potential bias and mean square error.

In the above forms, the contribution of any unit j to the estimate does not depend on the values of other units (k) in the sample. It does for some other statistics, e.g. of the type involved in the study of income distribution inequality and poverty from the ECHP or EU-SILC: the median income, measures of income disparity such as Gini coefficient, poverty rates, etc. Here the useful distinction is whether the dependence is on units only within the country, or on all units in the pooled populations.

(3) Distributional measures defined within countries;

Most commonly, measures of income distribution, disparity, poverty etc. are defined in relative terms, i.e. within each subpopulation (country) separately – for example 'the poor' maybe be defined as persons with income below a certain proportion of the national median income. Obviously, such measures can only be computed separately by country, and then pooled using (3.3.1).

(4) Measures in terms of the common EU-level distribution.

There is also a policy interest in the EU concerning measures of income disparity and poverty which are obtained with reference to the pooled EU-level income distribution, e.g. 'the poor' defined as persons with income below a certain proportion of the EU median income. Such measures are less 'relativistic' in that they depend not only in the income distribution within each country, but also on disparities in the among the countries average income level. Obviously, such measures can be computed only using the combined data using (3.3.2). Of course, once the pooled measure (such as the common EU poverty line) is defined, it may be possible and meaningful to use it to derive and compare other types of statistics by country (such as the 'proportion in poverty').

3.3.1.2. Meta-analysis

Insights gained from meta-analysis can be useful to resolve several issues faced in combining surveys, as survey heterogeneity, planning data collection, and pooling data across surveys (Morton, 1999). Laird and Mosteller (1990) define meta-analysis as "the practice of using statistical methods to combine the outcomes of a series of different experiments or investigations". It implies four steps: identifying all relevant studies; assessing study quality; dealing with study heterogeneity; and summarizing the results.

Kish (1998b), in the context of constructing an average birth rate for a continent using separate country birth rates, proposes three options for pooling multinational samples that are directly comparable to the three main meta-analytic models for combining study effect sizes: fixed effects (equal weight to each country's estimate); equal effects (all subjects are independent and of equal importance); and random effects (weighed averages of the study proportions).

3.3.1.3. Combining separate sites

When similar data are collected in several sites (cities, provinces or districts of one country) of a combined population, but *not in all of the sites*, alternative treatments of them are possible (Kish, 1999a). In combining separate sites three decisions must be made: the allocation of sample sizes, whether the samples should be combined and what weighting to use. These are expressed as follows by Kish.

i. Only separate survey estimates y_t may be presented;
ii. Comparisons between the separate sites require harmonization to render the differences $(y_t - \bar{y}_t)$ meaningful;
iii. Simple cumulations $\bar{y}_t = \sum y_t / \sum n_t$ of all sample cases amount to assuming that the populations N_t of the sites can be considered parts of the same population of $\sum N_t$ elements;
iv. Equal combination $\sum \bar{y}_t / k$ of k sites weight each of the sites equally, disregarding both the sample sizes n_t and the population sizes N_t;

v. Weighted combinations $\bar{y}_w = \sum W_t y_t / \sum W_t$ weight the sites with some measure of their relative importance;

vi. Post-stratification weights imply the construction of pseudo-strata composed of similar sites.

Multinational combinations may be viewed as special cases of multi-site combinations.

3.3.1.4. Multinational designs

Multinational designs arise "not only because of the development of new methods and techniques, but especially because of availability of funds needed for these large enterprises, emerging effective demand for valid international comparisons, and also the improved national statistical and research institutions that are able to implement this complex of coordinated research" (Kish, 1999a).

From a theoretical perspective, combining the provinces of a country is similar to combining the nations of a continent, but from a practical view multinational combinations differ from multi-domain designs for five main reasons (Kish, 1999a): (i) in the former, the centres of decisions reside in separate national offices, and within any nation the agencies for policy setting and for resource allocation may be separate; (ii) technical resources are national and the separate offices may have very different technical development, organizational structures and social connections; (iii) survey variables (education, income, health etc.) depend heavily on national boundaries that vary by culture, religions, economic and educational levels, etc.; (iv) translations of concepts and of questionnaires are daunting challenges that need ingenuity, knowledge and devoted effort; (v) separate samples must be designed and operated to meet distinct national conditions.

Combination of national statistics can occur in six distinct ways and weights (Kish, 1999a): (1) do not combine but publish only separate national statistics; (2) do not combine populations but "harmonize" designs for multinational comparisons in survey measurement methods (Kish, 1994); (3) use equal weights (1/*H*) for every country; (4) weight with sample size (n_h) when elements are drawn essentially from the same population or when per-element variance is the only component of variation; (5) use population weights W_h; (6) use post-stratification weights.

3.3.2 Examples of category D.s (b): disaggregation for separate reporting by domain

3.3.2.1. Multi-domain designs

Statistics and data from national samples commonly provide the basis for separate reporting by sub-national domains.

Kish (1994) defines domains as partitions (non–overlapping, mutually exclusive) of the population, and subclasses as their representation in the sample. He distinguishes *design domains* that designate subpopulations for which separate samples can be planned and selected like regions, provinces and states, from *cross-classes*, meaning domains and subclasses that cut across sample designs, across strata and across sampling unit (classes of age, gender, occupation, income, health, education, etc. Kish,1987).

The diversity of domains may be recognized within national sample designs like provinces that in most countries can number from 5 to 20. In samples of smaller populations like cities or institutions, similar partitions into major domains are typical, but for smaller and more numerous domains deliberate sample designs are not feasible for most samples of limited size; in this situation methods of small area estimation have been developed.

3.4. Scenario S.d. Same population, different data sources

Estimates for a given population, from different data sources. Here as well, it is useful to distinguish two important subtypes. (a) One refers to the situation when the same variables or statistics are being estimated by pooling together multiple sources, such as two sample surveys on the same topic, two different types of surveys but with a common subset of variables (such as household income in income surveys versus income in budget/expenditure surveys), or two sources of different types but providing information on a common set of variables (for example, income from interviews versus from administrative sources). In such situations, the pooling essentially involves aggregation by giving weights to different sources in proportion to their expected degrees of reliability. An example of this category can be found in Di Marco (2006).

(b) The second type of situation involves pooling of substantively different types of data or indicators so as to construct more complex, composite indicators. The different type of data may come from different sources, or from different parts of the same source - they may even refer to the same individual units at the micro level. Typically, the pooling involves the construction of new variables or estimates for a given sample, rather than of the same measures over different samples. A good example is provided by the construction of indicators of multi-dimensional deprivation from indicators of monetary and non-monetary aspects of poverty (see Betti *et al.,* 2006).

3.5. Scenario S.s. Same population, similar data sources

This is the most important scenario in the present context. A number of possible designs and applications are noted in this section. Illustrations from EU surveys and technical aspects of pooling under a rotational design are discussed in more details in the following sections.

A good example of pooling of similar sources for a given population is provided by a periodic survey, repeated frequently at regular intervals using the same methodology and covering essentially the same population. (Of course, the population is not the “same” in the literal sense because it changes over time; but in many practical situations, such as in the context of repeated national surveys, the target population can be considered “essentially” the same).

A number of examples will be given below based on multiple waves of a panel survey such as ECHP or EU-SILC. For instance, poverty rates may be computed for each wave, and then appropriately averaged over time to give more stable measures covering a numbers of years. Poverty rates defined using different thresholds in terms of the mean or median income (e.g. 50%, 60% or 70% of the median) may be averaged for the same purpose. Similarly, poverty analysis may be carried out at different levels of aggregation (e.g. at the level of EU, country, NUTS1, NUTS2,…) and the results pooled in some appropriable manner. Note also that the concept of "pooling" also incorporates putting together of information for the purpose of comparisons such as in the study of time trends or regional differentials.

Another example of this scenario is provided by the "rolling sample" concept promoted by Kish (1990). As described below, here the emphasis is on cumulation of data from independent samples over time in order to improve sampling precision and permit more detailed geographical disaggregation.

3.5.1. Periodic Surveys

Periodic surveys, i.e. repeated surveys over time, have been designed and used mainly for measuring periodic changes, exploiting the advantages of partial overlaps.

They have some common fundamental aspects with combining data from spatial units, but they show also some practical differences (Kish, 1999a): (i) they are designed for the "same" population, which tends to retain some stability between periods; (ii) similar methods and designs are feasible, simpler, and usual over different periods than over different geographical domains; (iii) these stabilities encourage designs with "overlapping" selection of units, in order to reduce unit costs and the variances (from positive correlations); (iv) many periodic surveys employ widely known and used, quite standard, methods.

Kish noted that there now exist several cumulated representative samples of national populations. In order to reduce field costs, they are often restricted within fixed selections of primary samplings units. The Health Household Interview Surveys of the USA consist separate weekly samples of about 1,000 households, cumulated yearly to 52,000 households (National Centre for Health Statistics, 1958). These samples are selected by the US Census Bureau within their large sample of PSUs. The Australian Population Monitors have quarterly non-overlapping samples that are cumulated to yearly samples and these are also confined within fixed primary sampling units (Australian Bureau of Statistics, 1993). The new Labour Force Surveys of the United Kingdom publishes each month the cumulation of three separate non-overlapping monthly samples (Caplan, Haworth and Steele, 1999).

Periodic surveys are the common form of repeated surveys and of longitudinal studies. The periods can be annual, quarterly, monthly or short like daily or less. We can distinguish collection periods from reference periods and from reporting periods, and distinguish panels of individual elements from overlapping sampling units and from non-overlapping or independent selections.

Most periodic surveys use partially overlapping samples with some kind of rotation design in order to reduce variances per sample element and to measure changes between periods and make current estimates.

On the other hand, separate new samples are preferred for cumulations in order to avoid positive correlations.

Panels denote samples in which the same elements (persons, families, households) are measured on two or more occasions for the purpose of obtaining individual changes. From the mean of these individual changes the net population change can be estimated. However, from the net changes of means we cannot estimate (directly) the gross change of individuals. Only panels can reveal the gross changes behind the net changes generally. (Exceptions can be found with strong models; Kish, 1987).

3.5.2. Split panel designs

Another variation, called the Split Panel Design, replaces the overlaps of rotating designs and provides the useful correlations for measuring net changes. Moreover, it serves to measure individual changes. Split Panel Designs (Kish 1981, 1987, 1990, 1998a, 1999a) displace partial overlaps with two samples: a panel *p* added to the independent rolling samples (*a-b-c-d-…).* Thus the periodic samples will consist of *pa-pb-pc-pd* etc. It has two critical advantages over the classical partial overlaps: first, it provides true panels of elements (e.g., persons or households), which are missing for the moving elements in designs of mere overlaps; second, in Split Panel Designs the correlations are present for all periods, not only for the pairs arbitrarily designed in the classical symmetrical rotation designs.

3.5.3. Symmetrical rotations

In many surveys, the pattern of rotation is "symmetrical", that is, new sets of units are introduced into the sample at regular intervals, and once introduced, each set is retained or dropped from the sample following the same pattern (Verma, 1991).

Many surveys use a straightforward pattern of rotation. The sample consists of "n" sub-samples; at the beginning of each survey period, one new sub-sample is introduced; and each sub-samples remains in the survey for n consecutive periods (rounds). The overlap between rounds decreases linearly as the interval separating them increases. For two samples introduced i interval apart the overlap is $(n-i)/n$, up to $i=(n-1)$; after which ($i \geq n$) the overlap becomes zero.

More complicated rotation patterns can be used to vary the degree of sample overlap and how it changes with time.

In many situations, the sample is rotated slowly (or not at all) at higher stage units, and more rapidly as we move to lower stage units. This is done to reduce the cost and inconvenience of changing the primary sampling units and other higher stage units.

3.5.4. Asymmetrical Cumulations

Asymmetrical cumulations are associated with cumulated sample. They denote a strategy of combining several periods for small domains, but reporting large domains frequently, for example, annual reports for small domains, but monthly national reports (Kish, 1997).

Their usefulness is due to three main reasons: (i) the principal divisions of most countries tend to vary greatly in size; (ii) statistics are also wanted for subdivisions of principal divisions; (iii) cumulations are often needed for rare items. However, asymmetrical cumulations can present practical problems of inconsistencies (Kish, 1998a).

3.5.5. Rolling samples and censuses

The 'rolling samples and censuses' methods may be considered as special types of sample cumulations, but they are designed for different and specific functions.

Kish (1998a) define rolling samples as a combined (joint) design of k separate (non-overlapping) periodic samples, each a probability sample with selection fraction $f = 1/F$ of the entire population, designed such that the cumulation of k periods yields a detailed sample of the whole population with $f = k/F$. For example, if we imagine a weekly national sample each designed with same selection rates of $f = 1/520$, the cumulations of 52 such weekly samples would yield an annual sample of 52/520 = 10 percent and then, ten of these annual samples would yield a census of 520/520 (Kish, 1999a).

Rolling samples have been proposed for combining data from periodic surveys into annual data. Data are often collected weekly, or monthly, or quarterly in many countries to provide periodic comparisons, but these same data can also be combined for annual statistics. For efficient cumulation the best designs would be without the overlaps that benefit comparisons, but good compromises are feasible that are nearly optimal for both aims. For both comparisons and for cumulations all survey aspects (variables and populations) must be standardized (Kish, 1999b).

The American Community Survey (ACS) is the largest and best actual national design for rolling samples. Its aim is to bridge the gap in timeliness for the full range of estimates that have traditionally come from the census (in countries such as the USA, from the census "long form"). Starting from 2003 the ACS questionnaire has been mailed to 250,000 addresses each month, spread evenly across the country. A rolling sample is used without overlaps, so that the annual sample is 3 million different addresses and the 5-year cumulated sample is 15 million addresses, compared to about 17 million for the 1990 census long form sample (Alexander, 1999). It is expected that in the next US census, only the short form on a full coverage basis will be used; the traditional long form will be entirely replaced by the rolling ACS sample.

3.5.6. More robust poverty measures

Finally, we consider cumulation specifically for constructing more reliable or stable poverty measures.

3.5.6.1. Poverty rates cumulated over time

Where the information comes from sample surveys of limited size, a trade-off is required between temporal detail and geographical breakdown. In order to achieve greater geographical disaggregation (e.g. by region), the emphasis has to be shifted away from the study of trends over time and longitudinal measures to essentially cross-sectional measures aggregated over suitable time periods, so as to illuminate the more stable aspects of the patterns of variation across geographical areas. Simple average of wave-specific poverty rates over waves provides an indicator reflecting the overall situation over the period covered. Such measures constructed from averaging over waves tend to be more robust than results based only on one wave. They increase precision, that is the effective sample size, help to smooth out short-term fluctuations and bring out more clearly the underlying patterns and relationships.

3.5.6.2. Poverty rates at different thresholds

In the standard analysis, poverty line is defined as a certain percentage (x%) of the median income of the national population; by *poverty line threshold* we mean the choice of different values of x. The three more commonly chosen thresholds are 50%, 60% and 70% of the median.

Irregularities in the empirical income distribution can arise especially in smaller samples. Computing poverty rates using different thresholds and then taking their weighted average using some appropriate pre-specified weights can reduce such irregularities and increase sampling precision.

Lower thresholds isolate the more severely poor and tend to be more sensitive in distinguishing countries or other population groups being compared in terms of the extent of extreme poverty. This sensitivity tends to fall as the threshold is raised.

3.5.6.3. Poverty rates with poverty lines at different levels

The level of a poverty line refers to the population level at which the income distribution is pooled for the purpose of defining the poverty line. Commonly used poverty-related indicators, such as in the Laeken list, are based on country poverty lines; that is, the poverty line used in these indicators is always determined on the basis of the national income distribution. The common procedure is to consider the income distribution separately at the level of each country, and pool the numbers poor over countries to obtain the overall EU poverty rate, but a rate still defined in terms of *national* poverty lines. Similarly, the numbers poor defined according to the *national* poverty line in each country can be disaggregated by region, obtaining regional poverty rates, but still in terms of *national* poverty lines.

It is necessary to consider other levels of the poverty line, especially for the construction of poverty rates at the regional level. Some examples are: EU poverty line determined on the basis of pooled income distribution for all EU countries; country-level poverty lines determined on the basis of pooled income distribution separately within each country; NUTS1 level poverty lines determined on the basis of pooled income distribution separately within each NUTS1 region, and NUTS2 level poverty lines determined on the basis of pooled income distribution separately within each NUTS2 region in each country and so on.

Hence, for deeper analysis it is useful to consider poverty lines defined at different levels, such as using a common EU-level poverty line for identifying the poor in each EU country. Different levels for the poverty line can also imply a different mix of relative measures (those concerning purely the distribution of income) and absolute measures (those involving the mean income levels as well). We can mix any level of analysis of aggregation, concerning the units for which the measures are computed, with any poverty line level that refers to the population of which the income distribution has been considered in defining the poverty line. The poverty line level chosen can make a major difference to the resulting poverty rates, in particular when that level (e.g. national) is higher than the level of analysis or aggregation (e.g. regional).

It is important to note that while consolidation over waves and poverty line thresholds (discussed in 3.5.6.1 and 3.5.6.2 above) increases sampling precision of the estimates, such consolidation or averaging is not meaningful over poverty line levels because different poverty line levels capture different aspects of the situation – varying from absolute to purely relative aspects - and help to separate out within and between regional variation. It is best, therefore, to keep them separate, each regarded as defining a different indicator of poverty (Verma *et al*., 2005).

3.6. Example of cumulation in European surveys

In this section we provide two detailed examples of scenario (S.s) from European social surveys: namely from the continuous or annual Household Budget Surveys, and the rotational design of EU-SILC.

3.6.1. Cumulation of data and measures in a continuous Household Budget Survey

Some individual Member States of the EU, have gradually moved towards annual household budget surveys, in place of surveys conducted once every few years. There are many advantages of continuous surveys. However, often it is not feasible to have large enough sample sizes for reporting the results by single years, even if in principle this can be done with a continuous survey. For instance, in Denmark (following from Norwegian experience) a new model of the Danish HBS was introduced a number of years ago. The idea consists of a survey of modest size conducted on a continuous basis, data from which can be cumulated over years to achieve more adequate sample sizes.

How can data and measures be cumulated and averaged over time to construct more reliable measures? The following methodology, based on Verma (2001c) attempts to provide a response of this question.

Suppose that in place of conducting one survey of, say, 5,000 households every five years, the survey is conducted on a continuous basis with a representative sample of 1,000 households every year. During the year the workload is also distributed more or less uniformly, e.g. enumerating around 80-100 households per month. The work can be conducted by a small team of interviewers (e.g. 8 or so) deployed permanently for the task. With a continuous flow of data from the field, data preparation and processing also becomes an ongoing operation. The sample is designed such that the information can be efficiently cumulated over time to achieve sufficient sample sizes, and the results are reported on a regular (annual) basis in the form of ‘moving averages’ over a number of most recent years.

Two main advantages of the model may be emphasized.

a. The relatively moderate but regular workload;
b. The regular updating of the results in a more timely manner.

Lack of flexibility can be a possible disadvantage of a continuous survey. Major redesigns are more easily accommodated in ad hoc surveys separated by long intervals. By contrast, in a continuous survey it is necessary to carefully regulate and control changes in content, design and procedures.

To cumulate the survey results over time it is necessary that the following hold.

i. The sample be representative simultaneously over space and time. This means that for annual surveys, for instance, the sample for each year separately should be representative of the whole country. Actually, it is desirable to divide the year into shorter (such as half-monthly, monthly, or at least quarterly) periods, each with a separately representative sample of the country;
ii. The annual samples should be independently selected, so as to avoid positive covariance and permit efficient cumulation over years. If a multi-stage sampling design is used, the samples for different periods should ideally use different, independently selected primary sampling units;
iii. The sample sizes should be equal or at least fairly similar from one period (year) to the next, even if some variation in the achieved sample sizes from year to year cannot be avoided in practice.

On the basis of these considerations, a good estimation procedure appears to be as follows. Weight each annual sample to be representative of the mid-year population of the year concerned (taking into account selection probabilities, response rates, external control totals etc.), and then put together the annual sample estimates with weights in proportion to their corresponding mid-year populations to produce cumulative results. In so far as the population does not change much over a few years, the above implies giving equal weights to the annual estimates in putting together the results.

For the following illustration of the details, we will assume that data are collected with the sample uniformly distributed over Y years and, for a particular set of items, with a moving reference period of X years preceding the survey interview. For instance, for major expenditures (such as purchase of motor vehicles) the reference period may be X = 1 year preceding the survey; for items such as clothing, we may have a reference period of six months (X = 0.5 years); while for items recorded on a continuous basis in a diary, we have effectively X = 0; and so on depending on the survey questionnaire. For a single-year survey we have Y = 1, while with data cumulated over three years in a continuous survey we have Y = 3. For the sample as a whole, cumulated over collection period Y with data collected with a moving reference period of X years, the resulting data would pertain to the time period

$$P = X + Y \qquad (3.6.1)$$

years preceding the last interview. The quantum (volume) of the information collected is distributed symmetrically, centred at the point

$$P_O = (X + Y)/2 - 1 \qquad (3.6.2)$$

years before the beginning of the most recent survey year, or

$$P_M = (X + Y)/2 - 0.5 \qquad (3.6.3)$$

before the mid-point of the most recent survey year.

It can be seen that with different values of the reference period for different types of items, the situation with cumulation over a number of years (Y > 1) is similar in form to that with a conventional survey conducted over a single year (Y = 1). Actually, with increasing Y, the period covered becomes less sensitive to differences in the reference periods for different types of items in the same survey. This greater uniformity of the time-periods covered for different types of items is in fact an advantage of increasing Y (the period of cumulation).

For a fixed reference period such as the preceding calendar year, the situation is similar but simpler. For a conventional one-year survey, the period covered is centred at the middle of the reference calendar year. We have for Y = 1:

$$P_M = 1; P_O = 1/2 \qquad (3.6.4)$$

More generally, with cumulation over Y > 1 years we have:

$$P_M = (P + 1)/2; P_O = P/2, \qquad (3.6.5)$$

giving, for instance,

P_M = mid-point of the second reference year when cumulated over Y = 3, and

P_M = end of the second reference year when cumulated over Y = 4 years.

In any HBS, prices (whether actual or imputed) for all items of consumption or expenditure need to be adjusted in accordance with the periods they refer to. At the individual level, the relevant price is the one prevailing at the mid-point of the reference period for the item concerned.

Exactly the same procedure as that for a single-year survey applies to any continuous survey involving cumulation over a number of years.

In adjusting prices, it is important to note that a *single, common adjustment factor* - reflecting the overall consumer price index for private households - applies to all types of items and all categories of households. (Using different adjustment factors for different categories will fail to reflect changes in the structure of consumption in terms of values.) This fact considerably simplifies the adjustment process. On the other hand, if certain items of consumption such as imputed rent are obtained from an external source and refer to a different period than the reference period of the survey, price adjustments to bring them in line with the survey period have to be made using appropriate quantity, price and quality indicators specific to the item concerned (Verma, 2001c).

3.6.2. Cumulation of cross-sectional and longitudinal data from EU-SILC

The following describes the EU-SILC rotational panel design, procedures for cumulating longitudinal data, and how the relative sample sizes of the cross-sectional and longitudinal components may be modified to affect this cumulation.

Full details of the following procedures are available in Verma (2001a, 2001b); see also Verma and Betti (2006).

EU-SILC rotational panel design

Consider two successive years with partially overlapping samples. For the cross-sectional sample for each year to be separately representative requires each of the following three parts to be a representative sample: (i) the dropped part to be representative of the population at year 1; (ii) the added part to be representative of the population at year 2; and (iii) the overlapping part to be representative of the population at both times.

Normally, the above is achieved in practice by selecting the total sample in the form of a number of replications. The scheme is illustrated in *Figure 3.6.1*. Each replication is in itself a representative sample, typically with the same design (structure, stratification, allocation, etc.) as the full sample, differing from the latter only in sample size. From one year to the next, some of the replications are retained, while others are dropped and replaced by new replications depending on the extent of the overlap desired.

Figure 3.6.1.

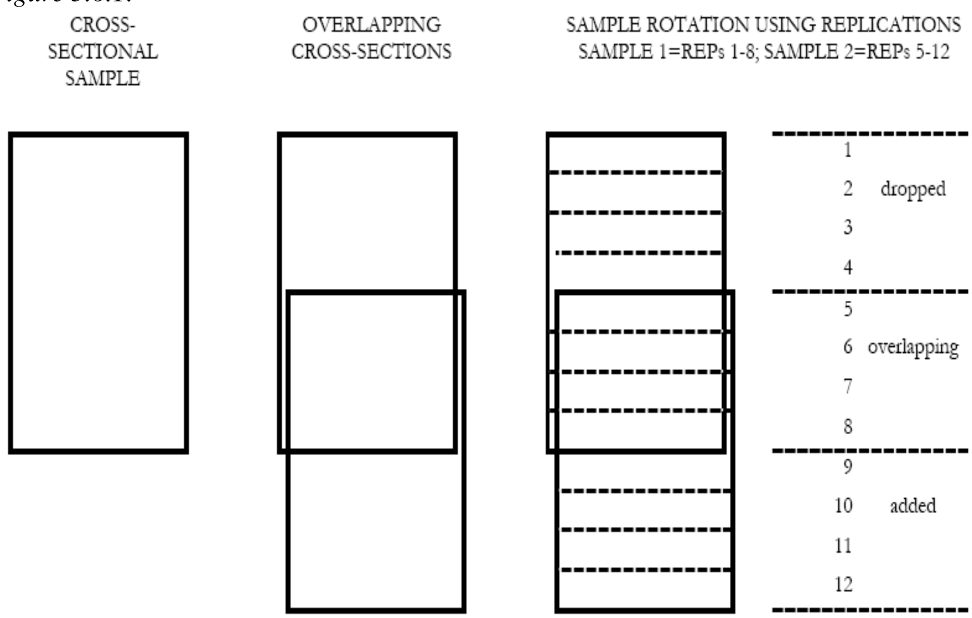


Figure 3.6.2 illustrates a simple rotational design (once the system is fully established). The sample for any one year consists of 4 replications, which have been in the survey for 1-4 years (as shown for 'Time=T' in the figure). Any particular replication remains in the survey for 4 years; each year one of the 4 replications from the previous year is dropped and a new one added, giving a 75% overlap from one year to the next. For surveys two years part, the overlap is 50%; it is reduced to 25% for surveys three years apart, and to zero for longer intervals. With *n* replications, each kept in the survey for *n* rounds, the overlap between

rounds declines linearly as the interval separating them increases. For two surveys i intervals apart the overlap is $(n - i)/n$, up to the time $i = (n - 1)$, after which $(i \geq n)$ the overlap becomes zero.

Figure 3.6.2.

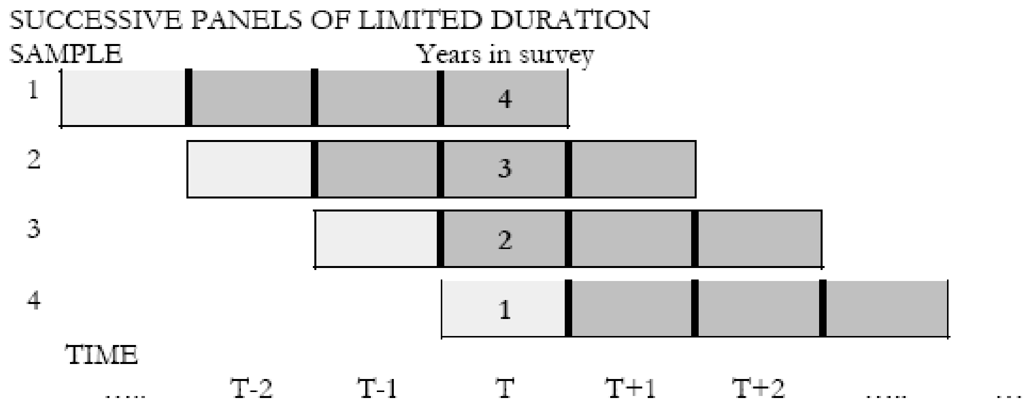
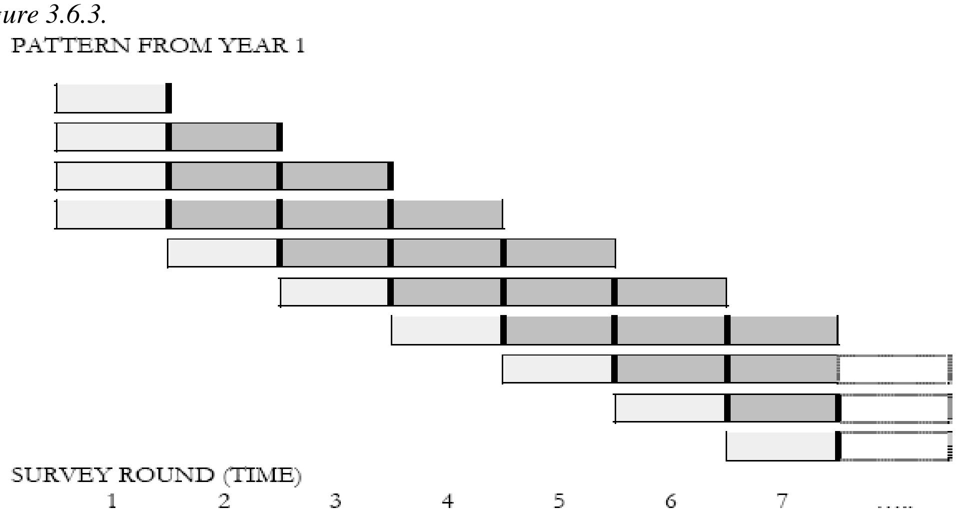


Figure 3.6.3 illustrates how a rotation pattern may be started from year 1. To obtain the full sample with 4 replications for the first year, it is necessary to begin with all the 4 replications. These replications are treated differently over time. One of these is dropped immediately after the first year, the second is retained for only 2 years, the third for 3 years, and only the fourth is retained for the full 4 years. The pattern becomes 'normal' from year 2 onwards: each year one new replication is introduced and retained for 4 years.

Figure 3.6.3.



Cumulation of longitudinal data

The main limitation of longitudinal sample is the smallness of the sample size available for studying special subgroups in the population: cumulation of data over time may be one simple method of increasing the available sample sizes.

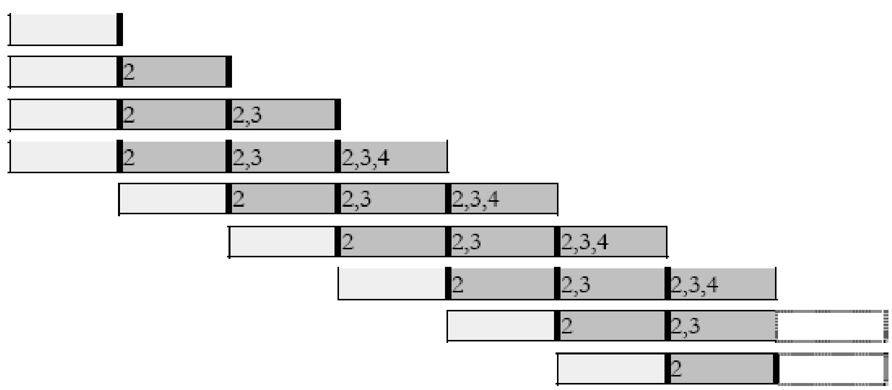
First consider year-to-year transitions in the design of *Figures 3.6.2-3.6.3*, with r subsamples for instance.

Each year starting with year 2, (r - 1) subsamples provide observations of year-to-year transitions. These can be cumulated over time to obtain (r - 1)*(y - 1) subsamples proving observations of year-to-year transitions over the years 1 to y. The resulting analysis provides an average picture of such transitions over the y years.

Figure 3.6.4 (based on *Figure 3.6.3*) provides an illustration with $r = 4$ which is by far the most common design used in EU-SILC. Each year starting with year 3, (r - 2) subsamples provide a set of longitudinal observations, each covering a three year period. These can be cumulated over time up to survey year y to obtain (r - 2)*(y - 2) subsamples proving observations, each covering 3 years. The resulting analysis provides an average picture of such observations over the y years.

Figure 3.6.4.

CUMULATION OF LONGITUDINAL OBSERVATIONS



CUMULATIVE NUMBER OF LONGITUDINAL OBSERVATIONS

1. SUBSAMPLES PROVIDING CROSS-SECTIONAL DATA								
4	8	12	16	20	24	28	…..	4*Y
2. SUBSAMPLES PROVIDING YEAR-TO-YEAR TRANSITIONS								
0	3	6	9	12	15	18	…..	3*(Y-1)
3. SUBSAMPLES PROVIDING 3 YEARS LONGITUDINAL OBSERVATIONS								
0	0	2	4	6	8	10	…..	2*(Y-2)
4. SUBSAMPLES PROVIDING 4 YEARS LONGITUDINAL OBSERVATIONS								
0	0	0	1	2	3	4	…..	(Y-3)
SURVEY YEAR								
1	2	3	4	5	6	7	…..	Y

Note: The numbers in the cells of the diagram indicate the type(s) of observations provided by the subsample.
The above numbers may be multiplied by the subsample size to obtain the cumulated number of observations.

As for sets of longitudinal observations each covering a 4-year period, each year starting with year 4 provides (r - 3) subsamples for the purpose. These can be cumulated over time up to survey year y to obtain (r - 3)*(y - 4) subsamples proving observations each covering 4 years.

<u>Adjusting the relative sample sizes of the two components</u>

The relative size of the panel component can be increased (reduced) only by increasing (reducing) its duration (r), but that duration is not a parameter which can be chosen merely on the basis of sampling considerations. More flexibility can be achieve by supplementing the basic structure by the *split panel*, i.e. the addition to the basic structure of a panel component of unlimited duration; by contrast, the size of the cross-sectional component can be increased by adding to the basic structure a fully rotational *cross-sectional booster*.

While this option has not be so far used in EU-SILC national survey, it remains potentially useful and interesting for regional and other special EU-SILC surveys.

For example, consider a rotational design with r replications or subsamples, each of size s. In the basic model, each subsample is retained in the survey for r years.
In any round:

i. the cross-sectional sample is of size $n_1=r*s$;
ii. the longitudinal sample linked over two years is of size $n_2 = (r - 1)*s$ (since all but the newly introduced panel provide such linkage with the previous year);
iii. the longitudinal sample linked over three years is of size $n_3 = (r - 2)*s$ (since all but the two most recently introduced panels provide such linkage with year y - 2);
iv. that linked over four years is of size $n_4 = (r - 3)*s$; and so on.

With the addition of a split panel of size p, each of the above is essentially increased by p, so that the longitudinal to cross-sectional sample size ratio, such as n_{i+1}/n_1 is increased from $\frac{n_{i+1}}{n_1} = \frac{r-i}{r}$ to $\frac{n_{i+1}}{n_1} = \frac{r-i+(p/s)}{r+(p/s)}$.

With the addition of a cross-sectional booster of size x, the available cross-sectional sample is increased by x without affecting the longitudinal components. The longitudinal to cross-sectional sample size ratio is therefore reduced from $\frac{n_{i+1}}{n_1} = \frac{r-i}{r}$ to $\frac{n_{i+1}}{n_1} = \frac{r-i}{r+(x/s)}$.

3.7. Effects of pooling on variance

We use EU-SILC standard sample structure to illustrate the effect of aggregation on resulting variance. In this design, each cross-section consists of four panels or subsamples, introduced one by one over the preceding years.

3.7.1. Reduction in variance by pooling data for subsamples

In aggregating over subsamples, variance decreases in inverse proportion to sample size, provided that the subsamples making up the total sample are independent. This is the case with EU-SILC samples where each subsample is based on a different set of clusters. There are also a number of designs in which different subsamples involve different households but all from a common set of clusters. Here the design effects tends to increase as the subsamples are pooled, so that the gain in precision is smaller than proportionate to sample size.

3.7.2. Reduction of variance from averaging different poverty thresholds

As noted, some gain in sampling precision can be obtained by computing poverty rates using different thresholds, and then taking their weighted average using some appropriate pre-specified (i.e., constant or external) weights. A quantitative indication of the magnitude of this gain may be obtained on the following lines. Consider three poverty line thresholds, with poverty rates p_i, $p_1 < p_2 < p_3$, and that with fixed weights W_i, $\sum W_i = 1$, a consolidated rate is computed as $p = \sum W_i p_i$. For simplicity, take the sample as SRS and approximate the complex statistic 'poverty rate' as an ordinary proportion. In case, since the design effects due to departures from SRS are likely to be very similar for the various statistics being considers, neglecting them should not substantially affect the conclusions. Under the above assumptions, variance of the consolidate poverty rate *p* is given by:

$$\text{var}(p) = \Sigma_i W_i^2 \text{ var}(p_i) + 2\Sigma_{j<i} W_i W_j \text{ cov}(p_i, p_j) \qquad (3.7.1)$$

By considering the poverty indicator variables $p_{i,k} = \{0,1\}$ for individuals *j* in the population, it can be easily seen that the above equation becomes:

$$\text{var}(p) = \Sigma_i W_i^2 p_i (1 - p_i) + 2\Sigma_{j<i} W_i W_j p_j (1 - p_i) \qquad (3.7.2)$$

Compared to variance of a rate (say, p_2) computed using a single poverty line such as 60% of the median, with $\text{var}(p_2) = p_2(1 - p_2)$, the ratio $g_V = (\text{var}(p)/\text{var}(p_2))^{1/2}$ gives the required factor by which the standard error is reduced.
The 'constant' weights may come from poverty rates estimated at the country level, and then the same weights applied to each region. An appropriate choice is (Verma *et al*. 2005):

$$W_1 = \frac{1}{3} \cdot \left(\frac{p_2}{p_1}\right), \quad W_2 = \frac{1}{3}, \quad W_3 = \frac{1}{3} \cdot \left(\frac{p_2}{p_3}\right) \qquad (3.7.3)$$

where subscripts 1, 2 and 3 refer to the rates computed at the national level with poverty line thresholds, respectively, as 50, 60 and 70% of the national median equivalised income.

3.7.3. Reduction due to aggregation over waves for a given panel (subsample)

Of course, we cannot merely add up the sample seizes over waves in a panel survey since there is a high positive correlation between waves which reduces the gain from cumulation. Consider two adjacent waves, with proportion poor as p and p', respectively, with the following individual-level overlaps between the two waves:

	Wave w+1		
Wave w	Poor (p'_i=1)	Non-poor (p'_i=0)	total
Poor (p_i=1)	a	b	p=a+b
Non-poor (p_i=0)	c	d	1-p=c+d
total	p'=a+c	1-p'=b+d	1=a+b+c+d

Indicating by p_j and $p_j^{'}$ the {1,0} indicators of poverty of individual *j* over the two waves, we have, with the sum over all (*g*) individuals:

$$\begin{aligned} \text{var}(p_j) &= \Sigma(p_j - p)^2/g = p(1-p) = v_1 \\ \text{cov}(p_j, p'_j) &= \Sigma(p_j - p)(p'_j - p')/g = a - pp' = c_1 \end{aligned} \qquad (3.7.4)$$

For data averaged over two adjacent years (and ignoring the difference between *p* and $p^{'}$), variance is given by: $v_2 = \frac{1}{4}(v_1 + v_1 + 2c_1) = \frac{v_1}{2}\left(1 + \frac{c_1}{v_1}\right)$. The correlation $(c_1/v_1) = R_1$ between two periods is expected to decline as the two become more widely separated. Let $(c_i/v_1) = R_i$ be the correlation between two points *i* waves apart. A simple and reasonable model of the attenuation with increasing *i* is: $(c_i/v_1) = (c_1/v_1)^i$. Now in a set of *Q* periods (waves) there are (*Q* - *i*) pairs exactly *i* periods apart, *i* = 1 to (*Q* - 1). It follows from the above that variance v_Q of an average over *Q* periods relates to variance v_1 of the estimate from a single wave as:

$$f_c^2 = \left(\frac{v_Q}{v_1}\right) = \frac{1}{Q}\cdot\left(1 + 2\sum_{i=1}^{Q-1}\left(\frac{Q-i}{Q}\right)\cdot\left(\frac{c_1}{v_1}\right)\right)^i, \text{ with } \left(\frac{c_1}{v_1}\right) \approx a - p^2 \quad (3.7.5)$$

where *a* is the overall rate of persistent poverty between pairs of adjacent waves (averaged over *Q*-1 pairs), and *p* is the (cross-sectional) poverty rate averaged over *Q* waves. Averaging over *Q* waves increases the effective sample size by $(1/f_c^2)$.

3.7.4. Reduction from averaging over rounds in a rotational design

Consider a rotational sample in which each unit stays in the sample for *n* consecutive periods, with the required estimate being the average over *Q* consecutive periods, such as

$Q = 4$ quarters for annual averages. The case $n = 1$ corresponds simply to independent samples each quarter and, under the simplifying assumptions of uniform variances, variance of the estimate of average over Q period is:

$$V_a^2 = \frac{V^2}{Q} \tag{3.7.6}$$

In the general case, the total sample involved in the estimation consists of ($n + Q$ - 1) independent subsamples. These correspond to the rows in the figures below. Each subsample is 'observed' over a certain number of consecutive periods within the interval (Q) of interest. In principle, for a given subsample the sample cases involved in these 'observations' are fully overlapping. The distribution of the ($n + Q$ - 1) subsamples according to the number of observation (m) provided is:

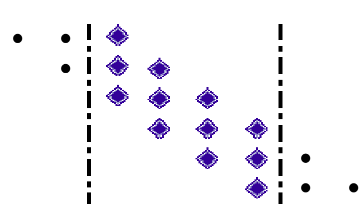
N° of observations (m) ⟶	provided by n° (x) of subsamples	Total n° of 'observations' provided by all subsamples
$m = 1,2,\ldots,(m_1 - 1)$	$x = 2$ for each value of m	$\sum_{i=1}^{(m_1-1)} 2i = (m_1 - 1)\cdot m_1$
$m = m_1$	$x = m_2 - (m_1 - 1)$	$m_1 \cdot m_2 - (m_1 - 1)\cdot m_1$
Total	n° of subsamples equal to $2\cdot(m_1 - 1) + m_2 - (m_1 - 1) =$ $= m_2 + m_1 - 1 = n + Q - 1$	n° of observations equal to $m_1 \cdot m_2 = n \cdot Q$

Here $m_1 = \min(n, Q)$ and $m_2 = \max(n, Q)$.

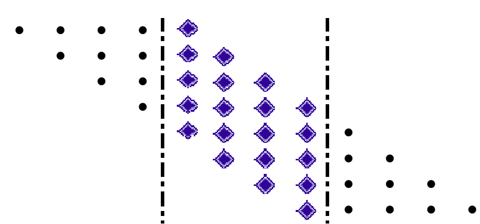
Note that the total number of 'observations' provided by all subsamples over interval Q is $m_1 \cdot m_2 = n \cdot Q$. This is consistent with the fact that, obviously, there are n subsamples observed at each of the Q periods in the interval being considered (see diagrams below).

Q=4

***n*=3** *('observations' provided=3*4=12)*



***n*=5** *('observations' provided=5*4=20)*

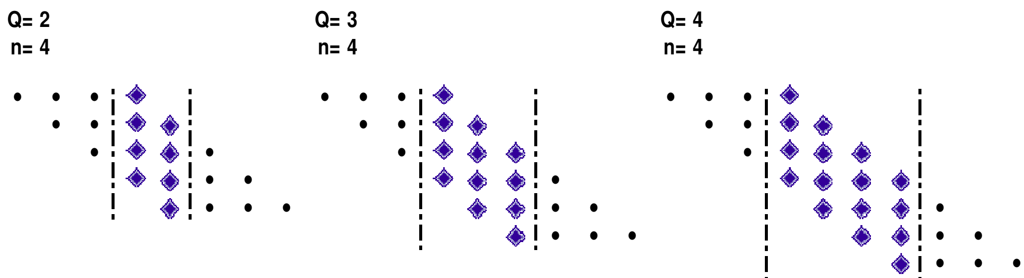


Note: The numbers on the left side of the figures represent the number of subsamples (n+Q-1).

For illustration, consider $Q = m_1 = 4$, $n = m_2 = 6$. There are 2 contributing subsamples for each number 1, 2 and $(m_1 - 1) = 3$ of observations; and in addition there are subsamples $m_2 - (m_1 - 1) = 2$, each contributing $m_1 = 4$ observations.

Similarly, for $Q = m_2 = 4, n = m_1 = 3$, we have 2 contributing subsamples for each number 1 and $(m_1 - 1) = 2$ of observations, and in addition $m_2 - (m_1 - 1) = 2$ subsamples each contributing $m_1 = 3$ observations.

In the EU-SILC survey in most countries, n is always equal to 4 (each survey rounds is made of 4 subsamples), and at the present stage Q could be equal to 2 (years 2003-2004), 3 (years 2003-2004-2005) and 4 (years 2003-2004-2005-2006).
So the previous figure could be adapted as follow:



In order to provide a simplified formulation of the effect of correlation arising from sample overlaps, we assume the following model. If R is the average correlation between estimates from overlapping samples in adjacent periods, then between points one period apart (e.g. between the 1st and 3rd quarters), the average correlations is reduced to R^2, the correlation between points two periods apart (e.g. the 1st and the 4th quarters) is reduced to R^3, and so on.

Consider a subsample contributing m observations during the interval (Q) of interest with full sample overlap. Considering all the pairs of observations involved and the correlations between them under the method assumed above, variance of the average over the m observations is given by

$$V_m^2 = \frac{V^2}{m} \cdot (1 + f(m)) \qquad (3.7.7)$$

where

$$f(m) = \frac{2}{m} \cdot \left\{ (m-1) \cdot R + (m-2) \cdot R^2 + ... + R^{m-1} \right\} \qquad (3.7.8)$$

The term $V_m^2 \Big/ \left(\frac{V^2}{m} \right) = 1 + f(m)$ reflects the loss in efficiency in cumulation or averaging over overlapping samples. The following illustrates its values for various values of m:

m	f(m)
2	R
3	$\frac{2}{3}(2R+R^2)$
4	$\frac{2}{4}(3R+2R^2+R^3)$
5	$\frac{2}{5}(4R+3R^2+2R^3+R)$

Repeated observations over the same sample are less efficient in the presence of positive correlations (*R*). The loss depends on the number of repetitions *m* and is summarised by the factor $(1+f(m))$.

In estimating the average using the whole available sample of $(n \cdot Q)$ subsample observations[3], we may simply give each observation the same weight. Taking into account the number of observations and the variances involved, the resulting variance of the average becomes:

$$V_a^2 = \left(\frac{V^2}{n \cdot Q}\right) \cdot \left\{ m_1 \cdot \left[m_2 - (m_1 - 1)\right] \cdot \left[1 + f(m_1)\right] + 2\sum_{m=1}^{m_1-1} m \cdot \left[1 + f(m)\right] \right\} \Big/ (n \cdot Q) = \left(\frac{V^2}{n \cdot Q}\right) \cdot F(R) \qquad (3.7.9)$$

The first factor is the variance to be expected from $(n \cdot Q)$ independent observations (with no sample overlaps or correlation), each observation with variance V^2. The other terms are the effect of correlation with sample overlaps. Thus effect, *F(R)* disappears when $f(i) = 0$ for all $i = 1$ to *m*, as can be verified in the above expression.

An alternative is to take a weighted average of the observations, with weights inversely proportional to their variance, i.e. to the corresponding factor $(1+f(m))$. The effect on the resulting variance, though may appear algebraically cumbersome, can be easily worked out, for any given rotation pattern and value of average correlation *R*.

It has the form

$$V_a^2 = \sum W_i^2 \cdot V_i^2 , \quad \text{with} \quad \sum W_i = 1 \qquad (3.7.10)$$

where W_i are the relative weights given to observations in a set involving *i* repetitions during the interval of interest.

[3] Obviously , we have n subsamples observed during each of *Q* periods in the rotational design assumed.

3.8. Concluding remark

The objectives of pooling include searching for measures which convey essentially the same information but in a *more robust* manner, reducing random variability or noise. A related objective of pooling is *trading dimensions* – gaining in some more needed directions by losing something less needed for particular purposes – such as permitting more detailed geographical breakdown but with less temporal detail. A third objective is to *summarise* over different dimensions, providing more consolidated and fewer indicators.

As noted above, reducing the variability is one of the objectives of pooling. However, if after pooling the variance is still high, *small area estimates* can be applied as possible solution. This method will be illustrated in details in the next chapter.

Chapter 4

Small area estimation

4.1. Introduction

In recent years, small area estimation methods have token a large importance due to the growing demand in public and private sectors of reliable statistics for small areas that cannot be satisfied by census, sample surveys or administrative data separately. For these reasons, sophisticated small area estimation methods are developed. They link the different sources of information in order to obtain more accurate estimators for small geographic areas or small domains of interest.

But what "small area" means? In literature several definitions can be found (Purcell and Kish, 1980; Brackstone, 1987; Rao, 1994); in what follows a domain is regarded as "small" if the domain-specific sample is not large enough to support direct estimates of adequate precision; they are likely to yield large standard errors due to the unduly small size of the sample in the area.

The methods used for SAE can be classified by the type of inference:

1) *Design based*: they make use of survey weights and the associated inferences are based on the probability distribution induced by the sampling design with the population values held fixed. The Horvitz - Thompson estimator is the most used in this category.

2) *Model assisted*: they make use of working models and are also design based, aiming at making the inferences "robust" to possible model misspecification. The role of the model is to describe the finite population point scatter, even if the assumption is never made that the population was really generated by the model. The basic property and the conclusion about finite population parameters are therefore independent of model assumptions. These procedures are thus model assisted, but they are not model dependent. The generalized regression estimator (GREG), synthetic and composite estimators are model assisted.

3) *Model based*: in literature it is called predictive approach. The parameter of interest, or its functions, is considered to be a realization of a random variable. The method starts from a specification of a super-population model that accounts for between area variation. The model permits empirical best linear unbiased prediction at small area level. Inferences from model based estimators refer to the distribution implied by the assumed model. Small area models, that make use of explicit linking models based on random area-specific effects, (*Area Level Random Effects Model* (Fay and Herriot, 1979), *Nested Error Unit Level*

Regression Model (Battese *et al.,* 1988)) and poverty mapping models, that link census data with survey data and auxiliary information, belong to this category.
In the next sections we will focus on this last category.

4.2. Mixed effects model

Methods based on models are largely used in small area estimation; they imply the introduction of probabilistic models that include area specific random effects to account for between area variation beyond that explained by the auxiliary information.

Small area models can be classified into two broad types: (a) area level random effect models, which are used when auxiliary information is available only at area level. They relate small area direct estimators to area-specific covariate; (b) nested error unit level regression models relate the unit values of a study variable to unit-specific covariates.

4.2.1. Area level random effects model

Area level random effects model implies a vector of p auxiliary variables $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ and of parameters of interest θ_i linked by a certain relationship as:

$$\theta_i = \mathbf{x}_i^T \boldsymbol{\beta} + z_i u_i, \quad i = 1,...,m \tag{4.2.1}$$

where the z_i's are known positive constants, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficient and u_i's are area-specific random effects assumed to be independent and identically distributed (i.i.d.) with mean 0 and variance σ_u^2. Normality of the random effect u_i is often used, but it is possible to make "robust" inferences by relaxing the normality assumption.

For make inferences about the small area means $\overline{\theta}_i$ under model (4.2.1) we assume that the direct estimators $\hat{\theta}_i$ exists and that the following model holds:

$$\hat{\theta}_i = \theta_i + e_i, \quad i = 1,...,m \tag{4.2.2}$$

where the sampling errors e_i are independent with $E(e_i \mid \theta_i) = 0$ and $\text{var}(e_i \mid \theta_i) = \psi_i$ often assumed known. Normality of the estimator is also often assumed, but this may not be as restrictive as the normality of the random effects.

Combining (4.2.1) and (4.2.2) we obtain the linear mixed model of Fay and Herriot (1979) that takes in account both area random effects u_i and sample errors e_i assuming their independence:

$$\hat{\theta}_i = \mathbf{x}_i^T \boldsymbol{\beta} + z_i u_i + e_i, \quad i = 1,...,m \tag{4.2.3}$$

4.2.2. Unit level random effects model

The unit level random effects model assumes that a vector of p auxiliary variables $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$ is known for each population unit j in each small area i. The variable of interest y_{ij} is assumed to be related to $\mathbf{x}_{ij}{}^T$ through a one-fold nested error linear regression model:

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + \varepsilon_{ij}, \quad j = 1, \ldots, N_i, \ i = 1, \ldots m \tag{4.2.4}$$

where u_i and ε_{ij} are random effects mutually independent with media 0 and variance σ_u^2 and σ_ε^2 respectively and are often assumed normally distributed.

We assume that a sample s_i of size n_i is taken from the N_i units in the i-th area and that selection bias is absent (assumption satisfied under simple random sampling).

The model (4.2.4) can be written in matrix form distinguishing between sampled (s) and non sampled (r) units as following:

$$\mathbf{y}_i^P = \begin{bmatrix} \mathbf{y}_{is} \\ \mathbf{y}_{ir} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{is} \\ \mathbf{X}_{ir} \end{bmatrix} \boldsymbol{\beta} + u_i \begin{bmatrix} \mathbf{1}_{is} \\ \mathbf{1}_{ir} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_{is} \\ \boldsymbol{\varepsilon}_{ir} \end{bmatrix} \tag{4.2.5}$$

If we write the mean of small are $\overline{Y}_i$ as:

$$\overline{Y}_i = f_i \overline{y}_{is} + (1 - f_i) \overline{Y}_{ir} \tag{4.2.6}$$

with $f_i = n_i / N_i$ and $\overline{y}_{is}$ (mean of sample units) and $\overline{Y}_{ir}$ (mean of non-sampled unit), the estimation of small area mean $\overline{Y}_i$ is equivalent to predict $\overline{Y}_{ir}$ given $\mathbf{y}_{is}$ and $\mathbf{X}_{is}$.

4.2.3. Generalized Linear Mixed Model

In small area estimation, many variables are not normally distributed, then linear mixed model cannot be applied. It is so necessary to use generalized linear mixed model, as logistic model for binary response or exponential model for count values.

In this model the vector $\mathbf{y}$ of values of the interest variable, is assumed to depend on a vector $\boldsymbol{\eta}$ related to covariates and random components as:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \tag{4.2.7}$$

Considering the partition sampled and non-sampled (as in the previous section) units the parameter of interest can be written as:

$$\boldsymbol{\theta} = \mathbf{a}_s \mathbf{y}_s + \mathbf{a}_r \mathbf{y}_r \tag{4.2.8}$$

and the predictor of $\boldsymbol{\theta}$ is given substituting $\mathbf{y}_r$ with the predict values (Saei and Chambers, 2003):

$$\hat{\boldsymbol{\theta}} = \mathbf{a}_s\mathbf{y}_s + \mathbf{a}_r\mathbf{y}_r = \mathbf{a}_s\mathbf{y}_s + \mathbf{a}_r f\left(\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\hat{\mathbf{u}}\right) \qquad (4.2.9)$$

4.2.4. Empirical Best Linear Unbiased Predictor (EBLUP)

Most small area models can be regarded as special cases of a general linear mixed models involving fixed effects, that determine the mean value of response variable *y*, and random effects that govern the variance-covariance structure (Mc Culloch and Searle, 2001). Moreover, small area parameters, as means or total, can be expressed as linear combinations of fixed and random effects. Following a classical approach, we can obtain best linear unbiased predictor (BLUP) estimators that minimize the mean square error among the class of linear unbiased estimators and that do not depend on normality of the random effects. However, BLUP estimators depend on the variance and covariance of random effects which can be estimated by moments method or, assuming normality, by maximum likelihood (ML) or restricted maximum likelihood (REML) methods. Using these estimated components in the BLUP estimator we obtain a two-stage estimator: the empirical BLUP or EBLUP estimator (Harville, 1991)[4].

In the next sub-sections EBLUP theory is applied to the basic area level and the basic unit level models and essential results are showed (for details see Rao 2003).

4.2.4.1 Basic area level model

The basic area level model is given by:

$$\hat{\theta}_i = \mathbf{x}_i^T\boldsymbol{\beta} + b_i u_i + e_i \qquad i = 1,...,m \qquad (4.2.10)$$

where $\mathbf{x}_i$ is a $p \times 1$ vector of area level covariates, u_i i.i.d.$(0,\sigma_u^2)$ and independent of sampling error e_i i.i.d.$(0,\psi_i)$ with known variance ψ_i, $\hat{\theta}_i$ is a direct estimator of *i*-th area

---

[4] The general linear mixed model can be defined as: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ where $\mathbf{y}$ is the *n* x 1 vector of sample observations, **X** and **Z** are known *n* x *p* and *n* x *h* matrices of full rank, and **u** and **e** are independently distributed with means **0** and covariance matrices **G** and **R** depending on some variance parameters $\boldsymbol{\delta}$. $V(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ^T} + \mathbf{R}$. The BLUP estimator of a linear combination $\mu = \tau(\boldsymbol{\delta},\mathbf{y}) = \mathbf{1^T}\boldsymbol{\beta} + \mathbf{m^T}\mathbf{u}$ is given by $\hat{\mu} = \hat{\tau}(\boldsymbol{\delta},\mathbf{y}) = \mathbf{1^T}\hat{\boldsymbol{\beta}} + \mathbf{m^T}\hat{\mathbf{u}} = \mathbf{1^T}\hat{\boldsymbol{\beta}} + \mathbf{m^T}\mathbf{GZ^T}\mathbf{V^{-1}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ with $\hat{\boldsymbol{\beta}}$ the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and $\hat{\mathbf{u}} = \hat{\mathbf{u}}(\boldsymbol{\delta}) = \mathbf{GZ^T}\mathbf{V^{-1}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$. Replacing $\boldsymbol{\delta}$ by the estimator $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}(\mathbf{y})$ we obtain the EBLUP estimator: $\hat{\tau}(\hat{\boldsymbol{\delta}},\mathbf{y}) = \mathbf{1^T}\hat{\boldsymbol{\beta}} + \mathbf{m^T}\hat{\mathbf{G}}\mathbf{Z^T}\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ with $MSE\left[\hat{\tau}(\hat{\boldsymbol{\delta}},\mathbf{y})\right] \approx g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) + g_3(\boldsymbol{\delta})$. The MSE components are expressed as follows: $g_1(\boldsymbol{\delta}) = \mathbf{m^T}(\mathbf{G} - \mathbf{GZ^T}\mathbf{V^{-1}}\mathbf{ZG})\mathbf{m}$, $g_2(\boldsymbol{\delta}) = \mathbf{d^T}(\mathbf{X^T}\mathbf{V^{-1}}\mathbf{X})^{-1}\mathbf{d}$ with $\mathbf{d} = \mathbf{1^T} - \mathbf{b^T}\mathbf{X}$ and $\mathbf{b^T} = \mathbf{m^T}\mathbf{GZ^T}\mathbf{V^{-1}}$, $g_3(\boldsymbol{\delta}) = tr\left[\left(\partial\mathbf{b}^T/\partial\delta\right)\mathbf{V}\left(\partial\mathbf{b}^T/\partial\delta\right)^T\overline{\mathbf{V}}(\hat{\boldsymbol{\delta}})\right]$ where $\overline{\mathbf{V}}(\hat{\boldsymbol{\delta}})$ is asymptotic covariance matrix of $\hat{\boldsymbol{\delta}}$.

parameter θ_i and b_i is a known positive constant. Model (4.2.10) is a special case of the general linear mixed model with block diagonal covariance structure and: $\mathbf{y}_i = \hat{\theta}_i$, $\mathbf{X}_i = \mathbf{x}_i^T$, $\mathbf{Z}_i = b_i$, $\mathbf{v}_i = v_i$ $\mathbf{e}_i = e_i$, $\boldsymbol{\beta} = (\beta_1,...,\beta_p)^T$, $\mathbf{G}_i = \sigma_u^2$, $\mathbf{R}_i = \psi_i$, $\mathbf{V}_i = \psi_i + \sigma_u^2 b_i^2$. Moreover, $\mu_i = \theta_i = \mathbf{z}_i^T \boldsymbol{\beta} + b_i u_i$ so that $\mathbf{1}_i = \mathbf{z}_i$ and $\mathbf{m}_i = b_i$. Then, substituting the above values in the general for the BLUP estimator of μ_i (see note 1), we can obtain the BLUP estimator of θ_i as a weighted average of the direct estimator $\hat{\theta}_i$ and the regression-synthetic estimator $\mathbf{x}_i^T \hat{\boldsymbol{\beta}}$:

$$\hat{\tau}_i(\sigma_u^2, \hat{\theta}_i) = \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \gamma_i \left(\hat{\theta}_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}\right) = \gamma_i \hat{\theta}_i + (1-\gamma_i)\mathbf{x}_i^T \hat{\boldsymbol{\beta}} \tag{4.2.11}$$

where $\gamma_i = \dfrac{\sigma_u^2 b_i^2}{\psi_i + \sigma_u^2 b_i^2}, 0 \le \gamma_i \le 1$, so called *shrinkage factor*, measures the model variance $\sigma_u^2 b_i^2$ relative to the total variance $\psi_i + \sigma_u^2 b_i^2$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\sigma_u^2) = \dfrac{\sum_{i=1}^{m} \mathbf{x}_i \hat{\theta}_i / (b_i^2 \sigma_u^2 + \psi_i)}{\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^T / (b_i^2 \sigma_u^2 + \psi_i)}$.

$\hat{\tau}_i(\sigma_u^2, \hat{\theta}_i)$ estimator is valid for general sampling designs and it is design-consistent because γ_i tends to 1 as the sampling variance ψ_i tends to 0.

The MSE of the BLUP estimator $\hat{\tau}_i(\sigma_u^2, \hat{\theta}_i)$ is easily obtained from the general formula:

$$MSE[\hat{\tau}_i(\sigma_u^2, \hat{\theta}_i)] = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) \tag{4.2.12}$$

where

$$g_{1i}(\sigma_u^2) = \frac{\left(b_i^2 \sigma_u^2 \psi_i\right)}{\left(b_i^2 \sigma_u^2 + \psi_i\right)} = \gamma_i \psi_i \tag{4.2.13}$$

and

$$g_{2i}(\sigma_u^2) = (1-\gamma_i)^2 \mathbf{x}_i^T \left[\frac{\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^T}{\left(b_i^2 \sigma_u^2 + \psi_i\right)} \right]^{-1} \mathbf{x}_i \tag{4.2.14}$$

Replacing σ_u^2 by an estimator $\hat{\sigma}_u^2$ we obtain an EBLUP estimator of $\hat{\tau}_i(\sigma_u^2, \hat{\theta}_i)$:

$$\hat{\tau}_i(\hat{\sigma}_u^2, \hat{\theta}_i) = \hat{\gamma}_i \hat{\theta}_i + (1-\hat{\gamma}_i)\mathbf{x}_i^T \hat{\boldsymbol{\beta}} \tag{4.2.15}$$

where $\hat{\gamma}_i$ and $\hat{\beta}$ are the value of γ_i and β when σ_u^2 is preplaced by $\hat{\sigma}_u^2$.

Under regularity conditions and normality of the errors u_i and e_i, the MSE of $\hat{\tau}_i(\hat{\sigma}_u^2, \hat{\theta}_i)$ can be approximated as:

$$MSE[\hat{\tau}_i(\hat{\sigma}_u^2, \hat{\theta}_i] = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2) \tag{4.2.16}$$

where $g_{1i}(\sigma_u^2)$ and $g_{2i}(\sigma_u^2)$ are given by (4.2.13) and (4.2.14) and

$$g_{3i}(\sigma_u^2) = \psi_i^2 b_i^4 (\psi_i + \sigma_u^2 b_u^2)^{-3} \overline{V}(\hat{\sigma}_u^2) \tag{4.2.17}$$

4.2.4.2. Basic unit level model

In this section the basic unit level model and the correspondent EBLUP estimator are illustrated briefly. Noting that the i-th small area mean can be represented as $\overline{Y}_i = \overline{\mathbf{X}}_i^T \boldsymbol{\beta} + u_i + \overline{\varepsilon}_i$, $(i = 1,...,m)$, and utilising for sample unit the model:

$$\mathbf{y}_i = \overline{\mathbf{X}}_i^T \boldsymbol{\beta} + u_i \mathbf{1}_{n_i} + \varepsilon_i, \quad i = 1,...,m \tag{4.2.18}$$

we can obtain the EPLUP estimator of $\overline{\mathbf{X}}_{ir} \boldsymbol{\beta} + u_i$ where $\overline{\mathbf{X}}_{ir}$ is the mean of non sample units. Substituting $\mathbf{Y}_{ir}$ in (4.2.6) with the predictor, and assuming σ_u^2 and σ_ε^2 known and $f_i = n_i / N_i$ negligible, the BLUP estimator of $\overline{Y}_i$ is given by:

$$\hat{\tau}_i(\sigma_u^2, \sigma_\varepsilon^2) = \gamma_i [\overline{y}_{iw} + \left(\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw}\right)^T \hat{\boldsymbol{\beta}}] + (1 - \gamma_i) \overline{\mathbf{X}}_i^T \hat{\boldsymbol{\beta}} \tag{4.2.19}$$

where $\hat{\boldsymbol{\beta}}$ is the BLUE estimator of $\hat{\boldsymbol{\beta}}$ and $\gamma_i = \dfrac{\sigma_u^2}{\left(\sigma_u^2 + \dfrac{\sigma_\varepsilon^2}{w_{i.}}\right)}$ with $w_{i.} = \sum_{j=1}^{n_i} w_{ij}$ and $w_{ij} = k_{ij}^{-2}$, $\overline{y}_{iw}$ and $\overline{\mathbf{x}}_{iw}$ weighted mean with weights w_{ij} (Prasad and Rao, 1990).

The BLUP estimator (4.2.19), then, is a weighted average of the "survey regression" estimator $\overline{y}_{iw} + \left(\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw}\right)^T \hat{\boldsymbol{\beta}}$ and the regression synthetic estimator $\overline{\mathbf{X}}_i^T \hat{\boldsymbol{\beta}}$.

Replacing σ_u^2 and σ_ε^2 by estimators $\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$ we obtain an EBLUP estimator:

$$\hat{\tau}_i(\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2) = \hat{\gamma}_i [\overline{y}_{iw} + \left(\overline{\mathbf{X}}_i - \overline{\mathbf{x}}_{iw}\right)^T \hat{\boldsymbol{\beta}}] + (1 - \hat{\gamma}_i) \overline{\mathbf{X}}_i^T \hat{\boldsymbol{\beta}} \tag{4.2.20}$$

where $\hat{\gamma}_i$ and $\hat{\beta}$ are the value of γ_i and β when $(\sigma_u^2, \sigma_\varepsilon^2)$ is preplaced by $(\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2)$.

The MSE approximation of $\hat{\tau}_i(\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2)$, under regularity conditions and normality of the error u_i and ε_{ij} is given by:

$$MSE[\hat{\tau}_i(\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2)] = g_{1i}(\sigma_u^2, \hat{\sigma}_\varepsilon^2) + g_{2i}(\sigma_u^2, \hat{\sigma}_\varepsilon^2) + g_{3i}(\sigma_u^2, \hat{\sigma}_\varepsilon^2) \tag{4.2.21}$$

where

$g_{1i}(\sigma_u^2,\sigma_\varepsilon^2)=\gamma_i(\sigma_\varepsilon^2/w_{i.})$,

$g_{2i}(\sigma_u^2,\sigma_\varepsilon^2)=(\overline{\mathbf{x}}_i-\gamma_i\mathbf{x}_{iw})^T\left(\sum_i \mathbf{A}_i\right)(\overline{\mathbf{x}}_i-\gamma_i\mathbf{x}_{iw})$, $\mathbf{A}_i=\sigma_\varepsilon^{-2}\left(\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^T-\gamma_i\mathbf{x}_{iw}\mathbf{x}_{iw}^T\right)$ and

$g_{3i}(\hat{\sigma}_u^2,\hat{\sigma}_\varepsilon^2)=w_{i.}^2\left(\sigma_u^2+\frac{\sigma_\varepsilon^2}{w_{i.}}\right)^{-3}\left[\sigma_\varepsilon^2\overline{V}(\hat{\sigma}_u^2)+\sigma_u^2\overline{V}(\hat{\sigma}_\varepsilon^2)-2\sigma_u^2\sigma_\varepsilon^2\overline{COV}(\hat{\sigma}_u^2,\hat{\sigma}_\varepsilon^2)\right]$ where $\overline{V}(\hat{\sigma}_u^2)$ and $\overline{V}(\hat{\sigma}_\varepsilon^2)$ are the asymptotic variances of $\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$, $\overline{COV}(\hat{\sigma}_u^2,\hat{\sigma}_\varepsilon^2)$ is the asymptotic covariance of $(\hat{\sigma}_u^2,\hat{\sigma}_\varepsilon^2)$.

4.3. M-quantile models

The main limitations of small area models are summarized in the follows:

1. these models depend on parametric and distributional assumptions;
2. random effects are assumed to be normally distributed;
3. the unit levels models don't take in account of sampling design;
4. the estimation process of parameters of the model doesn't consider the possible presence of outliers;
5. the domains are defined when the model is estimated.

To overcome the problems associated with the distributional assumptions of random effects, Chambers and Tzavidis (2006) proposed a new approach to small area estimation based on modelling quantile-like parameters of the conditional distribution of the target variable given the covariates. This new approach allows that inter–domain differences are characterized by the variation of area-specific M-quantile coefficients, instead of random area effects as on mixed models. The Chambers and Tzavidis proposal has some practical advantages, as no distributional assumptions and specification of random effects, easy non-parametric specification, straightforward outlier robust inference and incorporation of survey weights, estimation of other small area quantities (medians, percentiles, etc.). However, M-quantile modelling presents some drawbacks in asymptotic theory, when variables are nominal and it is never as efficient as mixed models when the assumptions of the latter approach are true.

4.3.1. Quantile regression

As seen in the previous sections, a linear mixed effects model can be expressed as:

$$y_{ij}=\mathbf{x}_{ij}^T\boldsymbol{\beta}+\mathbf{z}_{ij}^T\mathbf{u}_i+\varepsilon_{ij},\qquad i=1,\ldots,m\quad j=1,\ldots,n \tag{4.3.1}$$

After the estimation of fixed and random effects by Maximum Likelihood or Restricted Maximum Likelihood and using the available auxiliary information, domain specific estimates can be calculated. For example, if we know the population size N_i of the small area *i* we can obtain the EBLUP estimator of the mean $\bar{y}_i$ as:

$$\hat{\bar{y}}_i = N_i^{-1}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T\hat{\mathbf{u}}_i\right) \tag{4.3.2}$$

This regression model, tracing the behaviour of the mean of *y* for each *x*, gives an incomplete picture of a distribution. Then, it is to be hoped that a so called *quantile regression* is used, that is to fit a family of regression models, each one summarising the behaviour of a different percentage point of *y* at each point in the set of *x*'s.

In particular, in the linear case, quantile regression leads a family of planes indexed by the value of the corresponding percentile coefficient $q \in (0,1)$. For each value of q the corresponding model shows how $Q_q(x)$, the q^{th} quantile of the conditional distribution of *y* given *x*, varies with *x*:

$$Q_q(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}_q \tag{4.3.3}$$

where $\boldsymbol{\beta}_q$ is estimated by minimising $\sum_{j=1}^{n}\left|y_j - \mathbf{x}_j^T\mathbf{b}\right|\left\{(1-q)I\left(y_j - \mathbf{x}_j^T\mathbf{b} \le 0\right) + q\left(y_j - \mathbf{x}_j^T\mathbf{b} > 0\right)\right\}$ with respect to **b**. Quantile regression can be viewed as a generalisation of median regression.

4.3.2. M-Quantile regression

Quantile regression models can be fitted using linear programming methods that not necessarily guarantee convergence and a unique solution. For this reason, it is preferable using M-quantile regression whose *iteratively reweighted least squares* (IRLS) algorithm is guaranteed to converge to a unique solution when a continuous monotone influence function is used.

M-quantile regression provides a "quantile-like" generalisation of regression based on influence functions (M-regression). Influence functions determine the effect that residuals have on the estimation procedure.

In particular, the M-quantile of order *q* for the conditional density of *y* given *x* is defined as the solution $Q_q(\mathbf{x},\psi)$ of the estimating equation $\int \psi_q(y-Q)f(y\,|\,\mathbf{x})dy = 0$, where ψ denotes the influence function associated with the M-quantile, often the Huber Proposal 2 influence function, $\psi(t) = tI(-c \le t \le c) + c\,\mathrm{sgn}(t)I(|t > c|)$ where *c* is a cut-off constant.

A linear M-quantile regression model is given by:

$$Q_q(\mathbf{x},\psi) = \mathbf{x}^T\boldsymbol{\beta}_\psi(q) \tag{4.3.4}$$

For specified q and ψ, estimates of these regression parameters can be obtained by solving the estimating equation:

$$\sum_{j=1}^{n} \psi_q \left[\nu^{-1} \left(y_j - \mathbf{x}_j^T \hat{\boldsymbol{\beta}}_\psi(q) \right) \right] \mathbf{x}_j = 0 \qquad (4.3.5)$$

where $\psi_q(t) = 2\psi(t)\{qI(t > 0) + (1-q)I(t \leq 0)\}$ and ν is a suitable robust estimate of scale.

4.3.3. M-Quantile regression in small area estimation

Suppose to characterise conditional variability across the population of interest by the M-quantile coefficient of the population units, that is, for unit *j* with value y_j and $\mathbf{x}_j$, the value q_j such that $Q_{q_j}(\mathbf{x}_j, \psi) = y_j$. By definition;

$$\begin{aligned} \bar{y}_i &= N_i^{-1} \left(\sum_{j \in s_i} y_i + \sum_{j \in r_i} \mathbf{x}_j^T \boldsymbol{\beta}_\psi(q_j) \right) \\ &= N_i^{-1} \left(\sum_{j \in s_j} y_j + \sum_{j \in r_i} \mathbf{x}_j^T \boldsymbol{\beta}_\psi(\bar{q}_i) \right) + N_i^{-1} \sum_{j \in r_i} \mathbf{x}_j^T \left[\boldsymbol{\beta}_\psi(q_j) - \boldsymbol{\beta}_\psi(\bar{q}_i) \right] \end{aligned} \qquad (4.3.6)$$

when the conditional M-quantiles follow a linear model, with $\bar{q}_i = N_i^{-1} \sum_{j \in i} q_j$ the average value of the M-quantile coefficients of the units in area *i* and s_i, r_i respectively the sampled and non-sampled units in area *i*. A predictor of $\bar{y}_i$ is given by:

$$\hat{\bar{y}}_i = N_i^{-1} \left(\sum_{j \in s_i} y_j + \sum_{j \in r_i} \mathbf{x}_j^T \hat{\boldsymbol{\beta}}_\psi(\hat{\bar{q}}_i) \right) \qquad (4.3.7)$$

that is equivalent to using $\mathbf{x}_j^T \hat{\boldsymbol{\beta}}_\psi(\bar{q}_i)$ to predict the unobserved value y_j for population unit $j \in r_i$. For fixed q, the estimator of the M-quantile regression coefficient $\boldsymbol{\beta}_\psi(q)$ is $\hat{\boldsymbol{\beta}}_\psi(q) = \left(\mathbf{X}_s^T \mathbf{W}_s(q) \mathbf{X}_s \right)^{-1} \mathbf{X}_s^T \mathbf{W}_s(q) \mathbf{y}_s$ with $\mathbf{X}_s$ the *n* x *p* matrix of sample covariates, $\mathbf{y}_s$ the *n*-vector of sample *y* values and $\mathbf{W}_s(q)$ the diagonal matrix of final weights produced by IRLS algorithm used to compute $\hat{\boldsymbol{\beta}}_\psi(q)$.

Moreover, in order to calculated (4.3.7) we need the estimated M-quantile coefficient for area *i*, i.e. $\hat{\bar{q}}_i$, that depends on the sample M-quantile coefficients, $\{q_{js}; j \in s\}$ which characterise the variation in the conditional distribution of *y* given **x** in the sample as the q_j characterise this distribution in the population. The q_{js} values are obtain by linear interpolation over a fine grid defined on the (0,1) interval and using the sample data to fit M-quantile regression lines at each value q on this grid.

Provided the sampling method is non-informative given **x**, $\hat{\bar{q}}_i$ can be calculated as the mean or other quantities (e.g. median) of the q_{js} values in area i.

4.3.4. Mean squared error

An approximation of the prediction variance of (4.3.7) is given by:

$$Var\left(\hat{\bar{y}}_i - \bar{y}_i\right) \approx N_i^{-2}\left(\sum_{j\in s} u_{ij}^2 Var(y_j) + \sum_{j\in r} Var(y_j)\right) \tag{4.3.8}$$

where $\mathbf{u}_i = (u_{ij}) = \mathbf{W}_s(\hat{\bar{q}}_i)\mathbf{X}_s\left(\mathbf{X}_s^T\mathbf{W}_s(\hat{\bar{q}}_i)^{-1}\mathbf{X}_s\right)^{-1}\mathbf{t}_{ri}$ with $\mathbf{t}_{ri}$ the sum of the non-sample covariates in area i. If we take this variance to be unconditional, i.e. not specific to the area from which y_j is drawn, the estimator of the prediction variance of (4.3.7) becomes:

$$\hat{V}_i = \sum_{j\in s} \theta_{ij}\left(y_j - \mathbf{x}_j^T\hat{\boldsymbol{\beta}}_\psi(0.5)\right)^2 \tag{4.3.9}$$

Otherwise, if we consider $Var(y_j)$ conditionally, i.e. specific to the area k from which y_j is drawn, we have:

$$\hat{V}_i = \sum_k \sum_{j\in s_k} \theta_{ij}\left(y_j - \mathbf{x}_j^T\hat{\boldsymbol{\beta}}_\psi(\hat{\bar{q}}_k)\right)^2 \tag{4.3.10}$$

In both formulas (4.3.9) and (4.3.10) $\theta_{ij} = N_i^{-2}\left(u_{ij}^2 + I(j\in i)(N_i - n_i)/(n_i - 1)\right)$.

It follows that the estimator of the mean square error of (4.2.7) is given by:

$$\hat{M}_i = \hat{V}_i + \hat{B}_i^2 \tag{4.3.11}$$

with $\hat{B}_i = N_i^{-1}\left(\sum_k \sum_{j\in s_k} w_{ij}\mathbf{x}_j^T\boldsymbol{\beta}(\hat{\bar{q}}_k) - \sum_{j\in i}\mathbf{x}_j^T\boldsymbol{\beta}(\hat{\bar{q}}_i)\right)$, $w_{ij} = \mathbf{1}_{si} + \mathbf{W}_s(\hat{\bar{q}}_i)\mathbf{X}_s\left(\mathbf{X}_s^T\mathbf{W}_s(\hat{\bar{q}}_i)\mathbf{X}_s\right)^{-1}\mathbf{t}_{ri}$

where $\mathbf{1}_{si}$ is the n-vector with i-th component equal to one whenever the corresponding sample unit is in area i and is zero otherwise.

4.4. General framework for robust bias adjusted small area predictors

Chambers and Tzavidis (2006) observed that M-quantile predictors of small area means are biased. In 2008 they proposed a bias adjustment based on representing this predictor as a functional of a corresponding predictor of the small area empirical distributional function using the Chambers and Dunstan (1986) smearing type predictor or the Rao-Kovar-Mantel (1990) predictor of the distributional function.

The area empirical distribution function of *y* for area *i* can be expressed by:

$$F_i(t) = N_i^{-1}\left\{\sum_{j\in s_i} I(y_j \le t) + \sum_{j\in r_i} I(y_j \le t)\right\} \tag{4.4.1}$$

Then the problem of predicting $F_i(t)$ reduces to predicting the values y_j for the non-sampled units in small area i replacing them by their predicted value $\hat{y}_j$ under an appropriate model. The predictor of 4.4.1 is given by:

$$\hat{F}_i(t) = N_i^{-1}\left\{\sum_{j\in s_i} I(y_j \le t) + \sum_{j\in r_i} I(\hat{y}_j \le t)\right\} \tag{4.4.2}$$

Consequently, the predictor of the mean is expressed as:

$$\hat{\bar{y}}_i = \int_{-\infty}^{+\infty} t d\hat{F}_i(t) = N_i^{-1}\left(\sum_{j\in s_i} y_i + \sum_{j\in r_i} \hat{y}_j\right) \tag{4.4.3}$$

We can note that the EBLUP is the mean functional defined by (4.4.2) when $\hat{y}_j = x_j^T \hat{\beta} + z_j^T \hat{u}_i$, while the M-quantile predictor is a mean functional with $\hat{y}_j = x_j^T \hat{\beta}_\psi(\hat{\bar{q}}_i)$.

In general, a predictor $\hat{m}_{pi}$ of the p^{th} quantile of the distribution of *y* in area *i* is defined as the solution to the estimating equation:

$$\int_{-\infty}^{\hat{m}_{pi}} d\hat{F}_i(t) = p \tag{4.4.4}$$

given a suitable predictor $\hat{F}_i(t)$ of the area *i* distribution of *y*.

As said above, Chambers and Tzavidis (2006) observed that M-quantile predictors of small area means can be biased. By combining a smearing argument (Duan, 1983) with a model for finite population distribution of *y*, Chambers and Dunstan developed a model-consistent predictor for a finite population distribution function. Assuming that the residuals $\varepsilon_j = y_j - \mu_j$ are homoskedastic within the small area of interest, formula (4.4.2) becomes:

$$\hat{F}_i^{CD}(t) = N_i^{-1}\left\{\sum_{j\in s_i} I(y_j \le t) + n_i^{-1} \sum_{j\in s_i}\sum_{k\in r_i} I\left\{\hat{\mu}_k + (y_j - \hat{\mu}_j) \le t\right\}\right\} \tag{4.4.5}$$

and the mean functional defined by 4.4.4 takes the value:

$$\hat{\hat{y}}_i^{CD} = \int_{-\infty}^{+\infty} t d\hat{F}_i^{CD}(t) = N_i^{-1}\left(\sum_{j\in s_i} y_j + \sum_{j\in r_i}\hat{\mu}_j + (f_i^{-1}-1)\sum_{j\in s_i}(y_j - \hat{\mu}_j)\right) \quad (4.4.6)$$

where $f_i = n_i / N_i$ is the sampling fraction in area *i*. We obtain a bias-adjusted alternative to the EBLUP when we substitute $\hat{\mu}_j = x_j^T\hat{\beta} + z_j^T\hat{u}_i$ in 4.4.6, while we obtain a bias-adjusted M-quantile predictor when we substitute $\hat{\mu}_j = x_j^T\hat{\beta}_\psi(\hat{\bar{q}}_i)$.

A predictor that is both design-consistent and model-consistent has been proposed by Rao-Kovar-Mantel. Under simple random sampling, for the finite population distribution function, it is expressed as:

$$\begin{aligned}\hat{F}_i^{RKM}(t) &= n_i^{-1}\sum_{j\in s_i} I(y_j \le t) + N_i^{-1}\sum_{k\in r_i} n^{-1}\sum_{j\in s_i} I(y_j - \hat{y}_j \le t - \hat{y}_k) \\ &- (n_i^{-1} - N_i^{-1})\sum_{k\in s_i} n_i^{-1}\sum_{j\in s_i} I(y_j - \hat{y}_j \le t - \hat{y}_k)\end{aligned} \quad (4.4.7)$$

4.5. Spatial models

4.5.1. The Spatial EBLUP estimator

As noted in the previous sections, model-based methods of small area estimation are often based on assuming a linear mixed models, with area-specific random effects to account for between area variation beyond that explained by auxiliary variables included in the fixed part of the model. Independence of these random effects is assumed in most cases, but it is often reasonable to suppose that the effects of neighbouring area are correlated, with the correlation decaying to zero as the distance between these areas increases.

As illustration of models with spatial dependence, only area level random effect model will be take into consideration.

Let $\boldsymbol{\theta}$ be the $m \times 1$ vector of the parameter of interest and assume that the $m \times 1$ vector of the direct estimator $\hat{\boldsymbol{\theta}}$ is available and design unbiased, that is $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \mathbf{e}$ with **e** the vector of independent sampling errors with mean **0** and known diagonal variance matrix $\boldsymbol{\psi}$. The spatial dependence among small areas is introduced by specifying a linear mixed model with spatially correlated random effects for the $\boldsymbol{\theta}$ parameter:

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} \quad (4.5.1)$$

where **X** is the $m \times p$ matrix of the area specific auxiliary covariates, $\boldsymbol{\beta}$ is the regression parameters vector $p \times 1$, **Z** is a $m \times m$ matrix of known positive constants, **v** is the $m \times 1$ vector of the second order variation. Basically there are two approaches to describe the spatial second order variation: Simultaneously Autoregressive Models (SAR) and

Conditional Autoregressive Models (CAR). In the follows, the deviations from the fixed part of the model $\mathbf{X\beta}$ are the result of a simultaneously autoregressive process with parameter ρ, spatial autoregressive coefficient and $m \times m$ proximity matrix $\mathbf{W}$ (Cressie, 1993; Anselin, 1992):

$$\mathbf{v} = \rho\mathbf{W}\mathbf{v} + \mathbf{u} \Rightarrow \mathbf{v} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u} \tag{4.5.2}$$

where $\mathbf{u}$ is a $m \times 1$ vector of independent error terms with zero mean and constant variance σ_u^2 and $\mathbf{I}$ is the $m \times m$ identity matrix. Combining (4.5.1) and (4.5.2), with $\mathbf{e}$ independent of $\mathbf{v}$, we obtain the model with spatially correlated random area effects:

$$\hat{\boldsymbol{\theta}} = \mathbf{X\beta} + \mathbf{Z}(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u} + \mathbf{e} \tag{4.5.3}$$

The error terms $\mathbf{v}$ and $\mathbf{e}$ have, respectively, $m \times m$ covariance matrices, $\mathbf{G} = \sigma_u^2\left[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)\right]^{-1}$, that is the SAR dispersion matrix, and $\mathbf{R} = \boldsymbol{\psi} = diag(\psi_i)$. Then, the covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by:

$$\mathbf{V} = \mathbf{R} + \mathbf{ZGZ^T} = diag(\psi_i) + \mathbf{Z}\sigma_u^2\left[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)\right]^{-1}\mathbf{Z}^T \tag{4.5.4}$$

The W matrix describes the neighbourhood structure of the small areas whereas ρ defines the strength of the spatial relationship among the random effects associated with neighbouring areas. The spatial weight matrix represents the potential interaction between locations. A general spatial weight matrix can be defined by a symmetric binary contiguity matrix which can be generated from topological information provided by the geographical information system (GIS) based on adjacency criteria: the element of the spatial weight matrix $\{w_{ij}\}$ is one if location i is adjacent to location j , and zero otherwise. Generally, for ease of interpretation, the general spatial weight matrix is defined in row standardized form, in which the row elements sum to one. In this ρ is called a spatial autocorrelation parameter.

The Spatial Best Linear Unbiased Predictor (Spatial BLUP) estimator of θ_i is defined as:

$$\begin{aligned}\tilde{\theta}_i^S(\sigma_u^2, \rho) = \mathbf{x}_i\hat{\boldsymbol{\beta}} + \mathbf{b}_i^T\left\{\sigma_u^2\left[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)\right]^{-1}\right\}\mathbf{Z}^T \\ \times\left\{diag(\psi_i) + \mathbf{Z}\sigma_u^2\left[(\mathbf{I} - \rho\mathbf{W})(\mathbf{I} - \rho\mathbf{W}^T)\right]^{-1}\mathbf{Z}^T\right\}^{-1}(\hat{\boldsymbol{\theta}} - \mathbf{X}\hat{\boldsymbol{\beta}})\end{aligned} \tag{4.5.5}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\hat{\boldsymbol{\theta}}$ and $\mathbf{b}_i^T$ i is $1 \times m$ vector (0, 0, . . . 0, 1, 0, . . . 0) with 1 in the i-th position. The Spatial BLUP is equal to the traditional BLUP under the random area specific effects model when $\rho = 0$ (Pratesi and Salvati 2005).

The estimator $\tilde{\theta}_i^S(\sigma_u^2, \rho)$ depends on the unknown variance components σ_u^2 and ρ. Replacing the parameters with estimators $\hat{\sigma}_u^2$, $\hat{\rho}$, a two stage estimator $\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\rho})$ is obtained and is called Spatial EBLUP:

$$\begin{aligned}\tilde{\theta}_i^S(\hat{\sigma}_u^2, \hat{\rho}) = \mathbf{x}_i\hat{\boldsymbol{\beta}} + \mathbf{b}_i^T\left\{\hat{\sigma}_u^2\left[(\mathbf{I} - \hat{\rho}\mathbf{W})(\mathbf{I} - \hat{\rho}\mathbf{W}^T)\right]^{-1}\right\}\mathbf{Z}^T \\ \times\left\{diag(\psi_i) + \mathbf{Z}\hat{\sigma}_u^2\left[(\mathbf{I} - \hat{\rho}\mathbf{W})(\mathbf{I} - \hat{\rho}\mathbf{W}^T)\right]^{-1}\mathbf{Z}^T\right\}^{-1}(\hat{\boldsymbol{\theta}} - \mathbf{X}\hat{\boldsymbol{\beta}})\end{aligned} \tag{4.5.6}$$

Assuming normality of the random effects, σ_u^2 and ρ can be estimated both by maximum likelihood (ML) and restricted maximum likelihood (REML) procedures. The ML estimators, $\hat{\sigma}_u^2$ and $\hat{\rho}$, can be obtained iteratively using the "Nelder-Mead" algorithm (Nelder and Mead, 1965) and the "scoring" algorithm in sequence. This is necessary because the log-likelihood function has multiple local maxima.

For the Spatial EBLUP, given normality of random effects, the $MSE[\tilde{\theta}_i^S(\hat{\sigma}_u^2,\hat{\rho})]$ is given by:

$$MSE[\tilde{\theta}_i^S(\hat{\sigma}_u^2,\hat{\rho})] = g_{1i}(\sigma_u^2,\rho) + g_{2i}(\sigma_u^2,\rho) + g_{3i}(\sigma_u^2,\rho) \qquad (4.5.7)$$

where $g_{1i}(\sigma_u^2,\rho)$ is due to the estimation of random effects, $g_{2i}(\sigma_u^2,\rho)$ depends on the estimation of $\boldsymbol{\beta}$ and $g_{3i}(\sigma_u^2,\rho)$ is due to the estimation of variance components (for details see Pratesi and Salvati, 2008).

4.5.2. M-Quantile geographically weighted regression

M-quantile models in small area estimation also implicitly assume independence of random area effects. As we saw in the previous section, SAR models allow for spatial correlation in the error structure. An alternative approach to incorporating the spatial information in the regression model is by assuming that the regression coefficients vary spatially across the geography of interest. *Geographically Weighted Regression* (GWR) (Brunsdon *et al.*, 1999; Fotheringham *et al.*, 1997; Yu and Wu, 2004) extends the traditional regression model by allowing local rather than global parameters to be estimated.
A GWR model for the conditional expectation of y given x at location u is given by[5]:

$$y_{jl} = x_{jl}\beta(u_l) + \varepsilon_{jl} \qquad (4.5.8)$$

with n observations at a set of L locations $\{u_l; l=1,...,L; L \le n\}$ and n_l data values $\{y_{jl}, x_{jl}; j=1,...,n_l\}$ observed at location u_l. $\beta(u)$ at arbitrary location u is estimated using weighted least squares as:

$$\hat{\beta}(u) = \left\{\sum_{l=1}^{L} w(u_l,u)\sum_{j=1}^{n_l} x_{jl}x_{jl}^T\right\}^{-1}\left\{\sum_{l=1}^{L} w(u_l,u)\sum_{j=1}^{n_l} x_{jl}y_{jl}\right\} \qquad (4.5.9)$$

where $w(u_l,u)$ is a spatial weighting function whose value depends on the distance from sample location u_l to u in the sense that sample observations with locations close to u have more weight than those further away. Usually, Euclidean distance is used.

Generalizing, the M-quantile GRW model is given by:

[5] The subscript i is dropped because M-quantile models do not depend on how areas are specified.

$$Q_q(x,\psi,u) = x^T \beta_\psi(u,q) \tag{4.5.10}$$

where $\beta_\psi(u,q)$ varies with u and q and it can be estimated by solving:

$$\sum_{l=1}^{L} w(u_l,u) \sum_{j=1}^{n_l} \psi_q \{ y_{jl} - x_{jl}^T \beta_\psi(u,q) \} x_{jl} = 0 \tag{4.5.11}$$

with $\psi_q(t)$ influence function, usually Huber proposal 2 function. The estimate $\hat{\beta}_\psi(u,q)$ is obtained using an iteratively re-weighted least squares algorithm that combines the iteratively re-weighted least squares algorithm used to fit 'spatially stationary' M-quantile model and the weighted least squares algorithm used to fit a GWR model.

An alternative spatial M-quantile model with a smaller number of parameters is one that combines local intercepts with global slopes as:

$$Q_q(x,\psi,u) = x^T \beta_\psi(q) + \delta_\psi(u,q) \tag{4.5.12}$$

where $\delta_\psi(u,q)$ is a real valued spatial process with zero mean function over the space defined by locations of interest. This model is fitted in two steps: first of all, we ignore the spatial structure in the data and estimate $\beta_\psi(q)$ directly via the iterative re-weighted least squares algorithm used to fit the standard linear M-quantile regression model, then we estimate $\delta_\psi(u,q)$ using geographic weighting.

The results just obtained of the spatial extensions of the M-quantile model can be applied on small area estimation. Assuming only one population value per location (index l dropped) and that the geographical coordinates of every unit in the population are known, the bias-adjusted M-quantile GWR predictor of the mean $\bar{y}$ in small area i is:

$$\hat{\bar{y}}_i^{MQGWR/CD} = N_i^{-1} \left[\sum_{j \in s_i} y_j + \sum_{j \in r_i} \hat{Q}_{\hat{\theta}_i}(x_j,\psi,u_j) + \frac{N_i - n_i}{n_i} \sum_{j \in s_i} \left\{ y_j - \hat{Q}_{\hat{\theta}_i}(x_j,\psi,u_j) \right\} \right] \tag{4.5.13}$$

where $\hat{\theta}_i$ is the average value of the sample M-quantile GWR coefficient in area i.

An estimator of a first order approximation to the mean square error of (4.5.13) under the model (4.5.10) has been proposed by Salvati *et al.* (2008).

4.6. Temporal models

Most of the research on small estimation has focused on cross-sectional data at a given point in time. However, often data are available for many small areas simultaneously, then, it is useful to borrow information both cross-sectionally and over time. Rao and Yu (1994) proposed a combined cross-sectional and time-series model involving auto-correlated random effects and sampling errors with an arbitrary covariance matrix over time. This extension of the basic Fay Herriot is given by:

$$y_{it} = x_{it}\beta + v_i + u_{it} + e_{it}, \qquad i = 1,...I, \quad t = 1,...,T \tag{4.6.1}$$

where y_{it} is a direct estimator of the indicator of interest for area *i* and time instant *t*, and x_{it} is a vector containing the aggregated values of *p* auxiliary variables. They assume that the random small area effects v_i are i.i.d. normal with variance σ_v^2, the random vectors u_{it}'s follow autoregressive processes of order 1 (i.i.d. AR(1)) with variance and auto-correlation parameters σ_u^2 and ρ respectively, the sampling errors e_{it}'s are normally distributed with zero mean and block diagonal covariance matrix $\mathbf{\Sigma}$ with arbitrary but known blocks $\mathbf{\Sigma}_i$ where $\mathbf{\Sigma}_i$ is a *T* x *T* matrix, and the v_i's, the u_{it}'s and the e_{it}'s are independent.

If σ_u^2, σ_v^2 and ρ are known, the BLUP estimator of the mean $\boldsymbol{\theta}_{iT} = \mathbf{x}_{iT}^{'}\boldsymbol{\beta} + v_i + u_{iT}$ is:

$$\tilde{\boldsymbol{\theta}}_{iT} = \mathbf{x}_{iT}^{'}\hat{\boldsymbol{\beta}} + (\sigma_v^2\mathbf{1}_T + \sigma^2\boldsymbol{\gamma}_T)^{'}(\mathbf{\Sigma}_i + \sigma^2\mathbf{\Gamma} + \sigma_v^2\mathbf{J}_T)^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}) \tag{4.6.2}$$

where $\hat{\boldsymbol{\beta}}$ is the generalized least-squares estimator of $\boldsymbol{\beta}$, $\mathbf{\Gamma}$ is a *T* x *T* matrix with elements $\rho^{|i-j|}/(1-\rho^2)$, $\boldsymbol{\gamma}_T$ is the *T*-th row of $\mathbf{\Gamma}$ and $\mathbf{J}_T = \mathbf{1}_T\mathbf{1}_T^{'}$. (4.6.2) can also be written as a weighted sum of the direct estimator y_{iT}, the synthetic estimator $\mathbf{x}_{iT}^{'}\hat{\boldsymbol{\beta}}$ and the residuals $y_{ij} - \mathbf{x}_{ij}^{T}\hat{\boldsymbol{\beta}}$, *j*=1,…,*T*-1:

$$\tilde{\boldsymbol{\theta}}_{iT} = w_{iT}^{*}y_{iT} + (1 - w_{iT}^{*})\mathbf{x}_{it}^{'}\hat{\boldsymbol{\beta}} + \sum_{j=1}^{T-1} w_{ij}^{*}(y_{ij} - \mathbf{x}_{ij}^{'}\hat{\boldsymbol{\beta}}) \tag{4.6.3}$$

where $(w_{i1}^{*},...,w_{iT}^{*}) = (\sigma_v^2\mathbf{1}_T + \sigma^2\boldsymbol{\gamma}_T)^{'}\mathbf{V}_i^{-1}$ with $\mathbf{V}_i = \mathbf{\Sigma}_i + \sigma^2\mathbf{\Gamma} + \sigma_v^2\mathbf{J}_T$.

A two-stage estimator (EBLUP) $\hat{\theta}_{iT}(\rho)$ of θ_{iT}, supposing ρ known, is obtained from (4.6.2) by substituting the consistent estimators $\hat{\sigma}^2(\rho)$ and $\hat{\sigma}_v^2(\rho)$ for σ^2 and σ_v^2 respectively. The variance components estimators are obtained by Rao and Yu (1994) extending the method of Pantula and Pollock (1985) to the model with both auto-correlated errors u_{it} and sampling errors e_{it}.

Moreover, a second order approximation to the $MSE[\hat{\theta}_{iT}(\rho)]$ under normality of the errors v_i, u_{it} and e_{it} is given by:

$$MSE[\hat{\theta}_{iT}(\rho)] \approx g_{1iT}(\sigma^2, \sigma_v^2, \rho) + g_{2iT}(\sigma^2, \sigma_v^2, \rho) + g_{3iT}(\sigma^2, \sigma_v^2, \rho) \tag{4.6.4}$$

where

$$g_{1iT}(\sigma^2, \sigma_v^2, \rho) = \sigma_v^2 + \frac{\sigma^2}{1-\rho^2} - (\sigma_v^2\mathbf{1}_T + \sigma^2\boldsymbol{\gamma}_T)^{'}\mathbf{V}_i^{-1}(\sigma_v^2\mathbf{1}_T + \sigma^2\boldsymbol{\gamma}_T) \tag{4.6.5}$$

$$g_{2iT}(\sigma^2,\sigma_v^2,\rho)=\left\{\mathbf{x}_{iT}-\mathbf{X}_i^{'}\mathbf{V}_i^{-1}(\sigma_v^2\mathbf{1}_T+\sigma^2\boldsymbol{\gamma}_T)\right\}^{'}(\mathbf{X}^{'}\mathbf{V}^{-1}\mathbf{X})^{-1}\left\{\mathbf{x}_{iT}-\mathbf{X}_i^{'}\mathbf{V}_i^{-1}(\sigma_v^2\mathbf{1}_T+\sigma^2\boldsymbol{\gamma}_T)\right\} \tag{4.6.6}$$

and

$$g_{3iT}(\sigma^2,\sigma_v^2,\rho)=tr(\boldsymbol{\Delta}^{'}\mathbf{V}\boldsymbol{\Delta}\boldsymbol{\Sigma}^{*}) \tag{4.6.7}$$

where $\boldsymbol{\Sigma}^{*}$ is the 2 x 2 covariance matrix of unbiased estimators of $\sigma^2(\rho)$ and $\sigma_v^2(\rho)$, and $\boldsymbol{\Delta}=(\partial\mathbf{b}/\partial\sigma^2,\partial\mathbf{b}/\partial\sigma_v^2)$ with $\mathbf{b}^{'}=(\sigma_v^2\mathbf{1}_T+\sigma^2\boldsymbol{\gamma}_T)^{'}\mathbf{V}_i^{-1}$.

A second-order approximation to estimator of $MSE[\hat{\theta}_{iT}(\rho)]$ for a small or moderate number of time points and a relatively large number of small area is given from (4.6.7) substituting the consistent estimators $\hat{\sigma}^2(\rho)$ and $\hat{\sigma}_v^2(\rho)$ for σ^2 and σ_v^2 respectively.

The results obtained above assume known ρ. However, in practice, ρ is often unknown. In this case, Rao and Yu (1994) proposed three different methods to estimate θ_{iT}. In the first one a two-stage estimator $\hat{\theta}_{iT}(\hat{\rho})$ based on a prior guess ρ_0 is used. In method 2, the sampling errors e_{it} is ignored and a naive estimator of ρ is obtained as:

$$\hat{\rho}_N=\frac{\sum_{i=1}^{I}\sum_{t=1}^{T-2}\hat{a}_{it}(\hat{a}_{i,t+1}-\hat{a}_{i,t+2})}{\sum_{i=1}^{I}\sum_{t=1}^{T-2}\hat{a}_{it}(\hat{a}_{i,t}-\hat{a}_{i,t+1})},\qquad T>2 \tag{4.6.8}$$

where $\hat{a}_{it}=y_{it}-\mathbf{x}_{it}^{'}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$ is the *it*-th ordinary least-squares residual. Even if this estimator is inconsistent and typically underestimates ρ, the resulting two-stage estimator $\hat{\theta}_{iT}(\hat{\rho}_N)$ remains unbiased. Finally, in method 3 a consistent moment estimator of ρ is obtained by taking into account the sampling errors in the follows way:

$$\hat{\rho}_N=\frac{\sum_{i=1}^{I}\sum_{t=1}^{T-2}\hat{a}_{it}(\hat{a}_{i,t+1}-\hat{a}_{i,t+2})-(\sigma_{t,t+1}^{(i)}-\sigma_{t,t+2}^{(i)})}{\sum_{i=1}^{I}\sum_{t=1}^{T-2}\hat{a}_{it}(\hat{a}_{i,t}-\hat{a}_{i,t+1})-(\sigma_{t,t}^{(i)}-\sigma_{t,t+1}^{(i)})},\qquad T>2 \tag{4.6.9}$$

where $\sigma_{t,t}^{(i)}=Var(e_{it})$, $\sigma_{t,t+1}^{(i)}=Cov(e_{it},e_{i,t+1})$ and $\sigma_{t,t+2}^{(i)}=Cov(e_{it},e_{i,t+2})$. This last method runs into difficulties since $\hat{\rho}$ often takes values outside the admissible range (-1, 1) especially for small T or small σ^2.For each of the three methods, the MSE of the two-stage estimator $\hat{\theta}_{iT}(\hat{\rho})$ is obtained by substituting the estimate of ρ in (4.6.4).

Morales et al. (2009), proposed a model related to the model (4.6.1) that considers only u_{it} to take into account the area by time variability through specific random effects. The model is:

$$y_{it} = x_{it}\beta + u_{it} + e_{it}, \quad i = 1,...I, \quad t = 1,...,m_i \qquad (4.6.10)$$

where y_{it} is a direct estimator of the indicator of interest for area *i* and time instant *t*, and x_{it} is a vector containing the aggregated values of *p* auxiliary variables. They assume that the random vectors u_{it}'s follow i.i.d. AR(1) processes with variance and auto-correlation parameters σ_u^2 and ρ respectively, the sampling errors e_{itj}'s are normally distributed with zero mean and known variance σ_{it}^2 and the u_{it}'s are independent of the e_{it}'s.

They proposed also a simplification of model (4.6.11) assuming $\rho = 0$ that is useful for those cases where survey data is only available for a reduced number of time instants.

4.7. Non-parametric M-quantile regression

In the previous sections we described basic and M-quantile regression on small area estimation assuming that the quantities of interest are linear combinations of the covariates. This method can lead to biased estimators of the small area parameters when the functional form of the relationship between the quantity of interest and the covariates is not linear. In this case, using nonparametric smoothing of this functional form can gives better results. In the following we focus on two different approaches: first of all Opsomer *et al.* (2008) employed penalized splines (p-splines) for small area estimation based on mixed effects models; on the other hand Pratesi *et al*. (2006a, 2006b) proposed a nonparametric M-quantile regression based on p-splines, applied it to the context of small area estimation. The choice of p-splines, that rely on a set of basis functions to handle nonlinear structures in the data, is due to their simplicity of implementation (Ruppert *et al.* 2003).

In general, the spline-based non-parametric model is given by:

$$y_i = m_0(x_i) + \varepsilon_i \qquad (4.7.1)$$

where the ε_i are independent random variables with mean 0 and variance σ_ε^2 and the function $m_0(.)$ is unknown but it can be approximated via p-splines by:

$$m(x,\beta,\gamma) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{K} \gamma_k (x - \kappa_k)_+^p \qquad (4.7.2)$$

Here, $(t)_+^p = t^p$ if $t > 0$ and 0 otherwise, *p* is the degree of the spline, κ_k for *k* = 1, …, *K* is a set of knots, $\boldsymbol{\beta} = (\beta_0,...,\beta_p)'$ and $\boldsymbol{\gamma} = (\gamma_1,...,\gamma_K)'$ are the coefficient vectors of the parametric and the spline part of the model respectively. To avoid the problem of over-parameterization a penalty on the magnitude of the spline parameters γ is put.

Opsomer *et al*. assume that in small area estimation the data follow the model:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{Du} + \varepsilon \qquad (4.7.3)$$

where $\mathbf{D} = (\mathbf{d}_1,...,\mathbf{d}_n)'$ with d_{ij} the indicator taking value 1 if observation *j* is in small area *i* and 0 otherwise, $\mathbf{x}_j = (1, x_j,..., x_j^p)$ and $\mathbf{z}_j = ((x_j - \kappa_1)_+^p,...,(x_j - \kappa_K)_+^p)$.

Model (4.7.3) includes the spline function as a non-parametric mean function specification, and the small area random effects **Du**. $\mathbf{Z}\gamma$ is a random-effect term. If the variances of the random components are known, BLUP estimators of parameters are obtained, otherwise variance components are estimated via restricted maximum likelihood minimization or related methods and EBLUP estimators are given.

For a given small area *i*, usually the quantity of interest is the mean:

$$\bar{y}_i = \bar{\mathbf{x}}_i\beta + \bar{\mathbf{z}}_i\gamma + u_i \qquad (4.7.4)$$

where $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{z}}_i$ are the means of x_j and of the spline basis functions over the small area and they are assumed to be known, and u_i is the small area effect. A predictor of y_i is given by:

$$\hat{y}_i = \bar{\mathbf{x}}_i\hat{\beta} + \bar{\mathbf{z}}_i\hat{\gamma} + \mathbf{e}_i\hat{\mathbf{u}} \qquad (4.7.5)$$

where $\mathbf{e}_i$ is a vector with 1 in the *i*-th position and 0 otherwise.

Opsomer *et al.* (2008) also discuss the prediction mean-squared error (PMSE) of the small area estimates and an estimator for that quantity and likelihood ratio testing for the significance of the spline term and the small area random effect. Finally, they propose a simple bootstrap method for both PMSE estimation and testing.

On the other hand, following the approach of Pratesi *et al*. (2006a), a p-spline model for the *q*-th conditional quantile of *y* given a single covariate *x* is given by:

$$Q_q(x,\psi) = \beta_{0\psi}(q) + \beta_{1\psi}(q)x + ... + \beta_{p\psi}(q)x^p + \sum_{k=1}^{K}\gamma_{(p+k)\psi}(q)(x-\kappa_k)_+^p \qquad (4.7.6)$$

where ψ is a specified influence function, $(t)_+^p = t^p$ if $t > 0$ and 0 otherwise, *p* is the degree of the spline and κ_k for $k = 1, \ldots, K$ is a set of knots, usually uniformly spread quantiles of the unique values of *x* (see Ruppert *et al.*, 2003, for details).

The vector $\boldsymbol{\beta}_\psi(q) = (\beta_{0\psi}(q),...,\beta_{p\psi}(q))'$ is the coefficient vector of the parametric part of the model and $\boldsymbol{\gamma}_\psi(q) = (\gamma_{1\psi}(q),...,\gamma_{K\psi}(q))'$ is the coefficient vector for the spline one. The number of knots *K* is chosen to be large and the influence of the knots is limited by putting a constraint on the size of the spline coefficients. Regression parameters of the nonparametric M-quantile regression model can be estimated solving through *iteratively reweighted penalized least squares* (IRPLS) the equations:

$$\sum_{j=1}^{n}\psi_q(y_j - \mathbf{x}_j\boldsymbol{\beta})\mathbf{x}_j + \lambda\sum_{k=1}^{K}\gamma_{(p+k)\psi} = \mathbf{0} \qquad (4.7.7)$$

where $\psi_q(t) = 2\psi(s^{-1}(t))\{qI[(t) > 0] + (1-q)I[(t) \le 0]\}$ is the first derivate of the function ρ that gives the contribution of each residual to the objective function, ψ is the Huber Proposal 2 influence function and $\mathbf{x}_j = (1, x_j, ..., x_j^p, (x_j - \kappa_1)_+^p, ..., (x_j - \kappa_K)_+^p)$. The value of the smoothing parameter λ is chosen as a nested step within the IRPLS procedure through optimization of a Generalized Cross Validation criterion.

Then, the mean estimator for small area i is given by:

$$\hat{\bar{y}}_i = \int t d\hat{F}_{CD,i}(t) = \frac{1}{N_i}\left(\sum_{j \in s_{n_i}} y_j + \sum_{j \in r_i} \mathbf{x}_j \hat{\boldsymbol{\beta}}_\psi(\hat{\bar{q}}_i) + \frac{N_i - n_i}{n_i} \sum_{j \in s_{n_i}} (y_j - \mathbf{x}_j \hat{\boldsymbol{\beta}}_\psi(\hat{\bar{q}}_i)) \right) \tag{4.7.8}$$

where $\hat{F}_{CD,i}(t)$ is the estimated cumulative distribution function for each small area and $\hat{\bar{q}}_i$ is the average value of the sample M-quantile coefficients of all the units in area i.

The generalisation to more than one covariate can be easily obtained by suitably changing the parametric and the spline part of the model. Moreover, other continuous or categorical variables can be inserted parametrically in the model (semi-parametric M-quantile regression, Pratesi *et al.* 2008).

Chapter 5

EB method for estimation of small domain traditional and *fuzzy* poverty measures

5.1 Introduction

The aim of this chapter is to compare the results of analyses based on different small area estimation methods, including a new methodology for obtaining empirical best predictors of general (linear or non-linear) domain parameters using unit level linear regression models and that can resolve computational problems due to big populations or more complex poverty measures. The target is the estimation of the head count ratio (HCR), fuzzy monetary indicator (FM) and fuzzy supplementary index (FS) as non-linear parameters.

The proposed approach is based on a modified version of Empirical Best (EB) prediction proposed by Molina and Rao (2009) and it is applied to the estimation of HCR, FM and FS indexes in Tuscany provinces.

5.2 Fuzzy monetary and supplementary indicators for small areas

Let $U=\{E_1,...,E_N\}$ a population of size *N*, E_i a welfare variable (equivalised income) for individual *i*, $F_{(M),i}$ the distribution function of E_i and $L_{(M),i}$ the value of the Lorenz curve of E_i. We define the so called *Fuzzy Monetary Index* (FM_i) for individual *i* following the IFR approach (*Integrated Fuzzy and Relative Approach*, Betti *et al*. 2006) as combination of the $(1-F_{(M),i})$ indicator proposed by Cheli e Lemmi (1995) and of the $(1-L_{(M),i})$ indicator proposed by Betti and Verma (1999). Formally:

$$FM_i = \left(1 - F_{(M),i}\right)^{\alpha-1}\left(1 - L_{(M),i}\right) = \left\{\frac{1}{N-1}\sum_{j=1}^{N} \mathrm{I}\{E_j > E_i\}\right\}^{\alpha-1} \left\{\frac{\sum_{j=1}^{N} E_j \mathrm{I}\{E_j > E_i\}}{\sum_{j=1}^{N} E_j}\right\} \tag{5.2.1}$$

where $\mathrm{I}\{E_j > x\} = 1$ if $E_j > x$, 0 otherwise, $(1 - F_{(M),i})$ is the proportion of individuals less poor than the person concerned and $(1 - L_{(M),i})$ is the share of the total equivalised income received by all individuals less poor than the person concerned.

For the whole population the poverty index defined above is given by:

$$FM = \frac{1}{N}\sum_{i=1}^{N} FM_i \tag{5.2.2}$$

For each domain d ($d = 1, \ldots, D$) we define the fuzzy monetary index as:

$$FM_d = \frac{1}{N_d}\sum_{i=1}^{N_d} FM_i \tag{5.2.3}$$

Given a random sample of size $n < N$ drown from that population, $s \subseteq U$, $s = \{E_1, \ldots, E_n\}$, the direct estimator of FM_i is expressed as:

$$F\hat{M}_i^{DIR} = \left(1 - F_{(M),i}\right)^{\alpha-1}\left(1 - L_{(M),i}\right) = \left\{\frac{\sum_{j=1}^{n} w_j \mathrm{I}\{E_j > E_i\}}{\sum_{j=1}^{n} w_j}\right\}^{\alpha-1} \left\{\frac{\sum_{j=1}^{n} w_j E_j \mathrm{I}\{E_j > E_i\}}{\sum_{j=1}^{n} w_j E_j}\right\} \tag{5.2.4}$$

where w_j is the sample weight for individual *j*. The overall index for the sample population is given by:

$$F\hat{M}^{DIR} = \frac{\sum_{i=1}^{n} w_i F\hat{M}_i^{DIR}}{\sum_{i=1}^{n} w_i} \tag{5.2.5}$$

Analogously, for a domain d we can define:

$$F\hat{M}_d^{DIR} = \frac{\sum_{i=1}^{n_d} w_i F\hat{M}_i^{DIR}}{\sum_{i=1}^{n_d} w_i} \tag{5.2.6}$$

Given the recognized multidimensionality of poverty, indicators of the standard of living of households and individuals can be considered. Steps to quantify and put together these indicators are described in chapter 2.

Given a population $U = \{s_1,...,s_N\}$, the so called *Fuzzy Supplementary Index*, according to the IFR approach, can be defined as:

$$FS_i = \left(1 - F_{(s),i}\right)^{\alpha-1}\left(1 - L_{(s),i}\right) = \left\{\frac{1}{N-1}\sum_{j=1}^{N} \mathrm{I}\{s_j > s_i\}\right\}^{\alpha-1}\left\{\frac{1}{N-1}\sum_{j=1}^{N} s_j \mathrm{I}\{s_j > s_i\}\right\} \tag{5.2.7}$$

where $\mathrm{I}\{s_j > x\} = 1$ if $s_j > x$, 0 otherwise. $(1 - F_{(S),i})$ is the proportion of individuals who are less deprived than the individual concerned, $F_{(S),i}$ is the value of the score distribution function evaluated for individual *i*, $(1 - L_{(S),i})$ is the share of the total lack of deprivation score assigned to all individuals less deprived than the person concerned and $L_{(S),i}$ is the value of the Lorenz curve of score for the individual *i*.

For a whole population the supplementary index defined above is given by:

$$FS = \frac{1}{N}\sum_{i=1}^{N} FS_i \tag{5.2.8}$$

For each domain d (d = 1, …,D) we define the fuzzy monetary index as:

$$FS_d = \frac{1}{N_d}\sum_{i=1}^{N_d} FS_i \tag{5.2.9}$$

Given a random sample of size $n < N$ drown from that population, $s \subseteq U$, $s = \{s_1,...,s_n\}$, the direct estimator of FS_i is expressed as:

$$F\hat{S}_i^{DIR} = \left(1 - F_{(S),i}\right)^{\alpha-1}\left(1 - L_{(S),i}\right) = \left\{\frac{\sum_{j=1}^{n} w_j \mathrm{I}\{s_j > s_i\}}{\sum_{j=1}^{n} w_j}\right\}^{\alpha-1}\left\{\frac{\sum_{j=1}^{n} w_j s_j \mathrm{I}\{s_j > s_i\}}{\sum_{j=1}^{n} w_j s_j}\right\} \tag{5.2.10}$$

where w_j is the sample weight for individual *j*. For the whole sample population we have:

$$F\hat{S}^{DIR} = \frac{\sum_{i=1}^{n} w_i F\hat{S}_i^{DIR}}{\sum_{i=1}^{n} w_i} \qquad (5.2.11)$$

Analogously, for a domain *d* we define:

$$F\hat{S}_d^{DIR} = \frac{\sum_{i=1}^{n_d} w_i F\hat{S}_i^{DIR}}{\sum_{i=1}^{n_d} w_i} \qquad (5.2.12)$$

In all formulas, the parameter α is estimated so that the FM and FS indicators are equal to the head count ratio computed for the official poverty line (60% of the median).

5.3 Empirical Best Prediction

5.3.1 EB prediction under a finite population

Consider a random vector $\mathbf{y}$ containing the values of a random variable in the units of a finite population such that $\mathbf{y} = (\mathbf{y}_s^{'}, \mathbf{y}_r^{'})$ where $\mathbf{y}_s$ is the sub-vector of sample elements and $\mathbf{y}_r$ the sub-vector of non-sample elements. The target is to predict the value of a real measurable function $\delta = h(\mathbf{y})$ of the random vector $\mathbf{y}$ using the sample data $\mathbf{y}_s$. The best predictor (BP) of δ is the function of $\mathbf{y}_s$ that minimizes the mean square error of the predictor $\hat{\delta}$. Formally:

$$\hat{\delta}^B = \delta^0 = E_{y_r}(\delta \mid \mathbf{y}_s) \qquad (5.3.1)$$

where the expectation is taken with respect to the conditional distribution of $\mathbf{y}_r$ and the result is a function of sample data $\mathbf{y}_s$.

Generally, $\hat{\delta}^B$ depends on an unknown parameter vector $\boldsymbol{\theta}$ that can be replaced by a suitable estimator, obtaining an empirical BP of δ.

Note that, when $\mathbf{y}$ follows a Normal distribution with mean vector $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ for a known matrix $\mathbf{X}$ and positive covariance matrix $\mathbf{V}$, and the quantity to predict, δ, is a linear function of $\mathbf{y}$, then the BP of δ is equal to the BLUP of δ.

5.3.2 EB prediction of fuzzy monetary indicators

Consider the fuzzy monetary FM indicator in formula (5.2.3). The BP of FM_d is defined as:

$$F\hat{M}_d^B = E_{\mathbf{y}_r}(FM_d \mid \mathbf{y}_s) \tag{5.3.2}$$

In order to obtain the BP of FM_d, we need to express FM_d in terms of a domain vector $\mathbf{y}_d$, for which the conditional distribution of the non-sampled vector $\mathbf{y}_{dr}$ given the sample data $\mathbf{y}_{ds}$ is known. The distribution of the welfare variable E_{di} is seldom Normal, however, many times it is possible to find a transformation of the E_{di}'s whose distribution is approximately Normal. Suppose that there exists a one-to-one transformation $Y_{di} = T(E_{di})$ of the welfare variable E_{di}, which follows a Normal distribution, $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$. Let $\mathbf{y}_d = (\mathbf{y}_{ds}', \mathbf{y}_{dr}')'$ be the values of the transformed variables Y_{di} for the sample and non-sample units within domain *d*. Then we can define FM_{di} as:

$$FM_{di} = \left\{ \frac{1}{N-1} \sum_{j=1}^{N} \mathrm{I}\left\{ T^{-1}(Y_j) > T^{-1}(Y_{di}) \right\} \right\}^{\alpha-1} \left\{ \frac{\sum_{j=1}^{N} T^{-1}(Y_j) \mathrm{I}\left\{ T^{-1}(Y_j) > T^{-1}(Y_{di}) \right\}}{\sum_{j=1}^{N} T^{-1}(Y_j)} \right\} =: h_\alpha(Y_{di}) \tag{5.3.3}$$

Then, $FM_d = \frac{1}{N_d} \sum_{i=1}^{N_d} FM_{di}$ is a non-linear function of $\mathbf{y}$.

Using the decomposition of FM_d in terms of sample and non-sample elements we have:

$$FM_d = \frac{1}{N_d} \left(\sum_{i \in s_d} FM_{di} + \sum_{i \in r_d} FM_{di} \right) \tag{5.3.4}$$

Then, the BP of FM_d becomes:

$$F\hat{M}_d^B = \frac{1}{N_d} \left(\sum_{i \in s_d} FM_{di} + \sum_{i \in r_d} F\hat{M}_{di}^B \right) \tag{5.3.5}$$

where $F\hat{M}_{di}^B$ is the BP of $FM_{di} = h_\alpha(Y_{di})$ given by:

$$F\hat{M}_{di}^B = E_{\mathbf{y}_r}(h_\alpha(Y_{di}) \mid \mathbf{y}_s) = \int_R h_\alpha(y) f_{Y_{di}}(y \mid \mathbf{y}_s) dy, \quad i \in r_d \tag{5.3.6}$$

where $f_{Y_{di}}(y \mid \mathbf{y}_s)$ is the conditional distribution of Y_{di} given $\mathbf{y}_s$. Due to the complexity of the function $h_\alpha(y)$, we cannot have an explicit expression of the expectation (5.3.6). However, since $\mathbf{y} = (\mathbf{y}_s^{'}, \mathbf{y}_r^{'})$ is Normally distributed with mean vector $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_s \\ \boldsymbol{\mu}_r \end{pmatrix}$ and covariance matrix $\mathbf{V} = \begin{pmatrix} \mathbf{V}_s & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_r \end{pmatrix}$, the distribution of $\mathbf{y}_r \mid \mathbf{y}_s$ is given by:

$$\mathbf{y}_r \mid \mathbf{y}_s \sim N(\boldsymbol{\mu}_{r|s}, \mathbf{V}_{r|s}) \tag{5.3.7}$$

where $\boldsymbol{\mu}_{r|s} = \boldsymbol{\mu}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \boldsymbol{\mu}_s)$ and $\mathbf{V}_{r|s} = \mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{V}_{rs}$.

Then, we can approximate expectation (5.3.6) by Monte Carlo simulations. We generate a large number *L* of vectors $\mathbf{y}_r$ from (5.3.7) and we attach the vector $\mathbf{y}_r^{(l)}$ generated in the *l*-th replication to the sample vector $\mathbf{y}_s$ to obtain the population vector $\mathbf{y}^{(l)} = (\mathbf{y}_s^{'}, (\mathbf{y}_r^{(l)})^{'})^{'}$. Using the elements of $\mathbf{y}^{(l)}$ for the *d*-th area, we calculate the small area parameter of interest $\delta_d^{(l)} = h(\mathbf{y}_d^{(l)})$. A Monte Carlo approximation to the BP of Y_{di} is given by:

$$F\hat{M}_{di}^{B} \approx \frac{1}{L}\sum_{l=1}^{L} h_\alpha(Y_{di}^{(l)}), \quad i \in r_d \tag{5.3.8}$$

Generally, $f_{Y_{di}}(y \mid \mathbf{y}_s)$ depends on an unknown vector of parameters previously estimated using maximum likelihood (ML) or restricted ML estimator and then we obtain the EBP.

5.3.3 EB method under a nested error model

A possible model for the elements of the population vector $\mathbf{y}$ that can be used to evaluate the EBP is the nested error regression model (Battese, Harter, Fuller, 1988) that relates linearly, for all areas, the transformed population variables Y_{di} to vectors $\mathbf{x}_{di}$ of p explanatory variables and includes a random area-specific effect u_d and residual errors e_{di}. Formally:

$$\begin{aligned} &Y_{di} = \mathbf{x}_{di}^{'}\beta + u_d + e_{di}, \qquad j = 1,\dots,N_d, \quad d = 1,\dots D, \\ &u_d \sim iid\, N(0,\sigma_u^2), \quad e_{di} \sim iid\ N(0,\sigma_e^2) \end{aligned} \tag{5.3.9}$$

The vectors $\mathbf{y}_d = \underset{1\le i\le N_d}{\text{col}}(Y_{di})$ are independent with $\mathbf{y}_d \sim N(\boldsymbol{\mu}_d, \mathbf{V}_d)$, where $\boldsymbol{\mu}_d = \mathbf{X}_d\boldsymbol{\beta}$ and $\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}_{N_d}^{'} + \sigma_e^2 \mathbf{I}_N$. Then, by formula (5.3.7) we can derive the distribution of $\mathbf{y}_{dr} \mid \mathbf{y}_{ds}$ and the respective mean $\boldsymbol{\mu}_{dr|s}$ and variance $\mathbf{V}_{dr|s}$. To avoid computationally problems due to complexity of the process, instead of model (5.3.9) we can utilize the following model, noting that the conditional covariance matrix $\mathbf{V}_{dr|s}$ corresponds to the covariance matrix of the vector $\mathbf{y}_{dr}$ given by:

$$\mathbf{y}_{dr} = \boldsymbol{\mu}_{dr|s} + v_d \mathbf{1}_{N_d - n_d} + \boldsymbol{\varepsilon}_{dr}, \qquad v_d \sim N(0, \sigma_u^2(1-\gamma_d)), \quad \boldsymbol{\varepsilon}_{dr} \sim N(\mathbf{0}_{N_d - n_d}, \sigma_\varepsilon^2 \mathbf{I}_{N_d - n_d}) \tag{5.3.10}$$

where $\gamma_d = \sigma_u^2(\sigma_u^2 + \sigma_e^2 / n_d)^{-1}$ and n_d sample size in domain *d*.

5.3.4 The proposed method

Due to the greater complexity of fuzzy measures instead of traditional poverty index and then consider some computational problems we decided to modify the EB method described in the previous sections, as follows: from the original sample we draw a sample with the same size of the latter and probability proportional to sample weights. Then, this sample is representative of the whole population. At this point, we generate the $\mathbf{y}_{di}$ values as in formula (5.3.10):

$$\mathbf{y}_{di} = \boldsymbol{\mu}_{di|s} + v_d \mathbf{1}_{N_d - n_d} + \boldsymbol{\varepsilon}_{di}, \qquad v_d \sim N(0, \sigma_u^2(1-\gamma_d)), \quad \boldsymbol{\varepsilon}_{di} \sim N(\mathbf{0}_{N_d - n_d}, \sigma_\varepsilon^2 \mathbf{I}_{N_d - n_d}) \tag{5.3.11}$$

replacing the parameters with their estimates. Following the steps described in section 5.3.2, we obtain a new “direct” estimator which is representative of the whole population. As showed in the next section, a model-based simulation study has been carried out to study the performance of the proposed method of small domain traditional (HCR) and fuzzy monetary (FM) index. Due to the same computational problems, for HCR we report results of simulation in which we compare direct estimators, original EB estimators and new EB estimators, whereas for FM index we restrict to direct estimators and new EB estimators. As we can see by the results, the new method keeps similar properties of the standard EB, but it allows to overcome computational problems due to big populations or to more complex poverty measures, like FM index.

5.3.5 Parametric bootstrap for MSE estimation

The MSE of the EB estimator $F\hat{M}_d^{EB}$ with respect to the model is given by:

$$MSE(F\hat{M}_d^{EB}) = E(F\hat{M}_d^{EB} - FM_d) = V(F\hat{M}_d^{EB} - FM_d) + \left[E(F\hat{M}_d^{EB} - FM_d)\right]^2 \tag{5.3.12}$$

Because of the difficulty to calculate this expression for poverty measures analytically, we obtain a parametric bootstrap MSE estimator as described in Molina and Rao (2009). This method implies the following steps:

1. Fit model 5.3.9 to sample data $(\mathbf{y}_s, \mathbf{X}_s)$ and obtain estimators $\hat{\beta}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ of β, σ_u^2 and σ_e^2 respectively, using a suitable method (for example REML method).
2. Generate $u_d^* \sim \text{iid}\, N(0, \hat{\sigma}_u^2),\ d = 1,...,D$, and independently, generate $e_{di}^* \sim \text{iid}\, N(0, \hat{\sigma}_e^2),\ i = 1,...,N_d$.
3. Construct the bootstrap super-population model using u_d^*, e_{di}^*, $\mathbf{x}_{di}$ and β:

   $$Y_{di}^* = \mathbf{x}_{di}^{'}\beta + u_d^* + e_{di}^* \qquad (5.3.13)$$
4. Under the bootstrap super-population model 5.3.13, generate a large number *B* of independent and identically distributed bootstrap populations $Y_{di}^{*(b)}$ and calculate bootstrap population parameters $FM_d^{*(b)}$, *b* = 1,…,*B*.
5. From each bootstrap population *b* generated in step 4, take the sample with the same indices as the initial sample and calculate the bootstrap EBPs, $F\hat{M}_d^{EB*(b)}$ as described in section 5.3.2 and using bootstrap sample data $\mathbf{y}_s^*$ and the known population values $\mathbf{x}_{di}$.
6. A Monte Carlo approximation to the theoretical bootstrap MSE estimator of $F\hat{M}_d^{EB}$ is given by:

   $$mse_*(F\hat{M}_d^{EB}) = \frac{1}{B}\sum_{b=1}^{B}(F\hat{M}_d^{EB*(b)} - F\hat{M}_d^{*(b)})^2 \qquad (5.3.14)$$

5.4. Model-based simulation experiment

To study the performance of the proposed new EB estimators, we simulated populations of size *N* = 20000, composed of *D* = 80 areas with $N_d = 250$ elements in each area *d* = 1, .., *D*. The response variables for the population units Y_{dj} were generated from the model (5.3.9) taking as auxiliary variables two dummies $X_1 \in \{0,1\}$ and $X_2 \in \{0,1\}$ plus an intercept. The values of these two dummies for the population units were generated from Bernoulli distributions with success probabilities increasing with the area index for X_1 and constant for X_2. Formally we have respectively:

$$p_{1d} = 0.3 + 0.5d/80, \quad p_{2d} = 2, \qquad d = 1,...,D \qquad (5.4.1)$$

The welfare variables are the exponential of the model responses, then we consider a log transformation. A set of sample indices s_d with $n_d = 50$ was drawn independently in each area *d* using simple random sampling without replacement. The values of the auxiliary variables for the population units and the sample indices were kept fixed over all Monte Carlo simulations. The intercept and the regression coefficients associated with the two auxiliary variables to generate population were $\beta = (3, 0.03, -0.04)'$. The random area effects variance was taken as $\sigma_u^2 = (0.15)^2$ and the error variance as $\sigma_e^2 = (0.5)^2$. The poverty line z was fixed as z = 12, which is equal to 0.6 times the median of the welfare variables for a given generated population. We considered I = 1000 Monte Carlo simulations. Then, I population vectors $\mathbf{y}^{(i)}$ were generated from the true model and for each population i, we carried out the following steps:

i. The true area poverty incidence $\left(HCR_d^{(i)} = \frac{1}{N_d} \sum_{j=1}^{N_d} I\left(E_{dj}^{(i)} < z\right),\ E_{dj}^{(i)} = \exp(Y_{dj}^{(i)}) \right)$ and fuzzy monetary indicator $\left(FM_d^{(i)} = \frac{1}{N_d} \sum_{j=1}^{N_d} FM_{dj}^{(i)} \right)$ were obtain for each population.

ii. Direct estimators of these poverty measures were calculated using the sample part of the i-th population vector $\mathbf{y}_s^{(i)}$.

iii. The nested-error model given in (5.3.9) was fitted to sample data $\left(\mathbf{y}_s^{(i)}, \mathbf{X}_s\right)$ and parameters were substituted by their estimates.

iv. L = 50 non-sampled vectors $\mathbf{y}_r^{(il)}, l = 1,...,L$ were generated from the conditional distribution (5.3.7) using (5.3.10) and the population vector $\mathbf{y}^{(il)}$ was formed attaching the sample data $\mathbf{y}_s^{(i)}$ to the generated non-sample data $\mathbf{y}_r^{(il)}$. Then the Monte Carlo approximations to the EBPs of poverty measures were calculated.

v. From the original sample, a sample with the same size of the closer and probability proportional to sample weights was drawn. L = 50 $\mathbf{y}_{di}$ values were generated from (5.3.11) and the Monte Carlo approximations to the new EBPs poverty measures were calculated.

vi. Means over Monte Carlo populations of the true values of the poverty measures, biases and MSEs over Monte Carlo populations i = 1, …, I of the three estimators were computed.

vii. ELL estimators (Elbers *et al.*, 2003) of the poverty measures were also calculated. Firstly, model (5.3.9) was fitted to sample data $\mathbf{y}_s$ and then A = 50 censuses were generated using parametric bootstrap algorithm (for details see Molina and Rao, 2009). For each population, the poverty measures were calculated and the results were averaged over the A populations.

As explained above, due to computational problems, steps iv and vii weren't computed for FM index.

Figures 5.4.1, 5.4.2 and 5.4.3 show respectively the trends, the biases and the MSEs of the estimators for the HCR. The true values and the four estimators have the same absolute values. We can see that performances of the standard EB estimators and of the new EB estimators are very similar, then the new method doesn't make loss of efficiency. Moreover, biases are not significant different in absolutes values among estimators (figure 5.4.2), but figure 5.4.3 shows, for EB estimators and new EB estimators, big improvements in mean squared error over direct estimators and estimators obtained by simulated censuses (ELL).

Figure 5.4.1. Trend over simulated populations of true values, EB, direct, ELL and new EB estimators of HCR for each area d

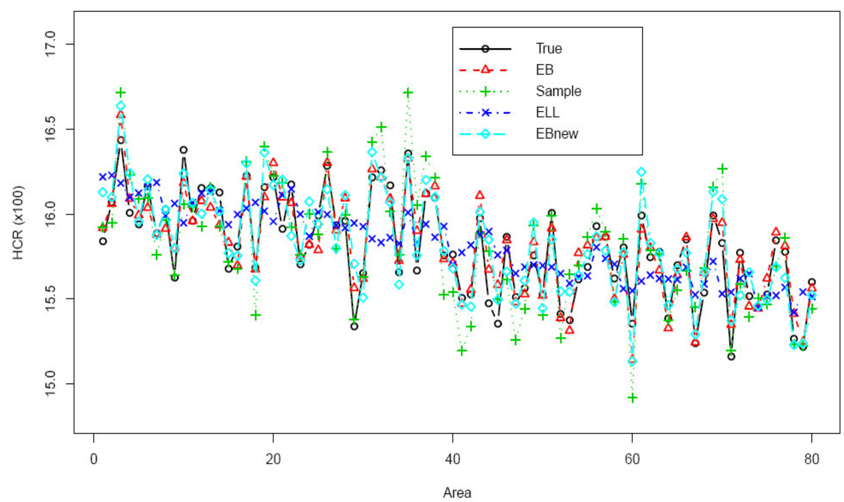


Figure 5.4.2. Bias ($\times 100$) over simulated populations of EB, direct, ELL and new EB estimators of HCR for each area d

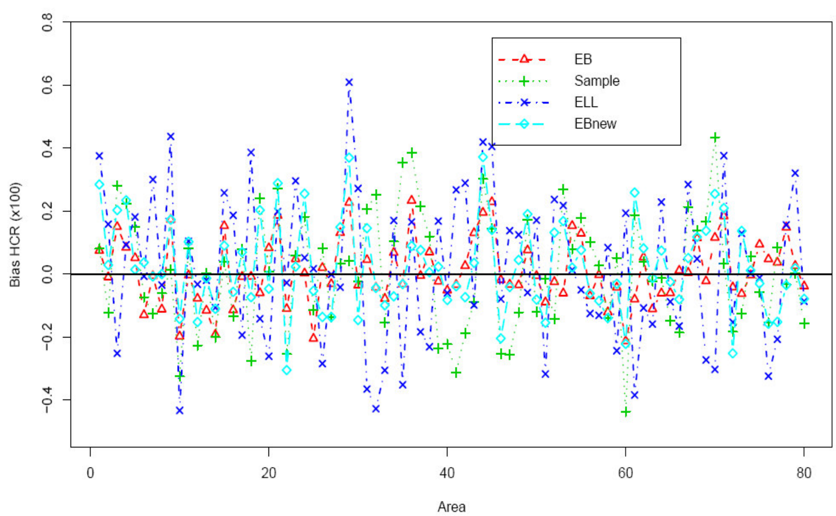
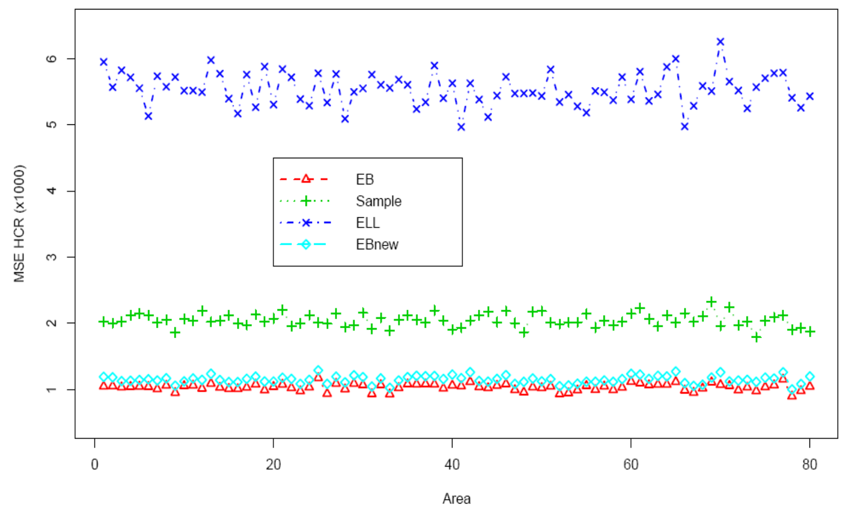


Figure 5.4.3. MSE ($\times 1000$) over simulated populations of EB, direct, ELL and new EB estimators of HCR for each area d



Analogously, figures 5.4.4, 5.4.5 and 5.4.6 show respectively the trends, the biases and the MSEs of the direct estimators and the new EB estimators for the FM index.

As for HCR, the true values and the two estimators have the same absolute values. Moreover, biases are not significant different in absolutes values among estimators (figure 5.4.5), but figure 5.4.6 shows for the new EB estimators, improvements in mean squared error over direct estimators.

Figure 5.4.4. Trend over simulated populations of true values, new EB and direct estimators of FM for each area d

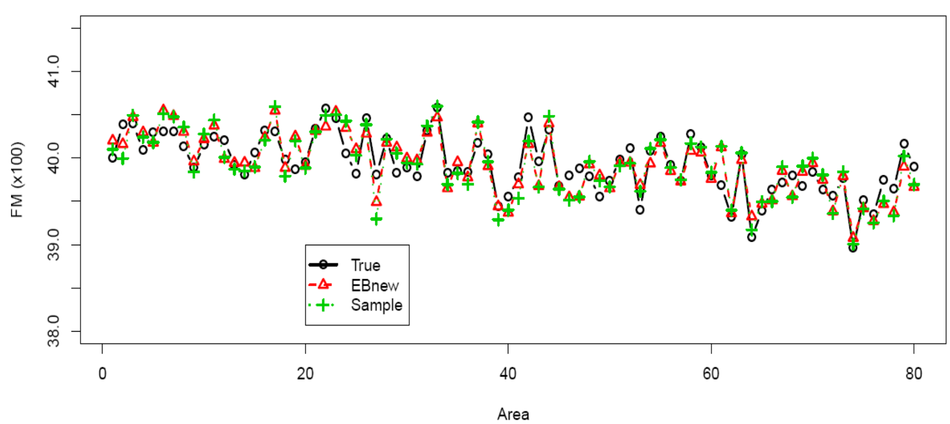


Figure 5.4.5. Bias ($\times 100$) over simulated populations of new E and direct estimators of FM for each area d

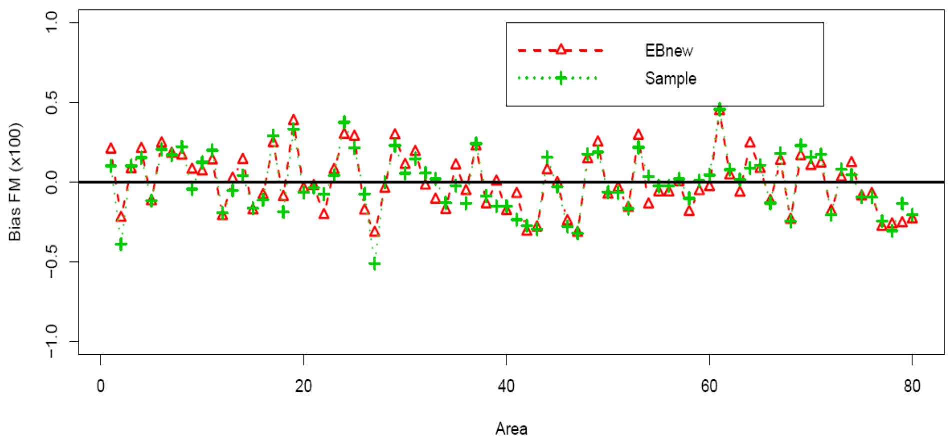
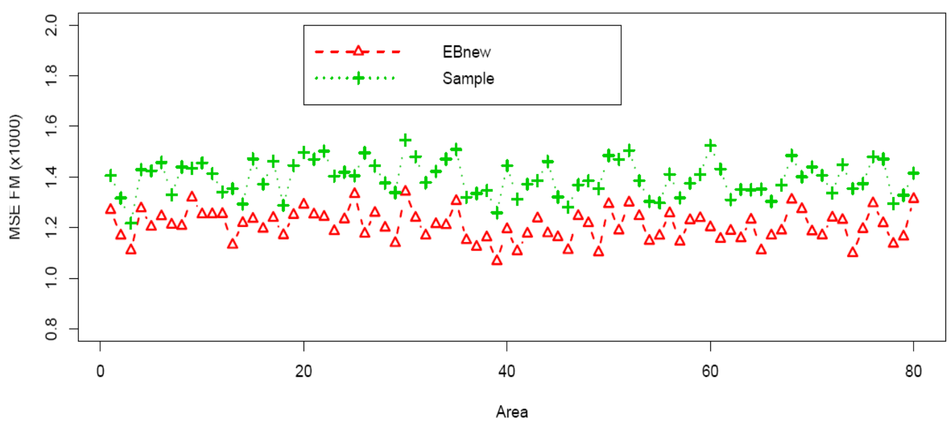


Figure 5.4.6. MSE ($\times 1000$) over simulated populations of new EB and direct estimators of FM for each area d



5.5 Application to Tuscany data

The modified EB method described in the previous section was applied to estimate head count ratio (HCR), fuzzy monetary (FM) and supplementary (FS) indicators in Tuscany provinces. Data from 2004 EU-SILC survey was used.

The regional sample in Tuscany is based on a stratified two stage sample design: in each province the municipalities are the primary sampling units (PSUs) divided into strata according to their dimension in terms of population size from which the households (SSUs) are selected by means of systematic sampling. Some provinces, generally the smaller ones, may have very few sampled municipalities and many municipalities are not even included in the sample at all. For example in 2004 survey only 53 municipalities out of 287 are present. Then, small area estimation techniques can be required given large errors of direct estimators at province level or the impossibility to compute them at municipalities level.

In our analysis the small areas of interest are the 10 Tuscany provinces, with sample sizes ranging from 155 (Province of Grosseto) to 1403 (Province of Firenze). The regional sample size is of 4426 individuals.

The welfare variable for the individuals is the equivalised annual net income. In order to overcome the problem of negative values of this variable, we followed the recommendation by Eurostat concerning this topic: a bottom coding strategy to the lowest values of the distribution have been applied. In particular, all values below 15% of the median household income have been set equal to the 15% of the median. This strategy has not effects on the poverty line and then on the direct estimators (Eurostat, 2006; Ciampalini *et al*., 2009; Neri *et al*., 2009). The equivalised annual net income has been transformed by taking logarithm to obtain a distribution approximately normal. This transformed variable acts as the response in the nested-error regression model (5.3.9). As auxiliary variables we have considered the indicators of 5 quinquennial groupings of variable age, the indicator of having Italian nationality, the indicators of 3 levels of the variable education level and 3 categories of the variable employment.

The poverty line for the calculation of HCR is computed as the 60% of the weighted median of the individual equivalent income at Regional level and is equal to 9,372.24 Euros.

Direct estimators and new EB estimators were calculated for HCR, FM and FS indicators. In the present analysis, we fixed the parameter alpha equal to 2, then we avoid any numerical link to the traditional approach. This is because the primary objective of this analysis is to develop methodologies for estimating fuzzy measures in small domains, rather than numerical comparisons with the conventional approach.

Values of direct estimators, new EB estimators of HCR and their associated coefficients of variation (CV) are shown in table 5.5.1 for each Tuscany provinces. The average over provinces is 16.4%. The poorer provinces concentrate mainly in the north-west of Tuscany. Province of Massa has the highest percentage of poor individuals (22.4%) followed by Lucca (18.2%) and Pisa (17.8%). On the other hand, Province of Arezzo (13.0%) and Province of Firenze (14.4) are the most rich. The MSEs of new EB estimators of HCR are calculated using the parametric bootstrap estimator 5.3.14 with B = 500. The coefficient of variation is given by $\mathrm{cv}(H\hat{C}R_d^{newEB}) = \{\mathrm{mse}(H\hat{C}R_d^{newEB})^{1/2} \big/ H\hat{C}R_d^{newEB}\}$. Results in table 5.5.1 show that the CVs of new EB estimators are much smaller than those of direct estimators and the reduction in CV tends to be greater for domains with smaller sample size. The only exception is Province of Firenze with a large sample size, for which the CVs of new EB estimators are much bigger than those of direct estimators. Due to computational problems, we couldn't calculate MSEs of new EB estimators of fuzzy poverty measures.

Table 5.5.1. Population size, sample size, direct and new EB estimators of HCR, CVs of direct and new EB estimators (x100) for Tuscany Provinces

Provinces	Population size	Sample size	$\hat{HCR}_d^{DIR}$	$\hat{HCR}_d^{newEB}$	cv $\hat{HCR}_d^{DIR}$	cv $\hat{HCR}_d^{newEB}$
Arezzo	304121	416	0.087	0.130	19.09	12.42
Firenze	1119377	1403	0.133	0.144	8.32	10.89
Grosseto	149082	155	0.124	0.147	26.10	11.10
Livorno	290122	339	0.131	0.149	15.61	10.30
Siena	278495	338	0.110	0.156	19.31	10.80
Prato	319320	416	0.170	0.159	14.06	10.77
Pistoia	267076	344	0.174	0.169	13.75	9.87
Pisa	335777	399	0.168	0.178	15.54	8.88
Lucca	265293	315	0.215	0.182	13.30	8.88
Massa Carrara	251471	301	0.260	0.224	13.09	7.30
average			**0.157**	**0.164**		

Table 5.5.2 shows respectively direct and new EB estimators of FM and FS indicators and the combination of two.

The new EB estimators of FM index give the same indication about the monetary poverty in the small areas of HCR and also the difference between provinces is similar according to either approach. The values of FM are bigger than those ones of HCR since there is a certain concentration in each province of individuals with equivalised income just above the poverty line. Province of Arezzo (36.5%) and Firenze (38.0%) remain the most rich, whereas on the other hand, Province of Massa has the highest percentage of poor individuals (47.5%) followed by Lucca (42.6%) and Pisa (42.3%).

Moreover, the overall fuzzy supplementary index has been calculated for Tuscany provinces. In this case the welfare variable is the overall score (constructed as explained in chapter 2) and as the response in the nested-error regression model we took the clog-log transformation. We used the same auxiliary variables employed for the calculation of HCR and FM indexes.

Concerning non-monetary deprivation, the ranking of some provinces is completely opposite respect to monetary poverty (table 5.5.2). For example, Province of Massa have high values of FM and low values of FS, whereas the opposite holds for Firenze and Livorno Provinces.

Non-monetary dimension is combined with the monetary dimension in order to obtain measures of manifest (MAN) and latent (LAT) deprivation which correspond respectively to intersection and union of the fuzzy sets. We can note some differences between provinces (table 5.5.2). The overlapping takes the lowest value of 27.6% for Grosseto and the biggest one of 39.4% for Pistoia. In general, the MAN/LAT ratio is lower in areas with lower levels of deprivation, and higher in areas with higher levels. High values of this ratio imply that different types of deprivation overlap and this means that in areas where levels of relative deprivation are already high, deprivation in the income and non-monetary domains is more likely to afflict the same individuals in the population. On the other hand, low values imply the absence of such overlap at micro level.

Table 5.5.2 .Sample size, direct and new EB estimators of FM and FS indicators, latent (LAT) and manifest (MAN) deprivation, ratio MAN/LAT for Tuscany Provinces

Provinces	Sample size	$F\hat{M}_d^{\alpha DIR}$	$F\hat{M}_d^{\alpha newEB}$	$F\hat{S}_d^{\alpha DIR}$	$F\hat{S}_d^{\alpha newEB}$	LAT	MAN	MAN/LAT
Arezzo	416	0.354	0.365	0.262	0.297	0.494	0.167	0.338
Firenze	1403	0.376	0.380	0.371	0.368	0.542	0.206	0.380
Grosseto	155	0.390	0.383	0.158	0.203	0.460	0.127	0.276
Livorno	339	0.379	0.392	0.370	0.372	0.552	0.212	0.384
Siena	338	0.381	0.396	0.306	0.321	0.526	0.191	0.363
Prato	416	0.402	0.404	0.332	0.347	0.545	0.206	0.377
Pistoia	344	0.424	0.414	0.411	0.380	0.570	0.225	0.394
Pisa	399	0.433	0.423	0.353	0.348	0.558	0.212	0.381
Lucca	315	0.424	0.426	0.398	0.360	0.566	0.220	0.388
Massa Carrara	301	0.496	0.475	0.330	0.316	0.578	0.213	0.368
average		**0.406**	**0.406**	**0.329**	**0.331**			

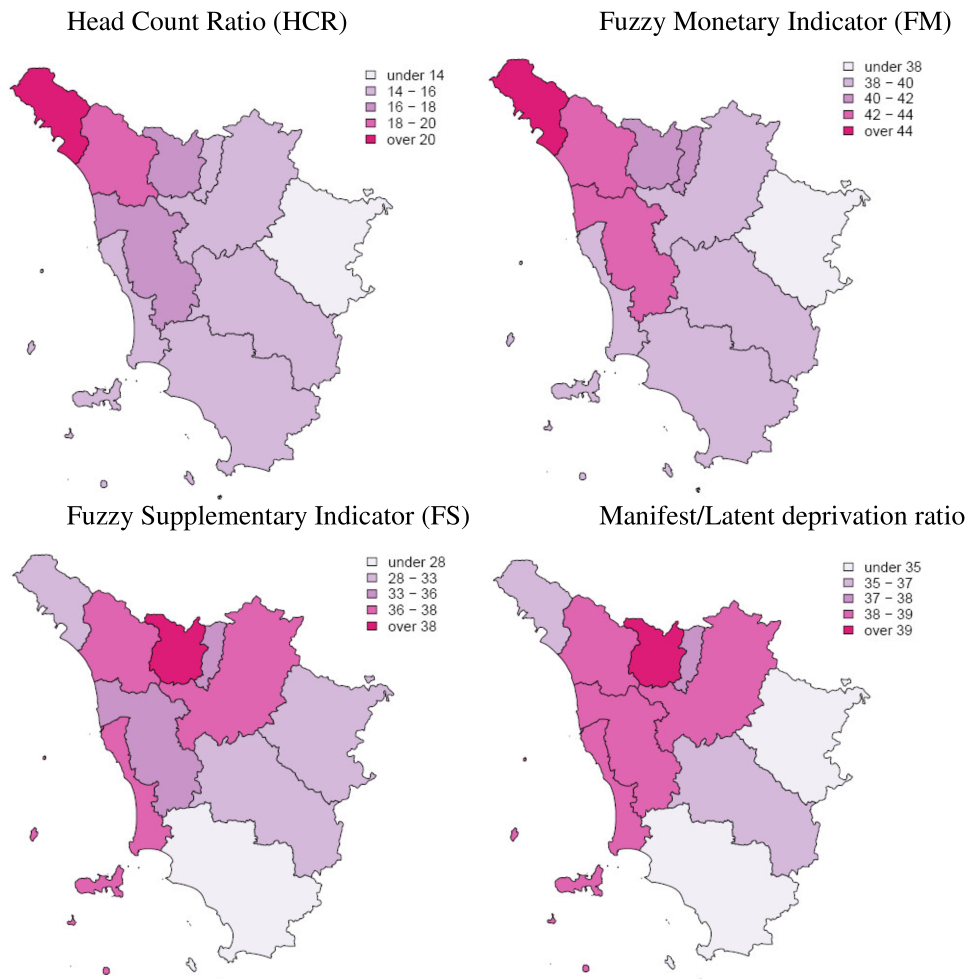
Figure 5.1 shows respectively the cartograms of the estimated head count ratio, fuzzy monetary indicators, fuzzy supplementary indicators and manifest/latent ratio in Tuscany provinces constructed using the new EB estimators.

5.6. Further Researches

Application of EB method to Tuscany data showed some limitations of this method where some assumptions are not respected. For example, we noted that not perfect normality of distribution causes a bias in the EB estimator. So the EB method deserves further research when the underlying distribution is not normal and we don't find a transformation to make it normal. Moreover, mean square errors of new EB estimators of fuzzy poverty measures have to be calculated.

Concerning fuzzy measures, as explained before, in our analysis the parameter alpha has been fixed equal to 2 don't making possible any comparison with the traditional approach. The idea is to determine alpha in order to make FM and FS indicators numerically identical to the head count ratio. Moreover, further research is requested to obtain a transformation of each score in order to calculate FS indicators for each dimension.

Figure 5.1. Cartograms of estimated percent Head count Ratio, Fuzzy Monetary indicators, Supplementary indicators and Manifest/Latent deprivation ratio in Tuscany Provinces.



Conclusions

This work takes place in a European Project inside the Seventh Framework Programme (S.A.M.P.L.E.) aimed at making a decisive impact on the eradication of poverty in European Countries by the year 2010, as established in Lisbon European Council (2000).

As well as treating poverty as a fuzzy state instead as a simple dichotomy poor non-poor, i.e. an attribute that characterizes an individual in terms of presence or absence, it highlights the complexity of this phenomenon and then, the demand of constructing indicators that take in account not only the income factor but also the multidimensional nature of poverty.

For this purpose, chapter 1, after a brief review of traditional poverty measures and its limitations, presents eight different multidimensional approaches focusing on the fuzzy set one, first proposed by Cerioli and Zani (1990).

In these last years, the multidimensional nature of poverty took a large importance at international level. One of the most recent examples is the Report by the Commission on the Measurement of Economic Performance and Social Progress (Commission Stiglitz-Sen-Fitoussi), demanded by French President Nicolas Sarkozy in 2008 in order to identify main causes of the growing divergence between current measures of economic performance and people's perceptions about the quality of their life. The Nobel Laureates highlight the importance to provide new measures of social well-being rather than the only GDP that is mainly a measure of market production and not of economic well-being. Conventional economic indicators have to be enriched by measures of quality of life, both objective and subjective, that, for their multidimensional nature, can be grouped in eight different dimensions: material living standard, health, education, personal activities including work, political voice and governance, social connections and relationships, environment and finally, insecurity. The third target of the Commission is measuring and assessing sustainability, i.e. "determining if at least the current level of well-being can be maintained for future generations", emphasizing the environmental aspects of sustainability.

In this work, we proposed the Integrated Fuzzy and Relative (IFR) approach, that combines the TFR approach of Cheli and Lemmi (1995) and the approach of Betti and Verma (1999). Following this approach, the fuzzy monetary indicator (FM) has been defined as the product of the share of individuals less poor than the person concerned and the share of the total equivalised income received by all individuals less poor than the person concerned. Fuzzy supplementary indicator has been determined, step by step, using the same membership function as that used for income poverty. Seven dimensions of deprivation have been constructed using a factor analysis and quantitative measures have been determined for each of them. Remark that, these dimensions, basic life-style, consumer durables, housing amenities, financial situation, environmental problems, work-education and health, are comparable with the groups of indicators defined in the Report of the Commission Stiglitz-Sen-Fitoussi.

Fuzzy monetary and non-monetary indicators and their overlapping degree have been calculated in a cross-sectional analysis from year 2004 to year 2007 in the European Members States using EU-SILC data. Results showed large differences among Countries, but a certain stability over years.

Subsequently, we applied the proposed monetary and non-monetary indicators at local level following two different approaches. The first one deals with pooling of data or estimates in order to obtain measures which convey essentially the same information as the "original" un-pooled measures, but reducing variability or noise. Moreover, pooling aims to trade dimensions such as permitting more detailed geographical breakdown but with less temporal details and to summarize over different dimensions, providing more consolidated and fewer indicators. Four different scenarios have been determined depending on whether the data sources and populations involved are same/similar or different. Due to the lack of all data required, in this work we didn't report any empirical results, but only methodological aspects, using examples of different scenarios that can arise. Simple model have been developed to illustrate the effect on variance of pooling over correlated samples, such as over waves in a rotational panel design.

The second method concerns small area estimation techniques following the approach of Molina and Rao (2009). The proposed new Empirical Best predictors allow to estimate poverty measures, as non-linear domain parameters, for small areas overcoming computational problems due to big populations or more complex poverty measures, as fuzzy indicators. Simulations results show good performance of new EB estimators in comparison with direct and ELL estimators. Results of traditional and fuzzy poverty measures have been obtained for Tuscany provinces in 2004. The poorer provinces concentrate mainly in the north-west of Tuscany. Province of Massa has the highest percentage of poor individuals followed by Lucca and Pisa. On the other hand, Province of Arezzo and Province of Firenze are the most rich.

A parametric bootstrap method has been used for mean square error estimation of HCR: it confirms a big improvement of new EB estimators in comparison with direct estimators. Due to computational problems, in this work we didn't calculate MSEs of new EB estimators of fuzzy measures, then further researches are required to overcome them. We noted that new EB is a model-based method that relies on the validity of the model. Model selection procedures and model diagnostics are essential in the practical application of this model and further researches are required when the underlying distribution is not normal or we don't find a transformation to make it normal. Moreover, some improvements are required in the calculation of parameter alpha in order to make FM and FS indicators numerically identical to head count ratio and in the determination of FS index for each dimensions of deprivation.

Annex

List of EU-SILC variables used to construct Fuzzy Supplementary Indicators.

HH040: Leaking Roof, Damp Walls/Floors/Foundation, Or Rot In Window Frames Or Floor
HH050: Ability To Keep Home Adequately Warm
HH080: Bath Or Shower In Dwelling
HH090: Indoor Flushing Toilet For Sole Use Of Household
HS010: Arrears On Mortgage Or Rent Payments
HS020: Arrears On Utility Bills
HS030: Arrears On Hire Purchase Instalments Or Other Loan Payments
HS040: Capacity To Afford Paying For One Week Annual Holiday Away From Home
HS050: Capacity To Afford A Meal With Meat, Chicken, Fish (Or Vegetarian Equivalent) Every Second Day
HS060: Capacity To Face Unexpected Financial Expenses
HS070: Do You Have A Telephone (Including Mobile Phone)?
HS080: Do You Have A Colour Tv?
HS090: Do You Have A Computer?
HS100: Do You Have A Washing Machine?
HS110: Do You Have A Car?
HS120: Ability To Make Ends Meet
HS160: Problems With The Dwelling: Too Dark, Not Enough Light
HS170: Noise From Neighbours Or From The Street
HS180: Pollution, Grime Or Other Environmental Problems
HS190: Crime Violence Or Vandalism In The Area
PE010: Current Education Activity
PE040: Highest Isced Level Attained
PH010: General Health
PH020: Suffer From Any A Chronic (Long-Standing) Illness Or Condition
PH030: Limitation In Activities Because Of Health Problems
PH040: Unmet Need For Medical Examination Or Treatment
PH060: Unmet Need For Dental Examination Or Treatment
PL030: Self-Defined Current Economic Status
PL070: Number Of Months Spent At Full-Time Work
PL072: Number Of Months Spent At Part-Time Work
PL080: Number Of Months Spent In Unemployment
PL085: Number Of Months Spent In Retirement
PL087: Number Of Months Spent Studying
PL090: Number Of Months Spent In Inactivity

The following tables show fuzzy measures of monetary poverty, non-monetary deprivation results and their combination respectively for 2006, 2005 and 2004 waves in European Countries.

Table 1. Fuzzy measures at Country level (2006)

	Rate of deprivation by dimension of deprivation								**mean**
Country	**FS0**	**FS1**	**FS2**	**FS3**	**FS4**	**FS5**	**FS6**	**FS7**	**FS1-FS7**
IS	0.096	0.083	0.030	0.042	0.078	0.064	0.080		0.063
CZ	0.098	0.100	0.064	0.053	0.058	0.115	0.089	0.089	0.081
NL	0.098	0.081	0.041	0.059	0.050	0.098	0.085		0.069
NO	0.108	0.080	0.044	0.052	0.078	0.077	0.088		0.070
SK	0.116	0.109	0.068	0.064	0.069	0.111	0.103	0.112	0.091
DK	0.116	0.095	0.059	0.061	0.059	0.097	0.094		0.078
SI	0.117	0.096	0.054	0.071	0.074	0.104	0.101		0.083
SE	0.122	0.095	0.042	0.061	0.084	0.093	0.100		0.079
FI	0.125	0.095	0.067	0.062	0.078	0.109	0.106		0.086
AT	0.126	0.101	0.055	0.073	0.049	0.104	0.110	0.090	0.083
DE	0.126	0.108	0.059	0.069	0.060	0.128	0.107	0.118	0.093
FR	0.131	0.101	0.064	0.076	0.080	0.122	0.110	0.105	0.094
LU	0.141	0.094	0.026	0.074	0.060	0.136	0.113	0.108	0.087
BE	0.147	0.130	0.070	0.087	0.083	0.137	0.124	0.094	0.104
CY	0.158	0.134	0.062	0.071	0.120	0.139	0.130	0.138	0.113
HU	0.159	0.162	0.106	0.104	0.088	0.130	0.134	0.167	0.127
EE	0.183	0.121	0.114	0.143	0.089	0.157	0.149	0.171	0.135
PT	0.185	0.137	0.117	0.120	0.094	0.162	0.154	0.158	0.135
IE	0.185	0.135	0.094	0.090	0.093	0.139	0.148	0.127	0.118
PL	0.191	0.231	0.114	0.123	0.107	0.144	0.158	0.189	0.152
UK	0.192	0.145	0.066	0.101	0.105	0.166	0.147	0.143	0.125
IT	0.196	0.159	0.068	0.099	0.118	0.185	0.158	0.170	0.137
ES	0.198	0.151	0.081	0.105	0.090	0.181	0.164	0.163	0.134
LT	0.199	0.186	0.137	0.164	0.087	0.141	0.160	0.195	0.153
GR	0.205	0.163	0.109	0.114	0.150	0.159	0.162	0.170	0.147
LV	0.231	0.247	0.149	0.195	0.101	0.217	0.181	0.268	0.194
average	0.152	0.128	0.075	0.090	0.085	0.131	0.125	0.146	0.109

NOTES FS0 stands for "HCR = FM = FS"

FS1 - FS7 refer to the seven dimensions of deprivation defined in section 2.4.6.

Table 2. "Normalised" Fuzzy measures at Country level (2006)

	'Normalised rates'								**mean**
Country	**FS0**	**FS1**	**FS2**	**FS3**	**FS4**	**FS5**	**FS6**	**FS7**	**FS1-FS7**
IS	1.00	1.03	0.62	0.74	1.45	0.77	1.02		0.937
CZ	1.00	1.21	1.32	0.91	1.06	1.36	1.10	0.94	1.130
NL	1.00	0.97	0.84	1.01	0.91	1.15	1.05		0.988
NO	1.00	0.87	0.82	0.81	1.29	0.83	0.99		0.934
SK	1.00	1.11	1.17	0.94	1.06	1.10	1.07	1.00	1.065
DK	1.00	0.97	1.03	0.89	0.92	0.97	0.98		0.958
SI	1.00	0.97	0.93	1.03	1.13	1.03	1.05		1.025
SE	1.00	0.92	0.69	0.84	1.23	0.88	1.00		0.925
FI	1.00	0.90	1.08	0.84	1.12	1.01	1.03		0.996
AT	1.00	0.95	0.88	0.98	0.70	0.96	1.06	0.75	0.897
DE	1.00	1.01	0.94	0.93	0.85	1.18	1.03	0.97	0.987
FR	1.00	0.92	0.98	0.99	1.10	1.08	1.02	0.84	0.988
LU	1.00	0.79	0.38	0.89	0.77	1.11	0.98	0.80	0.816
BE	1.00	1.05	0.96	1.00	1.02	1.08	1.02	0.67	0.971
CY	1.00	1.00	0.80	0.76	1.36	1.02	1.00	0.91	0.979
HU	1.00	1.20	1.34	1.11	0.99	0.95	1.02	1.09	1.098
EE	1.00	0.78	1.25	1.32	0.87	0.99	0.99	0.97	1.024
PT	1.00	0.88	1.27	1.10	0.92	1.01	1.01	0.89	1.013
IE	1.00	0.86	1.03	0.83	0.91	0.87	0.97	0.71	0.883
PL	1.00	1.43	1.21	1.09	1.01	0.87	1.00	1.03	1.092
UK	1.00	0.89	0.69	0.89	0.99	1.00	0.93	0.78	0.881
IT	1.00	0.96	0.70	0.85	1.08	1.09	0.98	0.90	0.937
ES	1.00	0.90	0.82	0.90	0.81	1.05	1.00	0.85	0.906
LT	1.00	1.11	1.39	1.40	0.79	0.82	0.97	1.02	1.071
GR	1.00	0.94	1.07	0.94	1.31	0.90	0.96	0.86	0.998
LV	1.00	1.26	1.30	1.43	0.78	1.09	0.95	1.21	1.146
average	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.000

NOTES 'Normalised rates' N_{ij} : all values scaled such that:

(1) for each dimension (j), average over countries rescaled to = 1.0; and
(2) for each country (i), FS_j values scaled to correspond to FS0 = 1.0.

$$N_{ij} = \left(\frac{FS_{ij}}{FS_{.j}}\right) \Big/ \left(\frac{FS_{i0}}{FS_{.0}}\right)$$

Table 3. Latent and Manifest deprivation at aggregated level and for each dimension of deprivation (2006)

Country	FS0	L0	M0	*M0/L0*	L1	M1	*M1/L1*	L2	M2	*M2/L2*	L3	M3	*M3/L3*	L4	M4	*M4/L4*	L5	M5	*M5/L5*	L6	M6	*M6/L6*	L7	M7	*M7/L7*
IS	0.096	0.165	0.027	*0.167*	0.159	0.021	*0.130*	0.117	0.009	*0.073*	0.128	0.011	*0.084*	0.153	0.020	*0.133*	0.147	0.013	*0.088*	0.161	0.016	*0.098*			
CZ	0.098	0.157	0.039	*0.250*	0.163	0.035	*0.217*	0.132	0.030	*0.229*	0.133	0.017	*0.130*	0.133	0.024	*0.178*	0.191	0.022	*0.117*	0.151	0.036	*0.235*	0.170	0.017	*0.100*
NL	0.098	0.166	0.031	*0.187*	0.148	0.032	*0.215*	0.123	0.016	*0.134*	0.144	0.013	*0.093*	0.131	0.017	*0.127*	0.177	0.019	*0.105*	0.155	0.028	*0.178*			
NO	0.108	0.181	0.036	*0.198*	0.159	0.030	*0.187*	0.135	0.018	*0.131*	0.147	0.013	*0.091*	0.163	0.023	*0.142*	0.170	0.016	*0.094*	0.170	0.026	*0.154*			
SK	0.116	0.193	0.039	*0.203*	0.189	0.036	*0.191*	0.157	0.027	*0.173*	0.161	0.019	*0.120*	0.163	0.022	*0.135*	0.207	0.019	*0.093*	0.179	0.039	*0.219*	0.204	0.024	*0.118*
DK	0.116	0.192	0.040	*0.209*	0.176	0.035	*0.201*	0.148	0.028	*0.186*	0.163	0.015	*0.090*	0.152	0.024	*0.157*	0.192	0.021	*0.112*	0.185	0.025	*0.134*			
SI	0.117	0.193	0.040	*0.206*	0.174	0.038	*0.221*	0.147	0.024	*0.164*	0.166	0.022	*0.132*	0.169	0.022	*0.128*	0.202	0.018	*0.091*	0.180	0.038	*0.212*			
SE	0.122	0.203	0.042	*0.204*	0.180	0.037	*0.207*	0.146	0.018	*0.123*	0.168	0.015	*0.091*	0.175	0.031	*0.180*	0.195	0.021	*0.107*	0.194	0.028	*0.146*			
FI	0.125	0.205	0.045	*0.219*	0.177	0.043	*0.243*	0.157	0.035	*0.220*	0.171	0.016	*0.093*	0.173	0.030	*0.171*	0.210	0.024	*0.113*	0.195	0.036	*0.183*			
AT	0.126	0.206	0.045	*0.219*	0.185	0.041	*0.221*	0.158	0.023	*0.143*	0.177	0.021	*0.119*	0.153	0.022	*0.144*	0.208	0.021	*0.101*	0.195	0.040	*0.206*	0.192	0.024	*0.122*
DE	0.126	0.207	0.045	*0.218*	0.190	0.044	*0.233*	0.161	0.024	*0.149*	0.177	0.019	*0.107*	0.167	0.019	*0.113*	0.225	0.029	*0.131*	0.195	0.039	*0.198*	0.212	0.032	*0.151*
FR	0.131	0.208	0.054	*0.258*	0.187	0.045	*0.240*	0.166	0.028	*0.169*	0.183	0.024	*0.130*	0.180	0.031	*0.174*	0.222	0.031	*0.138*	0.197	0.044	*0.222*	0.202	0.034	*0.170*
LU	0.141	0.223	0.058	*0.259*	0.181	0.054	*0.298*	0.152	0.015	*0.099*	0.191	0.023	*0.121*	0.164	0.037	*0.225*	0.240	0.036	*0.152*	0.205	0.049	*0.240*	0.214	0.034	*0.158*
BE	0.147	0.228	0.066	*0.290*	0.215	0.062	*0.288*	0.179	0.038	*0.210*	0.203	0.031	*0.153*	0.189	0.041	*0.215*	0.248	0.036	*0.146*	0.215	0.055	*0.257*	0.200	0.041	*0.204*
CY	0.158	0.249	0.066	*0.264*	0.231	0.060	*0.258*	0.192	0.027	*0.143*	0.201	0.028	*0.138*	0.237	0.040	*0.170*	0.264	0.033	*0.125*	0.236	0.052	*0.219*	0.245	0.050	*0.202*
HU	0.159	0.244	0.074	*0.305*	0.255	0.066	*0.260*	0.211	0.054	*0.259*	0.217	0.047	*0.217*	0.209	0.038	*0.181*	0.258	0.031	*0.121*	0.221	0.072	*0.325*	0.278	0.048	*0.171*
EE	0.183	0.279	0.087	*0.310*	0.233	0.071	*0.306*	0.237	0.060	*0.251*	0.264	0.062	*0.233*	0.224	0.048	*0.213*	0.294	0.046	*0.158*	0.265	0.066	*0.250*	0.289	0.064	*0.222*
PT	0.185	0.289	0.080	*0.277*	0.248	0.073	*0.296*	0.241	0.061	*0.251*	0.254	0.051	*0.202*	0.236	0.043	*0.182*	0.307	0.039	*0.127*	0.281	0.058	*0.207*	0.281	0.062	*0.219*
IE	0.185	0.284	0.085	*0.299*	0.241	0.079	*0.326*	0.225	0.053	*0.237*	0.243	0.032	*0.131*	0.229	0.049	*0.214*	0.281	0.043	*0.153*	0.260	0.073	*0.279*	0.262	0.049	*0.187*
PL	0.191	0.287	0.095	*0.332*	0.320	0.102	*0.317*	0.246	0.059	*0.240*	0.259	0.056	*0.215*	0.249	0.049	*0.197*	0.298	0.036	*0.122*	0.267	0.082	*0.306*	0.317	0.063	*0.198*
UK	0.192	0.299	0.085	*0.282*	0.261	0.075	*0.289*	0.225	0.033	*0.146*	0.258	0.035	*0.137*	0.244	0.053	*0.218*	0.305	0.052	*0.172*	0.268	0.071	*0.266*	0.284	0.051	*0.178*
IT	0.196	0.301	0.092	*0.306*	0.272	0.084	*0.309*	0.226	0.038	*0.166*	0.252	0.043	*0.169*	0.254	0.061	*0.240*	0.329	0.053	*0.160*	0.275	0.079	*0.289*	0.299	0.067	*0.225*
ES	0.198	0.314	0.083	*0.264*	0.273	0.076	*0.280*	0.240	0.039	*0.161*	0.264	0.040	*0.151*	0.244	0.045	*0.183*	0.328	0.051	*0.155*	0.288	0.074	*0.258*	0.302	0.059	*0.195*
LT	0.199	0.294	0.104	*0.352*	0.292	0.093	*0.320*	0.252	0.084	*0.334*	0.284	0.079	*0.280*	0.247	0.040	*0.160*	0.302	0.039	*0.128*	0.277	0.082	*0.294*	0.324	0.069	*0.213*
GR	0.205	0.310	0.100	*0.321*	0.279	0.089	*0.318*	0.261	0.053	*0.203*	0.267	0.052	*0.195*	0.280	0.075	*0.269*	0.325	0.039	*0.120*	0.291	0.076	*0.262*	0.308	0.067	*0.219*
LV	0.231	0.336	0.126	*0.376*	0.356	0.122	*0.344*	0.286	0.095	*0.333*	0.331	0.095	*0.287*	0.288	0.043	*0.151*	0.383	0.065	*0.170*	0.313	0.099	*0.316*	0.393	0.106	*0.270*

Table 4. Fuzzy measures at Country level (2005)

	Rate of deprivation by dimension of deprivation								**mean**
Country	**FS0**	**FS1**	**FS2**	**FS3**	**FS4**	**FS5**	**FS6**	**FS7**	**FS1-FS7**
SE	0.092	0.072	0.043	0.045	0.073	0.073	0.078		0.068
IS	0.096	0.083	0.039	0.035	0.088	0.067	0.082		0.070
CZ	0.104	0.105	0.066	0.064	0.070	0.112	0.093	0.097	0.089
NL	0.109	0.087	0.045	0.064	0.056	0.107	0.092		0.080
NO	0.115	0.085	0.055	0.056	0.089	0.082	0.095		0.083
FI	0.118	0.091	0.070	0.060	0.078	0.108	0.101		0.089
DK	0.118	0.095	0.059	0.060	0.067	0.093	0.099		0.084
SI	0.122	0.096	0.059	0.075	0.083	0.116	0.103		0.094
DE	0.123	0.103	0.067	0.065	0.064	0.113	0.103	0.120	0.095
AT	0.123	0.098	0.052	0.072	0.048	0.107	0.105	0.089	0.087
FR	0.130	0.101	0.067	0.076	0.081	0.128	0.109	0.107	0.100
SK	0.133	0.127	0.076	0.072	0.073	0.123	0.117	0.133	0.107
HU	0.134	0.147	0.088	0.106	0.078	0.127	0.113	0.165	0.120
LU	0.137	0.097	0.040	0.073	0.064	0.140	0.111	0.110	0.097
BE	0.148	0.128	0.074	0.089	0.086	0.138	0.123	0.099	0.111
CY	0.162	0.132	0.070	0.069	0.122	0.132	0.133	0.141	0.120
EE	0.183	0.121	0.123	0.145	0.094	0.164	0.150	0.168	0.144
IT	0.188	0.158	0.070	0.097	0.116	0.175	0.151	0.163	0.140
UK	0.191	0.149	0.071	0.104	0.105	0.166	0.145	0.146	0.135
LV	0.192	0.248	0.137	0.185	0.119	0.185	0.157	0.261	0.185
PT	0.194	0.148	0.115	0.124	0.102	0.175	0.161	0.166	0.148
GR	0.196	0.161	0.108	0.109	0.151	0.157	0.156	0.161	0.150
IE	0.197	0.146	0.103	0.091	0.118	0.135	0.159	0.134	0.135
ES	0.197	0.145	0.088	0.104	0.085	0.180	0.164	0.161	0.140
LT	0.205	0.202	0.149	0.175	0.097	0.150	0.165	0.184	0.166
PL	0.206	0.267	0.128	0.132	0.119	0.157	0.166	0.200	0.172
average	0.151	0.130	0.079	0.090	0.089	0.131	0.124	0.148	0.118

NOTES FS0 stands for "HCR = FM = FS"

FS1 - FS7 refer to the seven dimensions of deprivation defined in section 2.4.6.

Table 5. "Normalised" Fuzzy measures at Country level (2005)

	'Normalised rates'								**Mean**
Country	**FS0**	**FS1**	**FS2**	**FS3**	**FS4**	**FS5**	**FS6**	**FS7**	**FS1-FS7**
SE	1.00	0.90	0.88	0.80	1.34	0.90	1.03		0.976
IS	1.00	0.99	0.76	0.60	1.54	0.80	1.03		0.953
CZ	1.00	1.17	1.22	1.03	1.13	1.25	1.08	0.95	1.119
NL	1.00	0.93	0.78	0.99	0.86	1.13	1.02		0.951
NO	1.00	0.85	0.91	0.81	1.30	0.82	1.00		0.950
FI	1.00	0.89	1.12	0.84	1.11	1.05	1.04		1.009
DK	1.00	0.93	0.94	0.85	0.96	0.90	1.01		0.931
SI	1.00	0.91	0.92	1.03	1.15	1.09	1.02		1.021
DE	1.00	0.97	1.03	0.88	0.87	1.06	1.02	1.00	0.977
AT	1.00	0.92	0.80	0.98	0.65	1.00	1.03	0.74	0.875
FR	1.00	0.89	0.98	0.98	1.05	1.13	1.02	0.84	0.984
SK	1.00	1.10	1.08	0.90	0.92	1.06	1.06	1.01	1.020
HU	1.00	1.27	1.24	1.31	0.98	1.09	1.03	1.25	1.168
LU	1.00	0.81	0.56	0.89	0.79	1.17	0.98	0.81	0.859
BE	1.00	1.00	0.95	1.00	0.98	1.07	1.00	0.68	0.953
CY	1.00	0.94	0.83	0.71	1.27	0.94	0.99	0.89	0.938
EE	1.00	0.76	1.27	1.32	0.86	1.03	0.99	0.94	1.025
IT	1.00	0.97	0.71	0.86	1.03	1.07	0.97	0.89	0.928
UK	1.00	0.90	0.71	0.91	0.92	1.00	0.92	0.78	0.878
LV	1.00	1.49	1.36	1.61	1.04	1.10	0.99	1.39	1.283
PT	1.00	0.88	1.13	1.07	0.88	1.04	1.00	0.87	0.982
GR	1.00	0.94	1.04	0.92	1.30	0.92	0.96	0.84	0.990
IE	1.00	0.85	0.99	0.77	1.01	0.79	0.98	0.70	0.870
ES	1.00	0.85	0.85	0.88	0.72	1.05	1.01	0.83	0.883
LT	1.00	1.13	1.38	1.42	0.79	0.84	0.97	0.91	1.064
PL	1.00	1.50	1.18	1.07	0.98	0.87	0.98	0.99	1.081
average	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.000

NOTES

'Normalised rates' N_{ij} : all values scaled such that:

(1) for each dimension (*j*), average over countries rescaled to = 1.0; and
(2) for each country (*i*), FS_j values scaled to correspond to FS0 = 1.0.

$$N_{ij} = \left(\frac{FS_{ij}}{FS_{.j}}\right) \Big/ \left(\frac{FS_{i0}}{FS_{.0}}\right)$$

Table 6. Latent and Manifest deprivation at aggregated level and for each dimension of deprivation (2005)

Country	FS0	L0	M0	*M0/L0*	L1	M1	*M1/L1*	L2	M2	*M2/L2*	L3	M3	*M3/L3*	L4	M4	*M4/L4*	L5	M5	*M5/L5*	L6	M6	*M6/L6*	L7	M7	*M7/L7*
SE	0.092	0.157	0.028	*0.177*	0.139	0.025	*0.181*	0.121	0.015	*0.121*	0.129	0.008	*0.061*	0.143	0.023	*0.163*	0.152	0.013	*0.088*	0.151	0.019	*0.127*			
IS	0.096	0.166	0.027	*0.162*	0.155	0.024	*0.155*	0.124	0.011	*0.088*	0.123	0.008	*0.063*	0.162	0.022	*0.138*	0.152	0.012	*0.077*	0.163	0.016	*0.096*			
CZ	0.104	0.164	0.043	*0.261*	0.171	0.038	*0.224*	0.139	0.031	*0.226*	0.146	0.022	*0.149*	0.148	0.025	*0.166*	0.194	0.022	*0.113*	0.157	0.039	*0.246*	0.180	0.021	*0.115*
NL	0.109	0.185	0.033	*0.178*	0.161	0.036	*0.221*	0.137	0.016	*0.121*	0.158	0.015	*0.092*	0.144	0.021	*0.146*	0.195	0.020	*0.105*	0.171	0.030	*0.176*			
NO	0.115	0.191	0.040	*0.207*	0.165	0.035	*0.212*	0.146	0.025	*0.168*	0.156	0.015	*0.099*	0.177	0.027	*0.154*	0.180	0.017	*0.095*	0.182	0.028	*0.154*			
FI	0.118	0.196	0.040	*0.207*	0.171	0.038	*0.223*	0.154	0.033	*0.215*	0.162	0.016	*0.096*	0.171	0.025	*0.149*	0.205	0.022	*0.106*	0.186	0.033	*0.178*			
DK	0.118	0.195	0.042	*0.215*	0.178	0.035	*0.199*	0.151	0.026	*0.170*	0.164	0.015	*0.091*	0.160	0.026	*0.162*	0.188	0.023	*0.121*	0.189	0.028	*0.147*			
SI	0.122	0.201	0.043	*0.214*	0.180	0.039	*0.214*	0.153	0.028	*0.186*	0.173	0.024	*0.141*	0.180	0.025	*0.141*	0.217	0.021	*0.098*	0.185	0.039	*0.212*			
DE	0.123	0.197	0.048	*0.245*	0.181	0.045	*0.247*	0.158	0.032	*0.200*	0.168	0.019	*0.113*	0.161	0.025	*0.156*	0.209	0.027	*0.127*	0.188	0.037	*0.199*	0.208	0.035	*0.167*
AT	0.123	0.207	0.039	*0.191*	0.184	0.037	*0.202*	0.154	0.021	*0.136*	0.174	0.021	*0.123*	0.150	0.021	*0.140*	0.210	0.020	*0.094*	0.191	0.037	*0.195*	0.190	0.021	*0.111*
FR	0.130	0.207	0.053	*0.257*	0.186	0.044	*0.239*	0.166	0.031	*0.185*	0.181	0.025	*0.137*	0.179	0.032	*0.182*	0.226	0.031	*0.138*	0.196	0.043	*0.220*	0.203	0.033	*0.165*
SK	0.133	0.222	0.044	*0.200*	0.223	0.038	*0.170*	0.180	0.029	*0.163*	0.184	0.021	*0.115*	0.185	0.021	*0.116*	0.231	0.025	*0.109*	0.207	0.043	*0.208*	0.233	0.033	*0.139*
HU	0.134	0.211	0.057	*0.272*	0.228	0.054	*0.236*	0.182	0.040	*0.222*	0.197	0.043	*0.218*	0.183	0.030	*0.163*	0.236	0.025	*0.106*	0.195	0.053	*0.272*	0.259	0.040	*0.156*
LU	0.137	0.220	0.055	*0.249*	0.178	0.056	*0.316*	0.154	0.024	*0.158*	0.186	0.025	*0.133*	0.166	0.035	*0.213*	0.243	0.034	*0.140*	0.201	0.047	*0.236*	0.216	0.031	*0.142*
BE	0.148	0.228	0.069	*0.301*	0.216	0.061	*0.281*	0.184	0.039	*0.212*	0.204	0.033	*0.164*	0.191	0.044	*0.228*	0.248	0.038	*0.154*	0.215	0.056	*0.261*	0.206	0.041	*0.198*
CY	0.162	0.257	0.067	*0.260*	0.232	0.062	*0.269*	0.200	0.032	*0.161*	0.205	0.026	*0.126*	0.245	0.039	*0.159*	0.261	0.033	*0.125*	0.240	0.054	*0.224*	0.252	0.050	*0.200*
EE	0.183	0.280	0.086	*0.307*	0.237	0.068	*0.286*	0.239	0.067	*0.280*	0.265	0.062	*0.235*	0.235	0.042	*0.181*	0.302	0.045	*0.147*	0.265	0.068	*0.258*	0.289	0.062	*0.215*
IT	0.188	0.288	0.089	*0.308*	0.263	0.083	*0.315*	0.222	0.037	*0.165*	0.246	0.040	*0.163*	0.244	0.060	*0.246*	0.314	0.049	*0.157*	0.266	0.074	*0.279*	0.290	0.061	*0.211*
UK	0.191	0.297	0.085	*0.287*	0.265	0.074	*0.280*	0.227	0.035	*0.154*	0.258	0.037	*0.145*	0.244	0.051	*0.209*	0.306	0.052	*0.169*	0.264	0.071	*0.271*	0.283	0.054	*0.191*
LV	0.192	0.289	0.094	*0.326*	0.338	0.101	*0.299*	0.251	0.078	*0.312*	0.295	0.082	*0.276*	0.271	0.040	*0.147*	0.333	0.043	*0.129*	0.272	0.076	*0.279*	0.370	0.083	*0.224*
PT	0.194	0.303	0.086	*0.283*	0.261	0.081	*0.312*	0.248	0.061	*0.247*	0.264	0.054	*0.206*	0.248	0.048	*0.193*	0.327	0.043	*0.131*	0.292	0.063	*0.215*	0.296	0.065	*0.219*
GR	0.196	0.302	0.091	*0.300*	0.273	0.084	*0.307*	0.253	0.052	*0.204*	0.257	0.048	*0.185*	0.285	0.063	*0.219*	0.317	0.037	*0.115*	0.281	0.072	*0.255*	0.292	0.066	*0.225*
IE	0.197	0.299	0.095	*0.319*	0.256	0.086	*0.338*	0.242	0.058	*0.237*	0.252	0.036	*0.142*	0.253	0.063	*0.248*	0.289	0.043	*0.150*	0.275	0.081	*0.294*	0.279	0.053	*0.189*
ES	0.197	0.314	0.081	*0.256*	0.270	0.072	*0.266*	0.243	0.043	*0.176*	0.243	0.043	*0.176*	0.243	0.043	*0.176*	0.327	0.050	*0.153*	0.290	0.071	*0.245*	0.298	0.060	*0.200*
LT	0.205	0.302	0.108	*0.358*	0.307	0.100	*0.325*	0.263	0.092	*0.349*	0.298	0.083	*0.279*	0.260	0.042	*0.163*	0.318	0.037	*0.118*	0.285	0.086	*0.302*	0.324	0.066	*0.202*
PL	0.206	0.309	0.102	*0.330*	0.359	0.114	*0.318*	0.269	0.065	*0.240*	0.278	0.059	*0.214*	0.270	0.055	*0.205*	0.319	0.044	*0.137*	0.282	0.090	*0.319*	0.335	0.071	*0.211*

Table 7. Fuzzy measures at Country level (2004)

	Rate of deprivation by dimension of deprivation								**Mean**
Country	**FS0**	**FS1**	**FS2**	**FS3**	**FS4**	**FS5**	**FS6**	**FS7**	**FS1-FS7**
IS	0.101	0.087	0.039	0.033	0.097	0.072	0.086		0.074
DK	0.107	0.089	0.057	0.057	0.062	0.086	0.089		0.078
NO	0.108	0.083	0.051	0.056	0.084	0.081	0.088		0.079
FI	0.109	0.088	0.069	0.053	0.082	0.103	0.093		0.085
SE	0.111	0.087	0.053	0.052	0.085	0.083	0.092		0.080
LU	0.127	0.093	0.036	0.071	0.070	0.123	0.103	0.104	0.091
AT	0.128	0.103	0.064	0.075	0.058	0.109	0.108	0.093	0.092
FR	0.135	0.115	0.071	0.081	0.093	0.137	0.112	0.112	0.107
BE	0.143	0.117	0.079	0.087	0.075	0.138	0.119	0.104	0.108
IT	0.191	0.163	0.074	0.102	0.118	0.179	0.154	0.168	0.144
ES	0.199	0.144	0.094	0.108	0.087	0.173	0.162	0.168	0.142
GR	0.199	0.180	0.109	0.113	0.160	0.145	0.158	0.157	0.153
EE	0.202	0.145	0.133	0.155	0.111	0.237	0.161	0.178	0.165
PT	0.205	0.156	0.127	0.133	0.109	0.176	0.168	0.171	0.156
IE	0.209	0.153	0.107	0.098	0.121	0.146	0.166	0.144	0.143
average	0.152	0.120	0.078	0.085	0.094	0.133	0.124	0.140	0.116

NOTES FS0 stands for "HCR = FM = FS"
FS1 - FS7 refer to the seven dimensions of deprivation defined in section 2.4.6.

Table 8. "Normalised" Fuzzy measures at Country level (2004)

	'Normalised rates'								**Mean**
Country	**FS0**	**FS1**	**FS2**	**FS3**	**FS4**	**FS5**	**FS6**	**FS7**	**FS1-FS7**
IS	1.00	1.09	0.74	0.58	1.55	0.81	1.04		0.973
DK	1.00	1.04	1.03	0.96	0.93	0.92	1.01		0.983
NO	1.00	0.97	0.93	0.92	1.26	0.86	1.00		0.992
FI	1.00	1.01	1.24	0.87	1.21	1.07	1.04		1.063
SE	1.00	0.98	0.93	0.83	1.23	0.86	1.02		0.977
LU	1.00	0.92	0.56	1.00	0.88	1.11	0.99	0.88	0.917
AT	1.00	1.02	0.98	1.04	0.73	0.97	1.04	0.79	0.944
FR	1.00	1.08	1.03	1.07	1.12	1.17	1.02	0.90	1.048
BE	1.00	1.03	1.07	1.09	0.85	1.10	1.02	0.78	0.993
IT	1.00	1.07	0.76	0.95	0.99	1.07	0.98	0.95	0.973
ES	1.00	0.92	0.93	0.97	0.70	0.99	1.00	0.92	0.928
GR	1.00	1.14	1.08	1.02	1.30	0.84	0.97	0.86	1.024
EE	1.00	0.91	1.29	1.37	0.88	1.34	0.98	0.95	1.091
PT	1.00	0.96	1.21	1.16	0.86	0.99	1.00	0.91	1.011
IE	1.00	0.92	1.00	0.84	0.93	0.80	0.97	0.75	0.901
average	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.000

NOTES 'Normalised rates' N_{ij} : all values scaled such that:

(1) for each dimension (*j*), average over countries rescaled to = 1.0; and
(2) for each country (*i*), FS_j values scaled to correspond to FS0 = 1.0.

Table 9. Latent and Manifest deprivation at aggregated level and for each dimension of deprivation (2004)

Country	**FS0**	**L0**	**M0**	***M0/L0***	**L1**	**M1**	***M1/L1***	**L2**	**M2**	***M2/L2***	**L3**	**M3**	***M3/L3***	**L4**	**M4**	***M4/L4***	**L5**	**M5**	***M5/L5***	**L6**	**M6**	***M6/L6***	**L7**	**M7**	***M7/L7***
IS	0.101	0.172	0.030	*0.175*	0.165	0.024	*0.144*	0.129	0.011	*0.084*	0.125	0.009	*0.074*	0.173	0.025	*0.147*	0.157	0.016	*0.104*	0.168	0.019	*0.114*			
DK	0.107	0.181	0.034	*0.187*	0.168	0.028	*0.169*	0.141	0.024	*0.167*	0.152	0.013	*0.086*	0.147	0.022	*0.151*	0.174	0.019	*0.112*	0.175	0.021	*0.121*			
NO	0.108	0.182	0.033	*0.183*	0.160	0.030	*0.189*	0.137	0.022	*0.162*	0.150	0.013	*0.089*	0.170	0.021	*0.125*	0.173	0.016	*0.092*	0.170	0.026	*0.152*			
FI	0.109	0.184	0.035	*0.192*	0.166	0.032	*0.191*	0.149	0.030	*0.199*	0.151	0.011	*0.076*	0.166	0.026	*0.156*	0.192	0.020	*0.104*	0.176	0.026	*0.150*			
SE	0.111	0.183	0.039	*0.212*	0.165	0.033	*0.201*	0.143	0.021	*0.149*	0.150	0.013	*0.086*	0.166	0.029	*0.177*	0.176	0.019	*0.107*	0.178	0.026	*0.147*			
LU	0.127	0.206	0.049	*0.240*	0.169	0.051	*0.298*	0.143	0.020	*0.143*	0.178	0.021	*0.115*	0.163	0.034	*0.212*	0.222	0.029	*0.129*	0.186	0.045	*0.239*	0.204	0.027	*0.131*
AT	0.128	0.212	0.044	*0.206*	0.19	0.041	*0.216*	0.166	0.026	*0.157*	0.181	0.022	*0.121*	0.163	0.023	*0.141*	0.213	0.023	*0.109*	0.2	0.036	*0.179*	0.196	0.025	*0.130*
FR	0.135	0.213	0.056	*0.265*	0.196	0.054	*0.274*	0.174	0.032	*0.181*	0.189	0.026	*0.138*	0.192	0.036	*0.189*	0.238	0.034	*0.141*	0.200	0.047	*0.235*	0.211	0.035	*0.168*
BE	0.143	0.228	0.059	*0.257*	0.21	0.051	*0.242*	0.183	0.039	*0.213*	0.203	0.028	*0.136*	0.185	0.034	*0.182*	0.247	0.034	*0.137*	0.212	0.051	*0.238*	0.211	0.036	*0.168*
IT	0.191	0.293	0.090	*0.306*	0.271	0.083	*0.306*	0.227	0.039	*0.171*	0.249	0.044	*0.177*	0.249	0.060	*0.243*	0.318	0.052	*0.163*	0.270	0.075	*0.277*	0.297	0.062	*0.210*
ES	0.199	0.314	0.083	*0.264*	0.270	0.073	*0.272*	0.247	0.046	*0.186*	0.264	0.042	*0.159*	0.242	0.043	*0.177*	0.324	0.048	*0.148*	0.288	0.072	*0.250*	0.302	0.064	*0.213*
GR	0.199	0.302	0.096	*0.316*	0.289	0.090	*0.311*	0.255	0.053	*0.209*	0.263	0.049	*0.188*	0.293	0.066	*0.226*	0.309	0.035	*0.113*	0.285	0.071	*0.250*	0.291	0.065	*0.222*
EE	0.202	0.309	0.096	*0.309*	0.270	0.077	*0.285*	0.262	0.074	*0.283*	0.291	0.067	*0.230*	0.259	0.054	*0.209*	0.371	0.068	*0.184*	0.287	0.076	*0.265*	0.312	0.068	*0.219*
PT	0.205	0.313	0.096	*0.306*	0.275	0.086	*0.312*	0.262	0.069	*0.264*	0.277	0.060	*0.215*	0.260	0.054	*0.206*	0.334	0.047	*0.140*	0.306	0.066	*0.216*	0.307	0.069	*0.225*
IE	0.209	0.318	0.100	*0.315*	0.275	0.087	*0.317*	0.256	0.060	*0.233*	0.266	0.041	*0.153*	0.265	0.065	*0.246*	0.309	0.047	*0.152*	0.292	0.083	*0.283*	0.292	0.061	*0.210*

References

Alexander C.H. (1999). A rolling sample survey for yearly and decennial uses. Proceedings of the International Statistical Institute, Helsinki, Contributed papers, Book 1, pp. 29–30.

Alkire S., Foster J. (2008). Counting and Multidimensional Poverty Measurement. 30th General Conference of the International Association for Research in Income and Wealth, Portoroz, Slovenia, August 24-30, 2008.

Anselin L. (1992). Spatial econometrics: method and models. Kluwer, Boston.

Atkinson A.B. (1970). On the measurement of Inequality. Journal of Economic Theory, 2, pp. 244-263.

Atkinson A.B., Cantillon B., Marlier E., Nolan B. (2002). Social Indicators: The EU and Social Inclusion, Oxford University Press, Oxford.

Atkinson A.B. (2003). Multidimensional deprivation: contrasting social welfare and counting approaches. Journal of Economic Inequality, 1, pp. 51–65.

Australian Bureau of Statistics (1993). The Australian Population Monitor, Canberra, ABS.

Battese G.E., Harter R.M., Fuller W.A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. Journal of the American Statistical Association, 83, 401, pp. 28–36.

Berenger V., Verdier-Chouchane A. (2007). Multidimensional Measures of Well-Being: Standard of Living and Quality of Life Across Countries. World Development, 35, pp. 1259–1276.

Betti G. (1998). Intertemporal equivalence scales and cost of children using BHPS. ERSC Research Centre on Micro-social Change Working Papers, Paper 11/98, Colchester University of Essex.

Betti G., Cheli B., Lemmi A., Verma V. (2005). On the construction of fuzzy measures for the analysis of povertà and social exclusion. International Conference to Honour Two Eminent Scientists C GINI and MO LORENZ, University of Siena, May 23-26, 2005.

Betti G., Cheli B., Lemmi A., Verma V. (2006). On the construction of fuzzy measures for the analysis of poverty and social exclusion, Statistica & Applicazioni, 4(1), pp. 77-97

Betti G., Dourmashkin N., Rossi M.C., Verma V., Yin Y.P. (2001). Study of the Problem of Consumer Indebtedness: Statistical Aspects. Report to the Commission of the European Communities, Directorate-General for Health and Consumer Protection, Commission of the European Communities, Brussels.

Betti G., Verma V. (1999). Measuring the degree of poverty in a dynamic and comparative context: a multi-dimensional approach using fuzzy set theory. Proceedings, ICCS-VI, 11, pp. 289-301, Lahore, Pakistan, August 27-31, 1999.

Betti G., Verma V. (2002). Non-monetary or Lifestyle Deprivation. In. Eurostat, European Social statistics: Income, Poverty and Social Exclusion: 2nd Report, Luxembourg: Office for Official Publications of the European Communities, pp. 76-92.

Betti G., Verma V. (2004). A methodology for the study of multi-dimensional and longitudinal aspects of poverty and deprivation. Working Paper 49, Department of Quantitative Methods, University of Siena.

Betti G., Verma V. (2008). Fuzzy measures of the incidence of relative poverty and deprivation: a multi-dimensional perspective. Statistical Methods and Applications, 12(2), pp. 225-250.

Bourguignon F., Chakravarty S.R. (2003). The measurement of multidimensional poverty. Journal of Economic Inequality, 1, pp. 25–49.

Brackstone G.J. (1987). Small Area Data: Policy Issues and Technical Challenges. In Platek R., Rao J.N.K., Sardnal C.E. and Singh M.P. (eds.), Small Area Statistics, Wiley, New York, pp. 3–20.

Breckling J., Chambers R. (1988). M-quantiles. Biometrika, 75, pp. 761-71.

Brunsdon C., Fotheringham A.S., Charlton M. (1999). Some notes on parametric significance tests for geographically weighted regression. Journal of Regional Science, 39, pp. 497-524.

Caplan D., Haworth M., Steel D. (1999). UK labour market statistics: Combining continuous survey data into monthly reports. Proceedings of the 52nd Session of the International Statistical Institute, Helsinki.

Cerioli A., Zani S. (1990). A fuzzy approach to the measurement of poverty. In: Dagum C., Zenga M. (eds.) Income and wealth distribution, inequality and poverty, Springer Verlag, Berlin, pp. 272-284.

Chakravarty S. R., Mukherjee D., Ranade R.R. (1998). On the Family of Subgroup and Factor Decomposable Measures of Multidimensional Poverty. In D. J. Slottje (eds.), Research on Economic Inequality, 8, JAI Press, Stamford, CT and London.

Chambers R., Dunstan R. (1986). Estimating distribution functions from survey data. Biometrika, 73, pp. 597-604.

Chambers R., Tzavidis N. (2006). M-quantile models for small area estimation. Biometrika, 93, pp. 225-268.

Chambers R., Tzavidis N. (2008). Robust prediction of small area means and distributions. CCSR Working Paper 2007-08.

Cheli B. (1995). Totally Fuzzy and Relative Measures of Poverty in Dynamics Context. Metron, 53(1), pp.183-205.

Cheli B., Lemmi A. (1995). A Totally Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty. Economic Notes, 24, pp. 115-134.

Cheli B., Betti G. (1999). Totally Fuzzy and Relative Measures of Poverty Dynamics in an Italian Pseudo Panel, 1985-1994, Metron, 57(1-2), pp. 83-104.

Chiappero Martinetti E. (2006). Capability Approach and Fuzzy Set Theory: Description, Aggregation and Inference Issue. In A. Lemmi and G. Betti (eds.) Fuzzy Set Approach to Multidimensional Poverty Measurement, pp. 93–114, Springer, New York.

Ciampalini G., Betti G., Verma V. (2009). Comparability in self-employment income, Working Paper 82/09, Department of Quantitative Methods, University of Siena.

Coelli T., Prasada Rao D.S., Battese G.E. (1998). An Introduction to Efficiency and Productivity Analysis. Kluwer Academic Publishers, Boston.

Cressie N (1993) Statistics for spatial data. Wiley, New York.

Dagum C., Zenga M. (1989). Income and wealth distribution, inequality and poverty. Springer Verlag, Berlin.

Dalton H. (1920). The measurement of the inequality of income. The Economic Journal, 30, pp. 348-361.

Deutsch J., Silber J. (2006). The Fuzzy Sets Approach to Multidimensional Poverty Analysis: Using the Shapley Decomposition to Analyze the Determinants of Poverty in Israel. In A. Lemmi and G. Betti (eds.) Fuzzy Set Approach to Multidimensional Poverty Measurement, pp. 155–174, Springer, New York.

Di Marco M. (2006). Self Employment Incomes in The Italian EU-SILC: Measurement and International Comparability. Proceedings of the EU-SILC Conference on Comparative EU Statistics on Income and Living Conditions, Issues and Challenges.

Duan N. (1983). Smearing estimate: A nonparametric retransformation method. Journal of the American Statistical Association, 78, pp. 605-610.

Dubois D., Prade H. (1980). Fuzzy Sets and Systems. Academic Press, Boston, New York, London.

Elbers C., Lanjouw J.O., Lanjouw P. (2003). Micro-level estimation of poverty and inequality. Econometrica, 71, pp. 455-364.

Eurostat (2006). Treatment of negative income: empirical assessment of the impact of methods used. Report N. ISR I.04, Project EU-SILC (Community statistics on income and living conditions) 2005/S 116-114302 – Lot 1 (Methodological studies to estimate the impact on comparability of the national methods used).

Fay R.E., Herriot R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74, pp. 269–277.

Fine K. (1975). Vagueness, truth and logic. Syntheses, 30, pp. 265–300, reprinted in Rosanna Keefe and Peter Smith (eds) (1996) Vagueness: A Reader, Cambridge, MA. and London: MIT Press.

Foster J., Greer J., Thorbecke E. (1984). A Class of Decomposable Poverty Measures. Econometrica, 52, pp. 761-766.

Fotheringham A.S., Brunsdon C., Charlton, M. (1997). Two techniques for exploring non-stationarity in geographical data. Geographical Systems, 4, pp. 59-82.

Gini C. (1912). Variabilità e Mutabilità. Bologna, Tipografia di Paolo Cuppini.

Giorgi L., Verma V. (2002). European social statistics: income, poverty and social exclusion, 2nd report. Office for Official Publications of the European Communities, Luxembourg.

Giusti C., Pratesi M., Salvati N. (2009). Estimation of poverty indicators: a comparison of small area methods at LAU1-2 level in Tuscany. Presentation at NTSS Conference. Brussels 18-20 Febraury 2009.

Harville D.A. (1991). Comment. Statistical Science, 6, pp. 35-39.

Kakwani N.C. (1980). Income Inequality and Poverty. Methods of Estimation and Policy Applications, Oxford UP, New York, Oxford, London.

Kelloway E.K. (1998). Using LISREL for structural equation modelling. Sage, London.

Kish L. (1987). Statistical Research Design. New York, John Wiley & Sons.

Kish L. (1990). Rolling samples and censuses. Survey Methodology, 16(1), pp. 63-71.

Kish L. (1994). Multi-population survey designs: five types with seven shared aspects. International Statistical Review, 62, pp. 167-186.

Kish L. (1997). Designs and Uses for Multipopulation Samples. Proceedings of the 51nd Session of the International Statistical Institute, Istanbul.

Kish L. (1998a). Space/Time variations and rolling samples. Journal of Official Statistics, 14, pp. 31-46.

Kish L. (1998b). Combining multipopulation statistics. Journal of Statistical Planning and Inference, 102, pp. 109-118.

Kish L. (1999a). Cumulating/ Combining Population Surveys. Survey Methodology, 25(2), pp. 129-138.

Kish L. (1999b). Combining Surveys: A Framework. Proceedings of the 52nd Session of the International Statistical Institute, Helsinki.

Klir G.J., Yuan B. (1995). Fuzzy sets and fuzzy logic, theory and applications. Prentice Hall PTR, Upper Saddle River, New Jersey.

Koenker R., Bassett G. (1978). Regression quantiles. Econometrica, 46, pp. 33-50.

Koenker R., Hallock K.F. (2001). Quantile Regression: An introduction, Journal of Economic Perspectives, 51, pp. 143-53.

Kuklys W. (2005). Amartya Sen's Capability Approach: Theoretical Insights and Empirical Applications. Berlin, Springer-Verlag.

Laird N.M., Mosteller F. (1990). Some Statistical Methods for Combining Experimental Results. International Journal of Technology Assessment, 6, pp. 5-30.

Lelli S. (2001). Factor Analysis vs. Fuzzy Sets Theory: Assessing the Influence of Different Techniques on Sen's Functioning Approach. Discussion Paper Series DPS 01.21, Centre for Economic Studies, Catholic University of Leuven, Belgium.

Lemmi A., Betti G. (2006). Fuzzy Set Approach to Multidimensional Poverty Measurement. Springer, New York.

Lorenz M.O. (1905). Methods for measuring concentration of wealth. Journal of the American Statistical Association, 9, pp. 209-219.

Lovell C.A.K., Richardson S., Travers P., Wood P. (1994). Resources and Functionings: A New View of Inequality in Australia. In W. Eichhorn (eds.), Models and Measurement of Welfare and Inequality, Springer-Verlag, Heidelberg.

Maasoumi E. (1986). The measurement and decomposition of multi-dimensional inequality. Econometrica, 54(4), pp. 991-997.

Maasoumi E., Nickelsurg G. (1988). Multivariate Measures of Well-Being and an Analysis of Inequality in the Michigan Data. Journal of Business and Economic Statistics, 6, pp. 327-334.

Mc Culloch C.E., Searle S.R. (2001). Generalized, Linear, And Mixed Models. Wiley, New York.

Miceli D. (1997). Mesure de la pauvreté. Théorie et Application à la Suisse. Thèse de doctorat ès sciences économiques et sociale, Université de Genève.

Molina I., Rao J.N.K. (2009). Estimation of poverty measures in small area. Working Paper 09-15, Statistics and Econometrics Series 05. Departamento de Estadística, Universidad Carlos III de Madrid.

Molina I., Saei A., Lombarda M.J. (2007). Small area estimates of labour force partecipation under a multinomial logit mixed model. Journal of the Royal Statistical society A, 170, pp. 975-1000.

Morales D., Esteban M.D., Perez A., Santamarıa L. (2009). Small area estimation of poverty indicators under area-level time models. Paper presented at the Conference, on Small Area Estimation, 29 June - 1 July 2009, Elche, Spain.

Morton S.C. (1999). Combining Surveys from a Meta-analysis Perspective. Proceedings of the 52nd Session of the International Statistical Institute, Helsinki.

National Center for Health Statistics (1958). Statistical designs of the Health Household Interview Surveys. Public Health Series, pp. 584-A2.

Nelder JA, Mead R (1965). A simplex method for function minimization. The Computer Journal, 7, pp. 308–313.

Neri L., Gagliardi F., Ciampalini G., Verma V., Betti G. (2009). Outliers at upper end of income distribution, Working Paper 86/09, Department of Quantitative Methods, University of Siena.

Newey W.K., Powell J.L. (1987). Asymmetric least squares estimation and testing. Econometrica, 55, pp. 819-47.

Nolan B., Whelan C.T. (1996). Resources, deprivation and poverty. Clarendon Press, Oxford.

Opsomer J.D., Claeskens G., Ranalli M.G., Kauermann G., Breidt F.J. (2008). Nonparametric Small Area Estimation Using Penalized Spline Regression. Journal of the Royal Statistical Society, B, 70(1), pp. 265-286.

Pantula S.G., Pollock K.H. (1985). Nested analysis of variance with autocorrelated errors. Biometrics, 41, pp. 909-920

Prasad N. G. N., Rao J.N.K. (1990). The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association, 85, pp. 163-171.

Pratesi M., Salvati N. (2005). Small area estimation: the EBLUP estimator with autoregressive random area effects. Report no 261, Department of Statistics and Mathematics, University of Pisa.

Pratesi M., Salvati N. (2008). Small Area Estimation: the EBLUP estimator based on spatially correlated random area effects. Statistical Methods & Applications, 17(1), pp. 113-141.

Pratesi M., Ranalli M.G., Salvati N. (2006a). Nonparametric M-quantile regression via penalized splines. In ASA Proceedings on Survey Research Methods, Alexandria, VA, pp. 3596-3603.

Pratesi M., Ranalli M.G., Salvati, N. (2006b). P-splines M-quantile small area estimation: assessing the ecological conditions of lakes in the Northeastern US. Proceedings of the Conference on Spatial Data Methods for Environmental and Ecological Processes, Foggia, September 2006.

Purcell N.J., Kish L. (1980). Postcensal Estimates For Local Areas (Or Domains). International Statistical Review, 48, pp. 3-18.

Qizilbash M. (2006). Philosophical Accounts of Vagueness, Fuzzy Poverty Measures and Multidimensionality. In A. Lemmi and G. Betti (eds.) Fuzzy Set Approach to Multidimensional Poverty Measurement, pp. 9–28, Springer, New York.

Raghunathan T.E., Lepkowski J.M., Van Hoewyk J.V., Solenberger P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology, 27, 1, pp. 85-95.

Ram R. (1982). Composite indices of physical quality of life, basics needs fulfilment and income. A principal component representation. Journal of Development Economics, 11, pp. 227-248.

Rao, J.N.K. (1994). Estimation of totals and distributing functions using auxiliary information at the estimation stage. Journal of Official Statistics, 10, pp. 153-165.

Rao J.N.K. (2003). Small Area Estimation. John Wiley.

Rao J.N.K., Kovar J.G., Mantel H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. Biometrika, 77, pp. 365-75.

Rao J.N.K., Yu M. (1994). Small area estimation by combining time series and cross sectional data. Canadian Journal of Statistics, 22, pp. 511-528.

Ruppert D., Wand M.P., Carroll R. (2003). Semiparametric Regression. Cambridge University Press.

Salvati N., Tzavidis N., Pratesi M., Chambers R. (2008). Small Area Estimation via M-quantile Geographically Weighted Regression. Submitted for publication to Computational Statistics and Data Analysis.

Sen A.K. (1976). Poverty: an Ordinal Approach to Measurement. Econometrica, 44, pp. 219-231.

Sen A.K. (1985). Commodities and capabilities. Amsterdam, North Holland.

Sen A.K. (1992). Inequality rexamined. Clarendon Press, Oxford.

Shannon C. E. (1948). The Mathematical Theory of Communication. Bell System Tech Journal, 27, pp. 379–423 and 623–56.

Stiglitz J.E., Sen A., Fitoussi JP. (2009). Report by the Commission on the Measurement of Economic Performance and Social Progress.

Theil H. (1967). Economics and Information Theory. North Holland, Amsterdam.

Tsui K. Y. (1995). Multidimensional Generalizations of the Relative and Absolute Inequality Indices: The Atkinson-Kolm-Sen Approach. Journal of Economic Theory, 67, pp. 251-265.

Tsui K. Y. (2002). Multidimensional poverty indices. Social Choice and Welfare, 19(1), pp. 69-93.

Tzavidis N., Salvati N., Pratesi M., Chambers R. (2008). M-quantile Models with Application to Small Area Estimation and Poverty Mapping, Statistical Methods and Applications, 17(3), pp. 393-411.

Verma V. (1991). Sampling Methods: Training Handbook. Tokyo, Statistical Institute for Asia and the Pacific (SIAP).

Verma V. (1999). Combining national surveys for the European Union. Proceedings of the 52nd Session of the International Statistical Institute, Helsinki.

Verma V. (2001a). EU-SILC Sampling Guidelines. Report prepared for Eurostat.

Verma V. (2001b). EU-SILC: Proposals for a survey structure for those countries beginning a new survey. Proceedings, WG conference Rolling Samples and Sampling in Time - Problems of Data Accumulation and Data Quality, Trier, Germany.

Verma V. (2001c). The case for a Continuous Household Budget Survey. Proceedings, WG conference Rolling Samples and Sampling in Time - Problems of Data Accumulation and Data Quality, Trier, Germany.

Verma, V. (2002). Comparability in Multi-country Survey Programmes. Journal of Statistical Planning and Inference, 102(1), pp. 189-210.

Verma V. (2005). Indicators to reflect social exclusion and poverty Report prepared for Employment and Social Affairs DG - with contribution of Gianni Betti, Achille Lemmi, Anna Mulas, Michela Natilli, Laura Neri and Nicola Salvati.

Verma V., Clemenceau A. (1993). Methodology of the European Community Household Panel. Statistics in Transition, 2(7), pp. 1023-1062.

Verma V., Betti G. (2006). EU Statistics on Income and Living Conditions (EU-SILC): Choosing the survey structure and sample design. Statistics in Transition, 7(5), pp. 935-970.

Verma V., Gagliardi F., Ciampalini G. (2009). Methodology of labour force surveys: (3) Sample rotation patterns, Working Paper 80/09, Department of Quantitative Methods, University of Siena.

Whelan C.T., Layte R., Maitre B., Nolan B. (2001). Income, deprivation and economic strain: an analysis of the European Community Household Panel. European Sociological Review, 17, pp. 357-372.

Yu D.L., Wu C. (2004). Understanding population segregation from Landsat ETM+ imagery: a geographically weighted regression approach. GISience and Remote Sensing, 41, pp. 145-164.

Zadeh L.A. (1965). Fuzzy sets. Information and Control, 8, pp. 338-353.