SAMPLE
Small Area Methods for Poverty
and Living Condition Estimates

# SAMPLE DELIVERABLE 4
# LITERATURE REVIEW

# VOLUMES I, II and III

| | |
|---|---|
| Grant agreement No: | SSH - CT - 2008 - 217565 |
| Project Acronym: | SAMPLE |
| Project Full title: | Small Area Methods for Poverty and Living Conditions Estimates |
| Funding Scheme: | Collaborative Project - Small or medium scale focused research project |
| Deliverable n.: | 4 |
| Deliverable name: | Literature review |
| WP no.: | 1,2,3 |
| Lead beneficiary: | 1 |
| Nature: | Report |
| Dissemination level: | PU |
| Due delivery date from Annex I: | 30th September 2008 |
| Actual delivery date: | 7th May 2009 |
| Project co-ordinator name: | Mrs. Monica Pratesi |
| Title: | Associate Professor of Statistics - University of Pisa |
| Organization: | Department of Statistics and Mathematics Applied<br>to Economics of the University of Pisa (UNIPI-DSMAE) |
| Tel: | +39-050-2216252, +39-050-2216492 |
| Fax: | +39-050-2216375 |
| E-mail: | coordinator@sample-project.eu |
| Project website address: | www.sample-project.eu |

# LITERATURE REVIEW

# VOLUME I

| | |
|---|---|
| Project Acronym: | SAMPLE |
| Project Full title: | Small Area Methods for Poverty and Living Conditions Estimates |
| Project/Contract No: | EU-FP7-SSH-2007-1 |
| Grant agreement No: | 217565 |
| Work Package 1: | New indicators and models for inequality and poverty with attention to social exclusion, vulnerability and deprivation |
| Document title: | WP1 Literature review |
| Date: | 30th March 2009 |
| Type of document: | Deliverable D4 |
| Status: | Final |
| Editors: | A. Lemmi, V. Verma, T. Panek |
| Authors: | Achille Lemmi (lemmi@unisi.it), CRIDIRE-UNISI<br>Vijay Verma (verma@unisi.it), CRIDIRE-UNISI<br>Gianni Betti (betti2@unisi.it), CRIDIRE-UNISI<br>Laura Neri (neri@unisi.it), CRIDIRE-UNISI<br>Francesca Gagliardi (gagliardi10@unisi.it), CRIDIRE-UNISI<br>Caterina Ferretti (ferretti@ds.unifi.it), CRIDIRE-UNIFI<br>Kordos Jan (jan.kor@poczta.onet.pl), WSE<br>Panek Tomasz (tompa10@interia.pl), WSE<br>Szukiełojć-Bieńkuńska Anna (a.bienkunska@stat.gov.pl), GUS<br>Szulc Adam (a_szulc@interia.pl), WSE<br>Zięba Agnieszka (azieba@sgh.waw.pl), WSE |

# Contents

**Prologue** **IV**

**1 Poverty indicators in fuzzy and non fuzzy approach (coordinator Achille Lemmi) 1**

1.1 Traditional Poverty Approach ........ 1

Poverty measures ........ 1

Inequality measures........ 3

1.2. Multidimensional Approach 6

Social Welfare Approach ........ 6

Counting Approach ........ 8

Sen's Capability Approach ........ 8

Distance Function Approach........ 9

Information Theory Approach ........ 10

Axiomatic Approach ........ 11

Supervaluationist approach ........ 14

Shapley Decomposition ........ 15

Foster-Greer-Thorbecke index of multidimensional poverty........ 16

Fuzzy Set Approach ........ 17

1.3. The Fuzzy approach 18

Fuzzy monetary........ 18

Fuzzy supplementary ........ 19

Combination........ 22

1.4. The Leaken indicators ........ 26

Background of the Laeken indicators........ 26

1.5. References ........ 29

**2 Pooled estimates of indicators (coordinator Vijay Verma) 33**

2.1. Introduction ........ 33

Pooling and its fundamental objectives........ 33

Prerequisite for pooling: comparability........ 34

Diverse scenarios ........ 34

2.2. Scenario 1. Different population, different data sources ........ 35

2.3. Scenario 2. Different population, similar or same data source.....35
Examples of category 2.(a): aggregation of data or estimates.....35
*Pooling of national estimates*.....35
*Meta-analysis*.....37
*Combining separate sites*.....37
*Multinational designs*.....37
Examples of category 2.(b): disaggregation for separate reporting by domain.....38
*Multi-domain designs*.....38
2.4. Scenario 3. Same population, different data sources.....38
2.5. Scenario 4. Same population, similar data sources.....39
Periodic Surveys.....39
Split panel designs.....40
Symmetrical rotations.....41
Asymmetrical Cumulations.....41
Rolling samples and censuses.....41
More robust poverty measures.....42
*Poverty rates cumulated over time*.....42
*Poverty rates with different thresholds*.....42
*Poverty rates with poverty lines at different levels*.....42
2.6. Example of cumulation in European surveys.....43
Cumulation of data and measures in a continuous Household Budget Survey.....43
Cumulation of cross-sectional and longitudinal data from EU-SILC 46
2.7. Differences and averages under rotational design.....50
Illustrations from the EU Labour Force Survey.....50
Illustration of poverty measures of average and net differences from a household panel 55
2.8 References.....58

**3 Poverty and inequality measures for Regional and Local Governments (coordinator Tomasz Panek) 61**
3.1 Introduction.....61
Choice of appropriate indicators at regional and local levels.....63
Cross-sectional measures of income poverty at regional and local levels 64
Indicators of non-monetary deprivation at regional and local levels 64
Longit. indicators of income poverty and non-monetary deprivation at regional and local levels 65
Indicators of regional and local cohesion or disparities.....65
The third level indicators of poverty and deprivation.....66
3.2. National systems of social indicators.....84
The German System of Social Indicators.....84
The British Household Panel Survey (BHPS).....87
Measures of poverty and social exclusion in France for regional governments.....89
Measures of deprivation in the United Kingdom for local governments 89

3.3. Dimensions of poverty indicators ........................................ 93
3.4. Strategy for defining regional and local indicators ........................................ 98
Choice of units ........................................ 99
The available data ........................................ 100
*Sample surveys* ........................................ 100
*Employment of sample survey for poverty estimation beyond national level* ........................................ 111
*Meso data* ........................................ 111
*Sources of data at local level* ........................................ 112
3.5. Specific methods to estimate poverty and inequality measures at local level 113
Small Area Estimation (SAE) ........................................ 113
*Introduction* ........................................ 113
*The complementary roles of design and model* ........................................ 115
*Development of SAE methods* ........................................ 116
*Experiences in poverty estimation for small area in Poland* ........................................ 117
Poverty Mapping ........................................ 118
3.6 References ........................................ 120

# Prologue

The main objective of *Work Package 1* is to analyse the mechanisms and the determinants of poverty and inequality and to translate them into effective indicators.

We deal with two main tasks: the first one is the development of new poverty indicators covering both monetary and non monetary aspects and the second one is the construction of poverty and inequality measures at local level from several waves (pooled estimates) and the comparison between different EU-SILC waves results with focus on the local longitudinal changes.

This work aims to provide the literature review which is the basis of the whole *Work Package 1.* It is organized in three parts. We begin Part 1, *Poverty indicators in fuzzy and non-fuzzy approach*, with a review of traditional poverty measures and several multidimensional approaches as well as the fuzzy measures. We propose a new approach that combines the TFR approach of Cheli and Lemmi (1995) and the approach of Betti and Verma (1999) and then we conclude this part with a background of the Laeken indicators.

Next, in Part 2, *Pooled estimates of indicators*, we clarify the concept of pooling and its fundamental objectives. We illustrate, with many examples, four different scenarios that depend on whether the populations and data sources involved in the pooling are different or are the same.

Finally, Part 3, *Poverty and inequality measures for Regional and Local Governments*, deals with the choice of appropriate indicators at regional and local level, taking into account monetary and non-monetary cross-sectional measures as well as longitudinal measures. We conclude illustrating specific methods to estimate poverty and inequality measures at local level (SAE, poverty mapping).

# Chapter 1

# Poverty indicators in fuzzy and non fuzzy approach (coordinator Achille Lemmi)

## 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 income1 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.

**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.

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:

[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.

1. *Scale invariance*: poverty measures should be unchanged if, for example, a population doubles in size while everything else is maintained in the same proportions;
2. *Focus axiom*: changes among better-off people below the poverty line do not affect measured poverty;
3. *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;
4. *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;
5. *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;
6. *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:

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

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

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

where $\varepsilon$ is a measure of the sensitivity of the index to poverty. For $\varepsilon = 0$, *FGT*(0) coincides with the headcount index, when $\varepsilon = 1$ *FGT*(1) is the poverty gap index and for $\varepsilon = 2$, *FGT*(2) is the poverty severity index. For $\varepsilon > 0$ this measure is strictly decreasing in the living standard of the poor. Furthermore, for $\varepsilon > 1$ 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:

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

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:

$$S = H\lfloor I + (1-I)G_p \rfloor \tag{1.1.6}$$

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.

**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:

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

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:

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

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, 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[2]. 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 $\mu$ 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.

[2] N. C. Kakwani (1980), *Income Inequality and Poverty. Methods of Estimation and Policy Applications*, Oxford UP, New York, Oxford, London.

## 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.

**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.

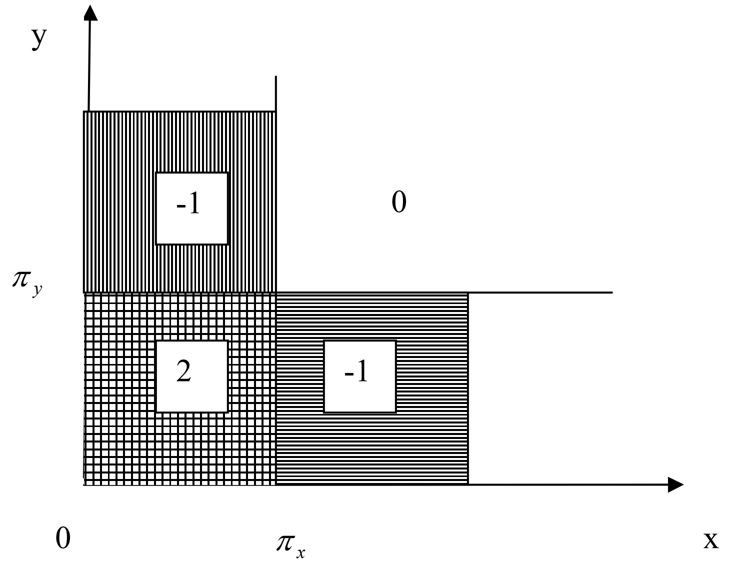
Dalton (1920) was the first to argue that economist were interested in the effects of inequality on economic welfare and that inequality in a distribution should be measured by the loss in welfare that it causes.

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 $\pi_x$ and $\pi_y$ 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 $\alpha$ is a measure of concavity of the function – *p*(*x*, *y*), $\beta$ 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.

**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:

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

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

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:

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

Alternatively, one can count a person poor if she is poor in any dimensions 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.

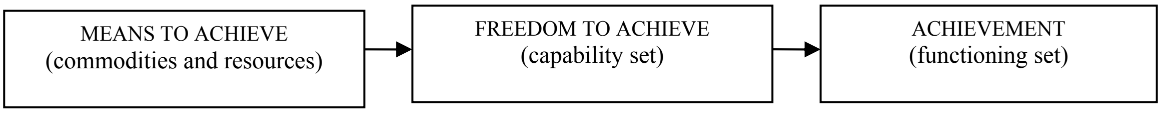
**Sen’s Capability Approach**

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:

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

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.

**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:

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

where

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

is the input set of all input vectors *x* which can produce the output vector *y*.

It holds (Coelli *et al.*, 1998) that:

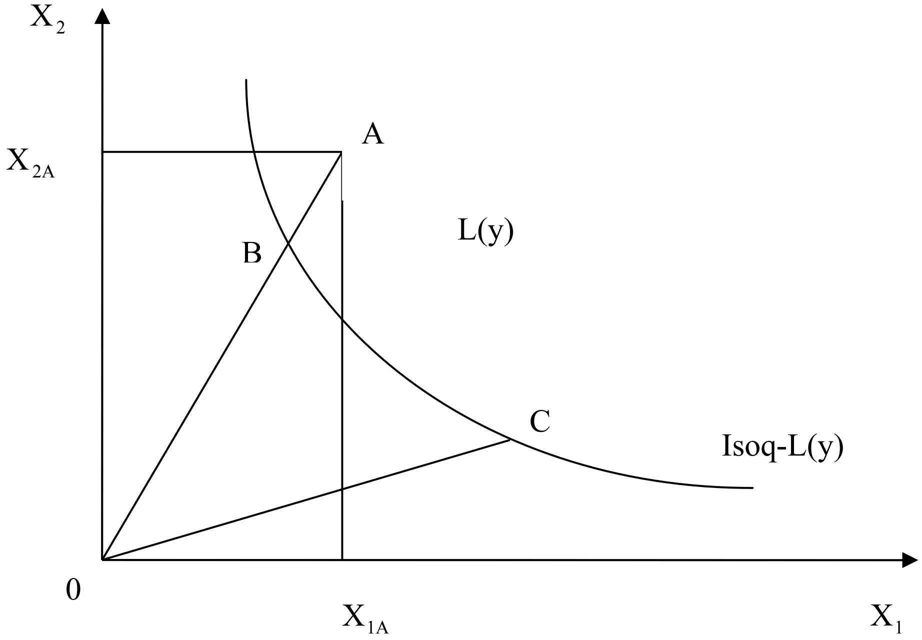
1. The input distance function is increasing in x and decreasing in y;
2. It is linearly homogeneous in x;
3. If x belongs to L(y) then $D_{in}(x, y) \geq 1$;
4. $D_{in}(x, y) = 1$ if x belong to the frontier of the input set (isoquant of y).

Graph 1.2.2 shows the concept of distance function. Here *q* and *q'* are respectively the input vectors corresponding to OA and OB. $\rho$ 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 $\rho$ 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



**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 $p_i = \Pr ob(x = x_i)$, $0 \le p_i \le 1$, 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:

$$h(p) = \log(1/p) = -\log(p) \tag{1.2.9}$$

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

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

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:

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

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 $x_{i1},\ldots,x_{im}$ 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 $x_c$.

### Axiomatic Approach

The axiomatic approach has been developed by Tsui (1995, 1999, 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):

1. *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.
2. *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 $x_{ij}$.
3. *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 $x_{ij}$.
4. *Principle of Population*: An $m$-fold replication of $X$ will not affect the value of the poverty index.
5. *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 $x_{ij}$.
6. *Non-Poverty Growth*: If the matrix $Y$ is obtained by adding a rich person to the population defined by $X$, then $P(Y; z) \leq P(X; z)$.
7. *Non-decreasingness in Subsistence Levels of Basic Needs*: If $z_j$ increases for any $j$, $P(X; z)$ does not decrease.
8. *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.
9. *Normalization*: $P(X; z) = 1$ whenever $x_{ij} = 0$ for all $i$ and $j$.
10. *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:

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

11. *Factor Decomposability*:

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

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

12. *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.
13. *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:

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

where *f* is continuous, non-increasing and convex with *f*(0) = 1 and *f*(*t*) = *c* for all *t* ≥ 1 and *c* < 1 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

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

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

### 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.

**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:

1. $I_{/k-1}^{k}(i)$ gives the value of the index I when only (k-1) variables have been dropped and k is the rank of the variable i;
2. $I_{/1}^{1}(i)$ gives the value of the index *I* when this variable is the first one to be dropped;
3. $I_{/0}^{1}(i)$ 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*);
4. $I_{/2}^{2}(i)$ 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;
5. $I_{/1}^{2}(i)$ gives the value of the index *I* when only one variable has been dropped and the variable *i* has the second rank;
6. $I_{/n-1}^{n}(i)$ corresponds to the (*n*-1)! cases where the variable *i* is dropped last and is the only one to be take into account;
7. $I_{/n}^{n}(i)$ 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:

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

**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):

i. Symmetry,

ii. Poverty and deprivation Focus,

iii. Monotonicity,

iv. Principle of Population,

v. Non-Poverty Growth,

vi. Non-decreasingness in Subsistence Levels of Basic Needs,

vii. Scale Invariance,

viii. Normalization,

ix. Subgroup Decomposability,

x. Factor Decomposability,

xi. Transfer Axiom,

xii. Nondecreasing Poverty under Correlation Increasing Arrangement.

Some axioms are passed only for particular ranks of FGT index ($\varepsilon$ 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:

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

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 $\mu_A(x) : X \to [0,1]$ 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.

**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 $y'$ up to which an individual is definitely poor and a value $y''$ above which an individual is definitely not poor. Formally:

$$\mu_A = 1 \qquad \text{if} \quad 0 \leq y \leq y' \tag{1.3.2}$$

and

$$\mu_A = 0 \qquad \text{if} \quad y > y'' \tag{1.3.3}$$

For incomes between $y'$ and $y''$ the membership function takes value in [0, 1] and declines linearly. Formally:

$$\mu_A = \frac{y''-y}{y''-y'} \qquad \text{if} \quad y' < y \leq y'' \tag{1.3.4}$$

The traditional approach is a particular case of the fuzzy approach with $y' = y'' = z$.

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:

$$\mu_i = (1 - F_i) \tag{1.3.5}$$

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 $\alpha \geq 1$. Formally:

$$\mu_i = (1 - F_i)^\alpha = \left( \frac{\sum_{\gamma=i+1}^{n} w_\gamma}{\sum_{\gamma=2}^{n} w_\gamma} \right)^\alpha \quad ; \quad \mu_n = 0 \tag{1.3.6}$$

where $F_i$ is the income distribution function and $w_\gamma$ is the sample weight of individual of rank $\gamma$ ($\gamma = 1,...,n$) in the ascending income distribution. $1 - F_i$ is the proportion of individuals less poor than the person concerned with mean ½ by definition.

The value of $\alpha$ is arbitrary, but Cheli and Betti (1999) have chosen the parameter $\alpha$ 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 (1.3.6) in order to define what they called Fuzzy Monetary indicator (FM):

$$\mu_i = (1 - L_i)^\alpha = \left( \frac{\sum_{\gamma=i+1}^{n} w_\gamma y_\gamma}{\sum_{\gamma=2}^{n} w_\gamma y_\gamma} \right)^\alpha \quad ; \quad \mu_n = 0 \tag{1.3.7}$$

where $y_\gamma$ is the equivalised income and $L_i$ represent the value of the Lorenz curve of income for individual *i*, then $1 - L_i$ 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 $1 - L_i$ values equals (1+G)/2, where G is the Gini coefficient of the distribution.

### 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 several steps are necessary. 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).

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:

$$\mu_A(i) = \frac{\sum_{j=1}^{k} g(x_{ij}) w_j}{\sum_{j=1}^{k} w_j} \qquad i = 1, \ldots, n \tag{1.3.8}$$

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 $\psi_j^{'}$, $\psi_j^{''}$, $\psi_{ij}$ represent the score of categories $x_j^{'}$, $x_j^{''}$, $x_{ij}$ respectively, then:

$$g(x_{ij}) = \begin{cases} 1 & \text{if } \psi_{\mathrm{ij}} \le \psi_j^{'} \\ \dfrac{\psi_j^{'} - \psi_{ij}}{\psi_j^{''} - \psi_j^{'}} & \text{if } \psi_{\mathrm{j}}^{'} \le \psi_{ij} \le \psi_{\mathrm{j}}^{''} \\ 0 & \text{if } \psi_{\mathrm{ij}} \ge \psi_j^{''} \end{cases} \tag{1.3.9}$$

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

$$w_j = \ln \frac{1}{p_j} \tag{1.3.10}$$

where $p_j$ is the proportion of individuals with deprivation in variable $X_j$. Substituting (1.3.10) in (1.3.9) we obtain:

$$\mu_A(i) = \frac{\sum_{j=1}^{k} g(x_{ij}) \ln \frac{1}{p_j}}{\sum_{j=1}^{k} \ln \frac{1}{p_j}} \tag{1.3.11}$$

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:

$$P = \frac{|A|}{n} \tag{1.3.12}$$

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

$$g(x_{ij}) = H(x_j) \tag{1.3.13}$$

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

$$g(x_{ij}) = g(x_j^{(k)}) = \begin{cases} 0 & \text{if } x_{ij} = x_j^{(1)}; k = 1 \\ g(x_j^{(k-1)}) + \dfrac{H(x_j^{(k)}) - H(x_j^{(k-1)})}{1 - H(x_j^{(1)})} & \text{if } x_{ij} = x_j^{(k)}; k > 1 \end{cases} \tag{1.3.14}$$

where $x_j^{(1)},\ldots,x_j^{(m)}$ 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:

$$w_j = \ln(1/\overline{g(x_j)}) \tag{1.3.15}$$

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 $p_j$.

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). 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. To avoid redundancy in the choice of weights, Betti and

Verma (1999) proposed the item weights to comprise two factor: i) 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; ii) the second factor is taken as a function of the correlation of any item with other items, in such manner that it is not affected by the introduction of variables entirely uncorrelated with the item concerned, but is reduced proportionately to the number of highly correlated variables present.

As in the Fuzzy Monetary approach, the individual's degree of non-monetary deprivation $FS_i$ can be defined in two alternative manners:

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

$$\mu_i = FS_i = (1 - F_{(S),i})^{\alpha_s} \tag{1.3.16}$$

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

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

$$\mu_i = FS_i = (1 - L_{(S),i})^{\alpha_s} \tag{1.3.17}$$

where $F_{(S),i}$ is the value of the Lorenz curve of *S* for individual *i*.

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

### Combination

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

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 paper, 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 $\overline{a} = 1 - a$, $a \cap b$, $a \cup b$ 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 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 → (1-*w*)(Standard) + *w*(Algebraic)
2. For sets representing dissimilar states → (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* < 1, 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 | $\min(1-FM_i, 1-FS_i) =$<br>$1-\max(FM_i, FS_i)$ | $\max(0, FS_i - FM_i)$ | $1-FM_i$ |
| | poor | $\max(0, FM_i - FS_i)$ | $\min(FM_i, FS_i)$ | $FM_i$ |
| | Total | $1-FS_i$ | $FS_i$ | 1 |

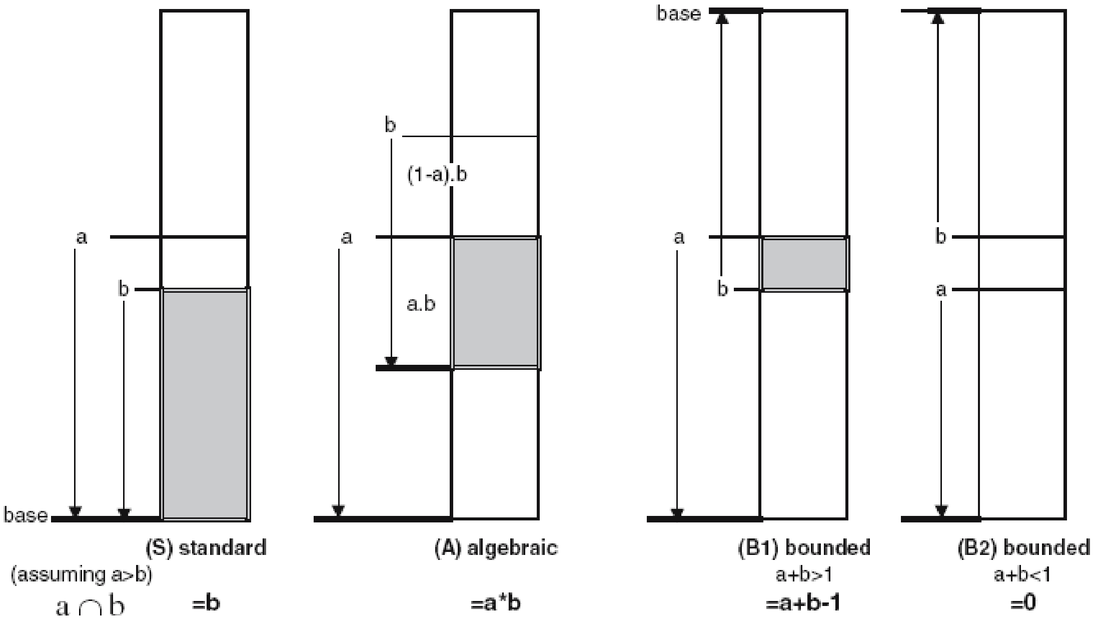
In the Graph 1.3.1, that shows intersections, the degree of membership in the "universal set" X is represented by a rectangle of unit length and the individual's memberships on the two subset (say, $0 \le a \le 1$, $0 \le b \le 1$ and their complements) have been placed within it. Different forms of fuzzy set operations (Table 1.3.1) are reproduced by different placements of the subset memberships within the rectangle for X. The Standard form, appropriate for similar sets, is represented by placing the two memberships (a, b) on the same base, so that their intersection is min(a, b) and union is max(a, b). In the Bounded form, appropriate for dissimilar sets, the two sets are placed et the opposite ends of X, thus their intersection is max(0, a+b-1) and union is min(1, a+b). Similarly, we can represented fuzzy sets unions.

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:

1. *Manifest deprivation* ($M_i$), representing the propensity to both income poverty and non-monetary deprivation simultaneously:

2. *Latent deprivation* ($L_i$), representing the individual being subject to at least one of the two, income poverty and/or non-monetary deprivation.

**Figure. 1.3.1**. The composite fuzzy set operations: a graphical representation of intersections



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 $FS_i$:

$$M_i = \min(FM_i, FS_i) \qquad (1.3.18)$$

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 $FS_i$:

$$L_i = 1 - \min(\overline{FM}_i, \overline{FS}_i) = \max(FM_i, FS_i) \qquad (1.3.19)$$

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 Laeken indicators

### Background of the Laeken indicators

We start providing some necessary background as regard to poverty indicators.

At the March 2000 Lisbon European Council, European Union (EU) Heads of State and Government declared 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”. In seeking to make this decisive impact on the eradication of poverty and social exclusion, they agreed to adopt the “open method of co-ordination”. This method involves the definition of a set of common objectives on poverty and social exclusion for the EU as a whole (which were agreed a few months later, at the December 2000 Nice European Council), the preparation of National Action Plans on social inclusion that Member States have to submit to the European Commission (the first NAPs/inclusion were submitted during the summer 2001), the exchange of good practices across Member States through so-called peer reviews; and the adoption of common indicators to monitor progress towards the common objectives and encourage mutual learning. Indicators ”can also prove useful for illustrating areas where more policy action is needed” (Atkinson *et al*., 2004).

To concretely implement this method in the area of social inclusion, and hence to allow an efficient fight against poverty and social exclusion at EU level, it is essential to be in a position to accurately measure where we are now and progress made towards the agreed objectives on the basis of comparable, quantitative information. It is precisely for this purpose that the Laeken European Council in December 2001 endorsed a first set of 18 common statistical indicators for social inclusion, organised in a two-level structure of 10 primary indicators – covering the broad fields considered to be the most important elements leading to social exclusion – and 8 secondary indicators – intended to support the lead indicators and describe other dimensions of the problem. The set of common indicators is supplemented by country specific indicators, according to data availability in individual countries.

The mentioned indicators take account the methodological research commissioned by the Belgian Presidency of the EU for this specific purpose (see Atkinson T. *et al*., 2002). The report on indicators for social inclusion prepared by the Social Protection Committee and endorsed in Laeken can be found on the web-site of Directorate General Employment and Social Affairs of the European Commission (www.europa.eu.int).

On the basis of the mentioned methodological principles of indicators construction, the Indicators Sub-Group has continued to refine and consolidate the original list of "Laeken indicators". The recent Indicators Sub-Group (ISG) work has resulted in a new list of common indicators, which by supplementing or modifying the Laeken indicators, is to reflect better the priorities given by the European Commission and member states to specific issues linked to social exclusion.

The modified list of indicators, which takes into account the re-launching of the Lisbon Strategy, as well as the streamlining (starting with 2006) of: the social inclusion processes, health care and social protection, and modification of the original list of Laeken indicators, reflect all the recent changes and policy reorientations that were proposed. This aims to support better interaction between all the processes defined in the Lisbon Strategy and Social Agenda.

The updated list, accepted in April 2008, covers three levels of indicators: primary indicators, secondary and context indicators (European Commission, 2008g). Level 1 covers the indicators considered as primary indicators, and cover the most crucial factors of social exclusion. Level 2 covers indicators which supplement level 1, providing the information necessary for a better understanding and interpretation of level 1 indicators. The list of context information is indicative and leaves room to other background information that would be most relevant to frame and understand better the national socio-economic context.

Newly adopted portfolio of social inclusion indicators are presented in the table 3.1. It indicates for each indicator the key dimension covered, the “name” and definition of each indicator and whether it is considered a commonly agreed EU indicator (EU) or a commonly agreed national indicator (NAT). Commonly agreed national indicators based on commonly agreed definitions and assumptions that provide key information to assess the progress of MS in relation to certain objectives, while not allowing for a direct cross-country comparison, and not necessarily having a clear normative interpretation. These indicators/statistics should be interpreted jointly with the relevant background information (exact definition, assumptions, representativeness).

These indicators need to be considered as a consistent whole reflecting a balanced representation of EU social concerns, rather than as a set of individual indicators. They now form a key basis for EU policy-making in the social area.

**Table 1.4.1**. The updated list of the Laeken indicators.

| **Symbol** | **Name**<br>**Commonly agreed EU indicator (EU)**<br>**Commonly agreed national indicators (NAT)** | **Source:** |
|---|---|---|
| **Primary indicators** | | |
| SI-P1 | EU: At-risk-of poverty rate + illustrative threshold values | EU-SILC |
| SI-P2 | EU: Persistent at-risk of poverty rate | EU-SILC |
| SI-P3 | EU: Relative median poverty risk gap | EU-SILC |
| SI-P4 | EU: Long term unemployment rate | LFS |
| SI-P5 | EU: Population living in jobless households | LFS |
| SI-P6 | EU: Early school leavers not in education or training | LFS |
| SI-P7 | NAT: Employment gap of immigrants | Relevant national data |
| SI-P8 | EU: Material deprivation (to be develop) | EU-SILC |
| SI-P9 | Housing (to be develop) | EU-SILC |
| SI-P10 | NAT: Self reported unmet need for medical care<br>NAT: Care utilisation | EU-SILC |
| SI-P11 | Child well-being (to be develop) | |

| Symbol | Name<br>Commonly agreed EU indicator (EU)<br>Commonly agreed national indicators (NAT) | Source: |
|---|---|---|
| **Secondary indicators** | | |
| SI-S1 | EU: At-risk-of poverty rate | EU-SILC |
| SI-S1a | EU: Poverty risk by household type | EU-SILC |
| SI-S1b | EU: Poverty risk by the work intensity of households | EU-SILC |
| SI-S1c | EU: Poverty risk by most frequent activity status | EU-SILC |
| SI-S1d | EU: Poverty risk by accommodation tenure status | EU-SILC |
| SI-S1e | EU: Dispersion around the at-risk-of-poverty threshold | EU-SILC |
| SI-S2 | EU: Persons with low educational attainment | |
| SI-S3 | EU: Low reading literacy performance of pupils | |
| **Context information** | | |
| SI-C1 | Income quintile ratio (S80/S20) | EU-SILC |
| SI-C2 | Gini coefficient | EU-SILC |
| SI-C3 | Regional cohesion: dispersion in regional employment rates | LFS |
| SI-C4 | Healthy Life expectancy and Life expectancy at birth, at 65, (by Socio-Economic Status when available | EUROSTAT |
| SI-C5 | At-risk-of-poverty rate anchored at a moment in time | EU-SILC |
| SI-C6 | At-risk-of-poverty rate before social cash transfers (other than pensions) | EU-SILC |
| SI-C7 | Jobless households by main household types | EU-SILC |
| SI-C8 | In-work poverty risk, breakdown full-time / part time | EU-SILC |
| SI-C9 | Making work pay indicators (unemployment trap, inactivity trap (esp. second earner case), low-wage trap | Joint EC-OECD project using OECD tax-benefit models |
| SI-C10 | Net income of social assistance recipients as a % of the at-risk of poverty threshold for 3 jobless household types | Joint EC-OECD project using OECD tax-benefit model |
| SI-C11 | Self reported limitations in daily activities by income quintiles, by sex, by age (0-17, 18-64, 65+) | |

The Joint Report on Social Inclusion (European Commission, 2003b; Section 10, *Use of Indicators*), and also the Report on Social Inclusion 2004 (European Commission, 2004) covering New Member States recommended to the construction of indicators at the national level, occasionally also concerning some subpopulations, such as children and minorities[3]. Nevertheless, they are equally pertinent to the development of appropriate indicators at the regional and local levels, and provide the necessary methodological framework and a starting point. Member states of the UE are encouraged to complement the set of common indicators with their own choice of country specified indicators (the third level indicators). These indicators should highlight national specificities as well as regional and local dimensions. In the same manner the common set of indicators should be complemented by specific regional and local indicators additionally taking into account national differences in this field. A deep discussion about ways in which the introduction

[3] See also European Commission, 2008b and 2008e.

of the regional dimension may make some fundamental differences in the choice of a "portfolio" of indicators is in Verma *et al*., (2005). The issues connected with indicator systems for monitoring poverty and social exclusion at regional and local levels were also taken up within the project carried out by UNDP-Poland (2006).

## 1.5. References

Alkire, S. and 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.

Atkinson A.B. (1970), On the measurement of Inequality. *Journal of Economic Theory*, 2, pp. 244-263.

Atkinson A.B. (2003), Multidimensional deprivation: contrasting social welfare and counting approaches. *Journal of Economic Inequality*, 1, pp. 51–65.

Berenger V. and 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., Cheli B., Cambini R. (2004), A statistical model for the dynamics between two fuzzy states: theory and application to poverty analysis. *Metron* 62, pp. 391-411.

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 23-26 May 2005.

Betti G., Cheli B., Lemmi A. and Verma V. (2006), Multidimensional and Longitudinal Poverty: an Integrated Fuzzy Approach, in A. Lemmi and G. Betti (eds.) *Fuzzy Set Approach to Multidimensional Poverty Measurement*, pp. 111–137, Springer, New York.

Betti G., Cheli B., Lemmi A., Verma V. (2007), The Fuzzy Approach to Multidimensional Poverty: the Case of Italy in the 1990s, in Kakwani N., Silber J. (eds.), *Quantitative Approaches to Multidimensional Poverty Measurement*, Palgrave Macmillan, pp. 30-48.

Betti G. and 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, Vol. 11, pp. 289-301, Lahore, Pakistan, August 27-31, 1999.

Betti G. and 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. and Verma V. (2004), A methodology for the study of multi-dimensional and longitudinal aspects of poverty and deprivation, Dipartimento di Metodi Quatitativi, University of Siena, Working paper 49.

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. and Chakravarty S.R. (2003), The measurement of multidimensional poverty. *Journal of Economic Inequality*, 1, pp. 25–49.

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.

Cheli B., (1995), Totally Fuzzy and Relative Measures of Poverty in Dynamics Context, *Metron*, Vol. 53, n.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.

Chiappero M.E. (2000), A Multidimensional Assessment of Well-being Based on Sen's Functioning Approach, *Rivista Internazionale di Scienze Sociali*, **108**, pp. 207–239.

Chiappero M.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.

Chakravarty S. (1983), A New Index of Poverty, *Mathematical Social Sciences*, **6**(3), 307–13.

Chakravarty S. R., D. Mukherjee, and R. R. Ranade (1998), On the Family of Subgroup and Factor Decomposable Measures of Multidimensional Poverty, in D. J. Slottje (ed.), *Research on Economic Inequality*, vol. 8, JAI Press, Stamford, CT and London.

Coelli T., D. S. Prasada Rao, and G. E. Battese (1998), *An Introduction to Efficiency and Productivity Analysis*, Kluwer Academic Publishers, Boston.

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. and Silber, J. (2005), Measuring multidimensional poverty: An empirical comparison of various approaches, *Review of Income and Wealth* **51**(1): 145-74.

Deutsch, J. and 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.

Dubois D., Prade H., (1980), *Fuzzy Sets and Systems*, Academic Press, Boston, New York, London.

Deutsch, J. and Silber, J. (2005), Measuring multidimensional poverty: An empirical comparison of various approaches, *Review of Income and Wealth* **51**(1): 145-74.

Dubois D., Prade H., (1980), *Fuzzy Sets and Systems*, Academic Press, Boston, New York, London.

Gini C. (1912), *Variabilità e Mutabilità*, Bologna, Tipografia di Paolo Cuppini.

Giorgi L. and Verma V. (2002), European social statistics: income, poverty and social exclusion, 2nd report. Office for Official Publications of the European Communities, Luxembourg.

Foster J., Greer J. and Thorbecke E. (1984), A Class of Decomposable Poverty Measures, *Econometrica*, 52, pp. 761-766.

Fine K. (1975), Vagueness, truth and logic, *Syntheses* **30**: 265–300, reprinted in Rosanna Keefe and Peter Smith (eds) (1996) *Vagueness: A Reader*, Cambridge, MA. and London: MIT Press.

Hagenaars A.J.M. (1986), *The Perception of Poverty*, North Holland, Amsterdam.

Klir G.J. and Yuan B. (1995), *Fuzzy sets and fuzzy logic, theory and applications*, Prentice Hall PTR, Upper Saddle River, New Jersey.

Kuklys W. (2005), *Amartya Sen's Capability Approach*: Theoretical Insights and Empirical Applications. Berlin: Springer Verlag.

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, Center for Economic Studies, Catholic University of Leuven, Belgium.

Lemmi A. and 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., S. Richardson, P. Travers, and L. Wood (1994), *Resources and Functionings: A New View of Inequality in Australia*, in W. Eichhorn (ed.), Models and Measurement of Welfare and Inequality, Springer-Verlag, Heidelberg.

Lugo M. A. and Maasoumi E. Multidimensional Poverty Measures from an Information Theory Perspective, ECINEQ WP 2008-85.

Maasoumi E. (1986), The measurement and decomposition of multi-dimensional inequality, *Econometrica*, **54**(4), pp. 991-997.

Maasoumi E. and 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.

Miceli D.(1997), *Mesure de la pauvreté. Théorie et Application à la Suisse*, Thèse de doctorat ès sciences économiques et sociale, Université de Genève.

Nolan B., Whelan C.T. (1996), *Resources, deprivation and poverty*. Clarendon Press, Oxford.

Qizilbash M. (2003), Vague Language and Precise Measurement: The Case of Poverty, *Journal of Economic Methodology*, **10**, pp. 41–58.

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.

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.

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, 379–423 and 623–56.

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**: 251-265.

Tsui, K. Y. (1999), Multidimensional Inequality and Multidimensional Generalized Entropy Measures: An Axiomatic Derivation, *Social Choice and Welfare*, **16**(1), 145–57.

Tsui, K. Y. (2002) Multidimensional poverty indices, *Social Choice and Welfare* **19**(1): 69-93.

United Nations Development Program (UNDP): Rapporto su "Lo sviluppo umano 8: Sradicare la Povertà (1997).

Whelan C.T., Layte R., Maitre B. and Nolan B. (2001), Income, deprivation and economic strain: an analysis of the European Community Household Panel, *European Sociological Review*, **17**, pp. 357-372.

Zadeh L.A. (1965), Fuzzy sets, Information and Control 8: 338-353.

# Chapter 2

# Pooled estimates of indicators (coordinator Vijay Verma)

## 2.1. Introduction

### 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*. 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: indeed, the problem of sample size requires a more sophisticated statistical approach than simply using direct estimates from one or more rounds of a sample survey.

(2) The second fundamental objective of pooling is to permit *comparisons*, for instance between different populations, between different 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.

Within each type of pooling, we can have a number of possibilities depending on whether the population and sample involved for the different elements in the pooling are the same or are different.

At the one extreme, we have the situation where both the population and the data sources involved are different: the data or estimates are being pooled across different population, using different sources of data in each.

Consider, for instance, pooled data sets or estimates produced from the Luxemburg Income Study (LIS, 1985). The LIS is a collection of micro-level data sets covering a large number of countries. The data sets are periodically updated. The characteristic feature of the data sets is that, while they are quite standardised in terms of the variables provided, the data may originally have been derived from different types of sources, differing to varying degrees in timing, coverage, concepts, etc.

### 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.

It is not possible to discuss the concept of comparability in detailed in this document except to provide a number of references in this literature review. See Verma (1992,1993, 1995a, 1995b, 1997, 1998a, 2002a, 2002b, 2004, 2006).

### Diverse scenarios

As noted above, different possibilities arise depending on whether the populations and data sources involved in the pooling are different or are the same.

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 2.1.1** Different types of situations involved in pooling

| | *Data source* | |
|---|---|---|
| *Population* | Same/Similar | Different /Dissimilar |
| Same/Similar | 4 | 3 |
| Different /Dissimilar | 2 | 1 |

Scenarios 2 and 4 are the ones most widely encountered in practice, but perhaps the other two scenarios present more complex technical problems.

## 2.2. Scenario 1. 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).

## 2.3. Scenario 2. 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.2.3.1.

### Examples of category 2.(a): aggregation of data or estimates

***Pooling of national estimates***

The Household Budget Survey, the Labour Force Survey, and the European Community Household Panel (ECHP) subsequently replaced by EU-SILC, are the three major social surveys conducted on a regular basis in countries of the European Union, in the order of increasing degree of inter-country comparability. ECHP and then EU-SILC which followed it form the most closely co-ordinated component of the EU system of social surveys: Though field implementation is handled by individual countries, all important aspects of design and statistical processing of the data are standardised through Eurostat (Verma 1995; Verma and Clemenceau 1996, Verma 2006). Much of the research using ECHP data has taken the form of inter-country comparisons and aggregations to construct the total EU picture.

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

$$\phi = \Sigma P_i.\phi_i \tag{2.3.1}$$

a choice has to be made of the weights $P_i$. The most common practice by far is to take the *Pi'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 $P_i$, (2.3.1) 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

$$w'_{ij} = w_{ij}.\left(P_i / \Sigma w_{ij}\right) \tag{2.3.2}$$

For ratios and relationships at the country level, estimates of $\phi_i$ are not affected by the scaling of the weights, and (2.3.2) gives the same results as obtained using the original weights $w_{ij}$. For aggregation over countries, (2.3.2) gives results identical to (2.3.1) when $\phi_i$ 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 (2.3.1) and (2.3.2) give identical results as except for the following.

The difference between (2.3.2) and (2.3.1) is that between 'combined' and 'separate' ratio or regression estimates. Normally form (2.3.2), 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 (2.3.1).

(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 (2.3.2). 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').

***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).

***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.

1. Only separate survey estimates $y_t$ may be presented.
2. Comparisons between the separate sites require harmonization to render the differences $(y_t - \bar{y}_t)$ meaningful.
3. Simple comulations $\bar{y}_t = \sum y_t / \sum n_t$ 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.
4. 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 $N_t$.
5. Weighted combinations $\bar{y}_w = \sum W_t y_t / \sum W_t$ weight the sites with some measure of their relative importance.
6. 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.

***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.

**Examples of category 2.(b): disaggregation for separate reporting by domain**

***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.

## 2.4. Scenario 3. 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)).

## 2.5. Scenario 4. 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.

### 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).

### 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.

### 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 $(n\text{-}i)/n$, up to $i$=$(n\text{-}1)$; 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.

### 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).

### 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.

## More robust poverty measures

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

### *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.

### *Poverty rates with 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.

### *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 2.5.6.1 and 2.5.6.2 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).

## 2.6. Example of cumulation in European surveys

In this section we provide two detailed examples of scenario (4) from European social surveys: namely from the continuous or annual Household Budget Surveys, and the rotational design of EU-SILC.

### 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.

1. The relatively moderate but regular workload;
2. 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

$$P = X + Y \tag{2.6.1}$$

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

$$P_O = (X + Y)/2 - 1 \tag{2.6.2}$$

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

$$P_M = (X + Y)/2 - 0.5 \tag{2.6.3}$$

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:

$$\text{for } Y = 1\text{: } P_M = 1;\ P_O = 1/2. \tag{2.6.4}$$

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

$$P_M = (P + 1)/2;\ P_O = P/2, \tag{2.6.5}$$

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).

### 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 2.6.1. 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 2.6.1.**

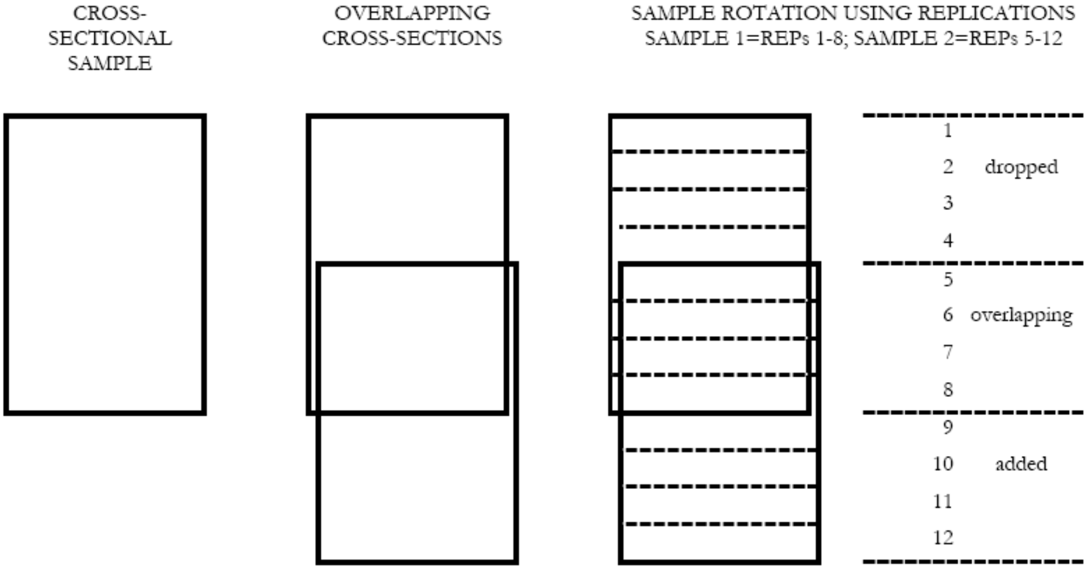

CROSS-SECTIONAL SAMPLE
OVERLAPPING CROSS-SECTIONS
SAMPLE ROTATION USING REPLICATIONS
SAMPLE 1=REPs 1-8; SAMPLE 2=REPs 5-12
1
2 dropped
3
4
5
6 overlapping
7
8
9
10 added
11
12

Figure 2.6.2 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 2.6.2**

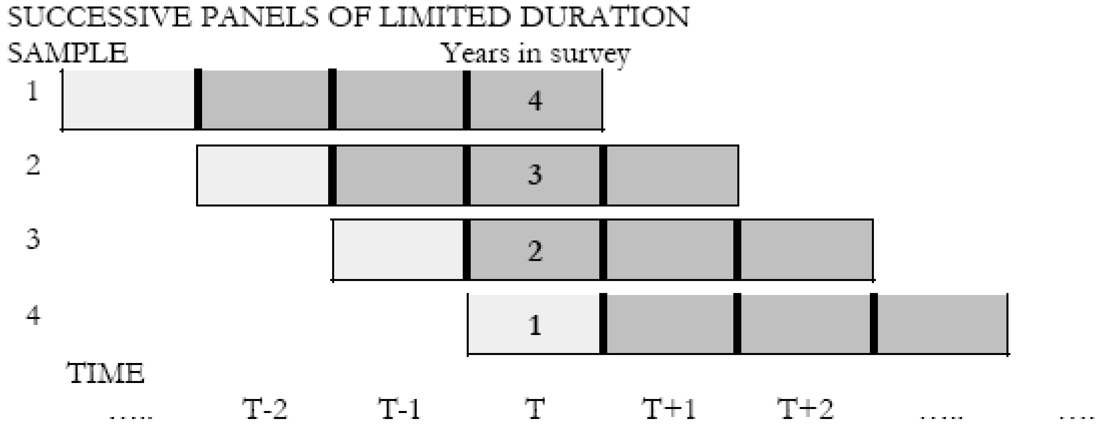
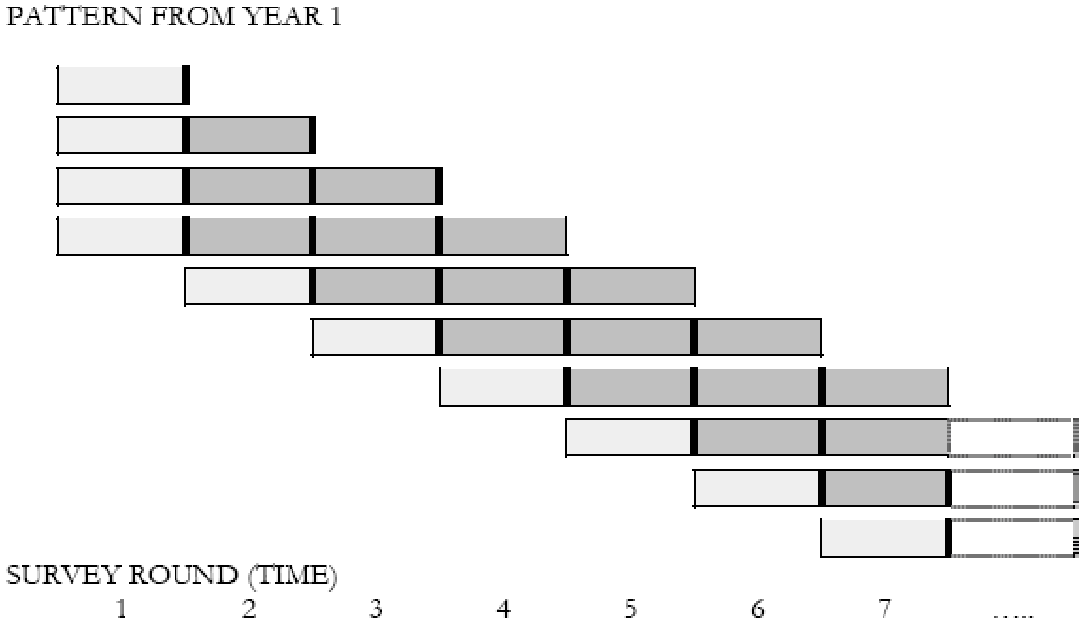


Figure 2.6.3 illustrates how a rotation pattern may be started from year 1. To obtain the full sample with 4 replications for the first year, it is necessary to begin with all the 4 replications. These replications are treated differently over time. One of these is dropped immediately after the first 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 2.6.3**



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.

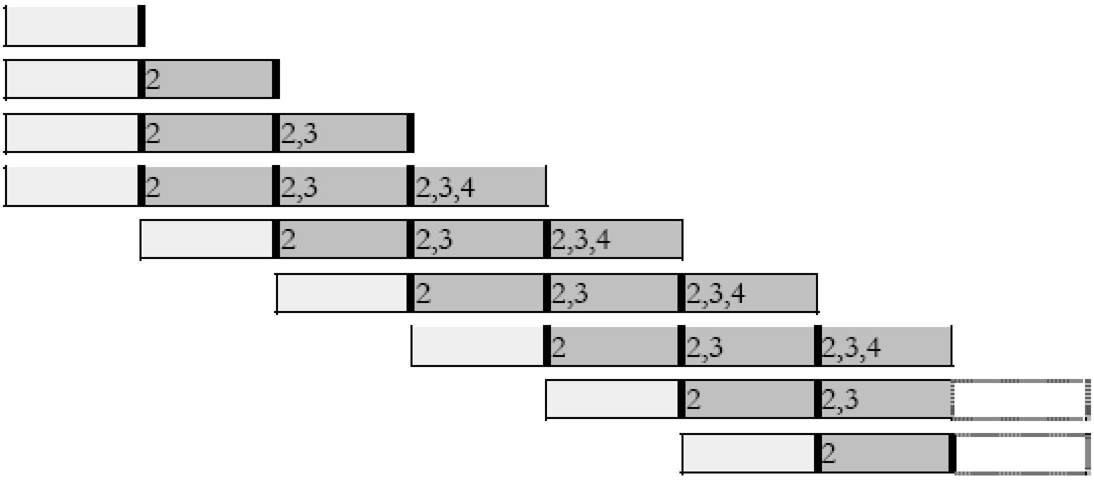
First consider year-to-year transitions in the design of *Figures 2.6.2-2.6.3*, with *r* subsamples for instance.

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 2.6.4 (based on Figure 2.6.3) 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 2.6.4**

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.

Adjusting the relative sample sizes of the two components

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 $\frac{n_{i+1}}{n_1} = \frac{r-i}{r+(x/s)}$.

## 2.7. Differences and averages under rotational design

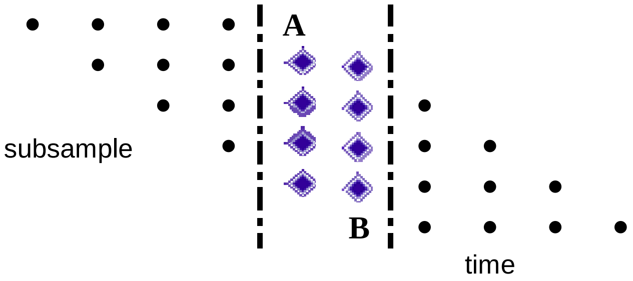
### Illustrations from the EU Labour Force Survey

An illustration of cumulating data and indicators over time in European labour force survey is based on Verma, Gagliardi and Ciampalini (forthcoming 2009).

Labour force survey are typically required to provide estimates of net change between two periods such as from one quarter to the next. Similarly, estimates are required of averages over a number of periods such as annual averages over consecutive quarters. With a rotational design, the total sample at any given period (say, quarter) consists of a number of subsample introduced into the survey at different points in the past. The sample overlap between two periods generally differs from one subsample to another. For example, *figure 7.1* shows a sample with a “linear” rotation pattern in which (1/5)th of the sample is replaced each quarter (i.e. any subsample remains in the survey for 5 consecutive quarters). Considering two consecutive quarters for instance, we see that the combined sample consists of 5 pairs, 4 of which are fully overlapping, while 1 is made up of two entirely independent samples (indicated by A and B in the figure).

In this section we consider technical issues in cumulating data under rotational and panel design, in particular how the variance and design effect are affected. The labour force surveys in EU countries provide the main illustrations. We conclude by considering variance and design effects of poverty trends and averages estimated in panel or rotational panel design, such as the ECHP and EU-SILC respectively.

**Figure 2.7.1**



Such differences between different subsamples in the sample overlap over time need to be taken into consideration in constructing measures of net change and averages over time for a rotational design.

For simplicity, we will assume that the rotation pattern is ‘linear’, where any unit once selected remains in the sample for *n* periods, thus giving a (*n*-1)/*n* sample overlap between successive periods. For any given period, the total sample is divided into *n* subsamples (each subsamples being representative of the population, just like the total sample), one of which is replaced by a different subsample in the following period. Next, we assume that variance of an estimate of interest from a single subsample at a given time has a constant value, say V$^2$; the average correlation between estimates, for two consecutive periods from a given subsample, is also assumed a constant, say R. This means that the difference between and the average over two consecutive periods estimated from a subsample have the following variances, respectively:

$$\begin{aligned} V_d^2 &= V_1^2 + V_2^2 - 2RV_1V_2 = 2V^2 \cdot (1-R) \\ V_m^2 &= (V_1^2 + V_2^2 + 2RV_1V_2)/2^2 = \frac{1}{2}V^2 \cdot (1+R) \end{aligned} \tag{2.7.1}$$

*Net change or difference between consecutive periods*

The situation is illustrated in the *figure 2.7.1* already given above. The total sample involved in the estimate consists of n pairs of subsamples, (*n*-1) of which are overlapping and 1 is a pair of independent subsamples. The resulting variances are as follows

**Table 2.7.1**

| Type of pair | No. of such pairs | Variance of estimate from one pair |
|---|---|---|
| Fully overlapping | (n-1) | $2V^2(1\text{-}R)$ |
| Independent | 1 | $2V^2$ |

Estimate of the difference form the whole sample may be obtained by simple average of the above estimates from individual pairs, *i*, of subsamples. Variance for such whole-sample estimate may be written as

$$V_d^2 = \sum V_i^2 / n^2 = 2V^2\left[(n-1)(1-R)+1\right]/n^2 = \frac{2V^2}{n}\left[1-\left(\frac{n-1}{n}\right)R\right] \qquad (2.7.2)$$

The coefficient (*n*-1)/*n* for R appears because one of the *n* pairs has no sample overlap. With full overlap in all the samples, the expression would have been

$$V_d^2 = \frac{2V^2}{n}(1-R) \qquad (2.7.3)$$

An alternative estimate with lower variance would be to take a weighted average, with the subsample estimates weighted inversely to their variance

$$V_d^2 = \sum W_i^2 V_i^2 \quad \text{with} \quad \sum W = 1. \qquad (2.7.4)$$

The appropriate weights each of the overlapping pairs are for: $W_i = 1/(n-R)$.

The non-overlapping pair is less efficient in estimating the difference, and consequently given a lower weight: $W_i = (1-R)/(n-R)$.

Variance of the resulting estimate is somewhat smaller than (2.7.2):

$$V_d^2 = 2V^2\left[(n-1)(1-R)+(1-R)\right]/(n-R)^2 = \frac{2V^2}{n}\cdot\frac{(1-R)}{\left(1-\frac{R}{n}\right)^2} \qquad (2.7.5)$$

*Averaging over quarters*

Consider a rotational sample with *n* subsamples as before. Let the required estimate be the average over Q consecutive periods, such as Q=4 quarters for annual averages. The subsamples contributing to the average estimate are illustrated in *figure 2.7.2* below for Q=4 and n values from 1 to 6.

The case n=1 corresponds simply to independent samples each quarter and, under the simplifying assumptions described above, the variance of the estimate of average over Q period is:

$$V_a^2 = \frac{V^2}{Q} \qquad (2.7.6)$$

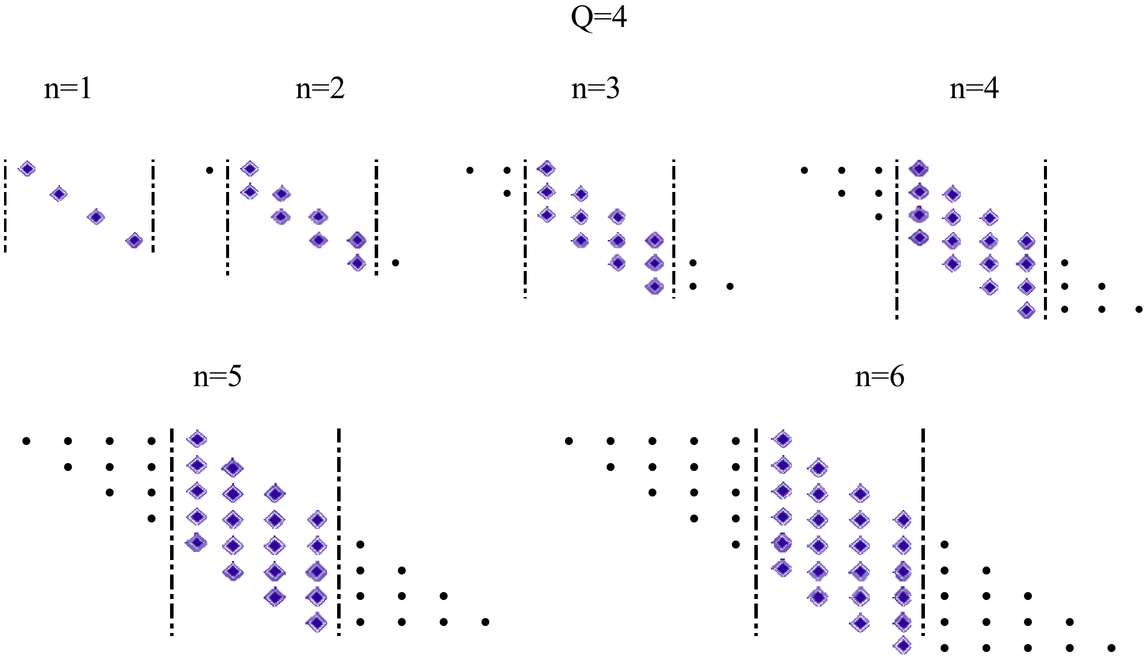
The total sample involved in the estimation of the average can be seen from the figure to consist of (*n*+Q-1) independent subsamples. Each subsample provides ‘observations’ with full sample overlap over a certain number of consecutive periods within the interval (Q) of interest. The distribution of the (*n*+Q-1) subsamples by the number of observation (*m*) provided can be seen to be as follows

**Table 2.7.2**

| m = | No. of subsamples |
|---|---|
| 1, 2, …, ($m_1$-1) | 2 for each value of m |
| m1 | $m_2$-($m_1$-1) |

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

**Figure 2.7.2**



For illustration, consider Q=m1=4, n=m2=6.

There are 2 contributing subsamples for each number 1, 2 and (m1-1)=3 of observations; and in addition there are m2-(m1-1)=3 subsamples, each contributing m1=4 observations.

Similarly, for Q=m2=4, n=m1=3, we have 2 contributing subsamples for each number 1 and (m1-1)=2 of observations, and in addition m2-(m1-1)=2 subsamples each contributing m1=3 observations.

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

$$V_m^2 = \frac{V^2}{m} \cdot (1 + f(m)) \tag{2.7.7}$$

Where

$$f(m) = \frac{2}{m} \cdot \{(m-1) \cdot R + (m-2) \cdot R^2 + \ldots + R^{m-1}\} \tag{2.7.8}$$

The term *f(m)* reflects the loss in efficiency in cumulation or averaging over overlapping samples. The following illustrates its values for various values of m:

**Table 2.7.3**

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

We may view the above as variance "per-observation", say $U_m^2$, for the subsample concerned

$$U_m^2 = m \cdot V_m^2 = V^2 \cdot (1 + f(m)) \tag{2.7.9}$$

Repeated observations over the same sample are less efficient in the presence of positive correlations; this is summarised by the factor *[1+f(m)]* where m is the number of repetitions.

In estimating the average using the whole available sample of $(n\cdot Q)$ subsample observations[4], 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

$$V_a^2 = \left(\frac{V^2}{n\cdot Q}\right)\cdot\left\{m_1\cdot\left[m_2-(m_1-1)\right]\cdot\left[1+f(m_1)\right]+2\sum_{m=1}^{m_1-1} m\cdot\left[1+f(m)\right]\right\}\Big/(n\cdot Q) = \left(\frac{V^2}{n\cdot Q}\right)\cdot F(R) \tag{2.7.10}$$

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

$$V_a^2 = \sum W_i^2\cdot V_i^2 \text{ , with } \sum W_i = 1 \tag{2.7.11}$$

where $W_i$ are the weights of observations *i*

**Illustration of poverty measures of average and net differences from a household panel**

Finally we report an example of the estimation of difference and averages of poverty measures using data from four consecutive waves of the ECHP survey (Betti et al., 2007). Suppose that W1, W2, W3 and W4 stand for the cross-sectional measures based on indicators of poverty/inequality for four consecutive years. Then nine measures of difference and average can be constructed as follows:

Measures of differences between waves: (W1-W2), (W2-W3), (W3-W4)

Measures of mean of two waves: (W1+W2)/2, (W2+W3)/2, (W3+W4)/2

Measures of mean of three waves: (W1+W2+W3)/3, (W2+W3+W4)/3

Measures of mean of four waves: (W1+W2+W3+W4)/4.

What is the relationship between the variance of a measure from a single wave, and what of more complex measures of averages and net changes of the above type?

This can be explained and estimated in terms of the “design effects” as follows.

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

The design effect is the ratio between the standard error of the measure considered based on the actual sample and the standard error of the same measure under the assumption of simple random sampling.

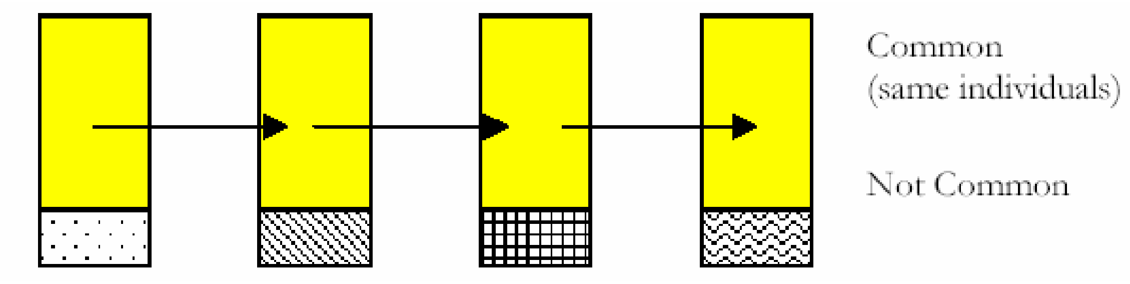
So, as example, the squared design effect for the measure of difference between first two waves can be expressed as:

$$Deft^2 = \frac{V(W1 - W2)}{V_{SRS}(W1) + V_{SRS}(W2)} \tag{2.7.12}$$

The design effect can be decomposed into components: the effect of clustering and stratification and the effect of weighting. Here, because of the panel nature of the data, we have also a third effect: the effect of correlation. This correlation arises from two sources. The first is the common structure (stratification and clustering) of the samples in different waves of a panel. This correlation exists even if there is no overlap between waves at the level of individual households or persons. Further correlation arises from overlap between waves at the individual level as follows.

We have four waves of a panel survey. Because of the panel nature of the survey, a large proportion of the individuals are common in these four waves. However, a small but generally non-negligible proportion of individuals are different from one wave to the other.

**Figure 2.7.3.**



As shown in the *figure 2.7.3*, the four cross-sectional samples largely overlap and are not independent. This causes correlation between measures from different waves.

For the measure of difference between two waves, *formula (2.7.12)* can be expressed as:

$$\frac{V(W1 - W2)}{V_{srs}(W1) + V_{srs}(W2)}$$

$$= \frac{V(W1 - W2)}{V_{rnd}(W1 - W2)} \cdot \frac{V_{rnd}(W1) + V_{rnd}(W2)}{V_{srs}(W1) + V_{srs}(W2)} \cdot \frac{V_{rnd}(W1 - W2)}{V_{rnd}(W1) + V_{rnd}(W2)} \tag{2.7.13}$$

where $V_{rnd}(.)$ is the standard error of the measure considered if we completely randomized the sample, i.e. all the elementary units put in a randomised order completely disregarding – hence completely removing the effect of – the sample structure.

The first term on the right hand side stands for the effect of clustering and stratification, the second term for the effect of weighting and the third term for the effect of correlation. In this expression we can estimate all the variance terms diversely, for example using a replication based variance estimation procedure such as JRR, except for $V_{srs}(.)$. For the calculation of the second term (the effects of weighting) that involves $V_{srs}$ (.), we proceed as follows. In Betti *et al.* (2007) we showed that $V_{srs}$ (.) can be substituted by $\frac{V_{rnd}(.)}{K(.)}$, where $K(.)$ stands for what we call the *"Kish factor"*. It can be calculated as follow, for instance, for poverty rate (p). Let

$$u_{1j} = (p_j - p) \tag{2.7.14}$$

In Verma et al. (2006), it has been empirically demonstrated that, at least for a wide variety of cross-sectional measures of poverty and inequality, the following expression very closely approximates the factor *K* expression:

$$K^2 = \left[\frac{n}{\sum w_j}\right] \cdot \frac{\sum w_j^2 \cdot u_{1j}^2}{\sum w_j \cdot u_{1j}^2} \tag{2.7.15}$$

Hence we can estimate the design effect of (W1-W2) as:

$$\frac{V(W1-W2)}{V_{srs}(W1)+V_{srs}(W2)} = \frac{V(W1-W2)}{V_{rnd}(W1-W2)} \cdot \frac{V_{rnd}(W1)+V_{rnd}(W2)}{\frac{V_{rnd}(W1)}{K(W1)}+\frac{V_{rnd}(W2)}{K(W2)}} \cdot \frac{V_{rnd}(W1-W2)}{V_{rnd}(W1)+V_{rnd}(W2)} \tag{2.7.16}$$

Similar expressions can be derived for the design effects for (W2-W3) and (W3-W4).

Similarly, the design effect for measures of average over waves can be expressed as follows:

$$\frac{V((W1+W2)/2)}{(V_{srs}(W1)+V_{srs}(W2))/4} = \frac{V((W1+W2)/2)}{V_{rnd}((W1+W2)/2)} \cdot \frac{V_{rnd}(W1)+V_{rnd}(W2)}{\frac{V_{rnd}(W1)}{K(W1)}+\frac{V_{rnd}(W2)}{K(W2)}} \cdot \frac{V_{rnd}((W1+W2)/2)}{(V_{rnd}(W1)+V_{rnd}(W2))/4} \tag{2.7.17}$$

Similar expressions can be derived for average of W2 and W3, and average of W3 and W4. For the average over three waves we have:

$$\frac{\mathrm{V}((\mathrm{W1}+\mathrm{W2}+\mathrm{W3})/3)}{(\mathrm{V}(\mathrm{W1})+\mathrm{V}_{\mathrm{srs}}(\mathrm{W2})+\mathrm{V}_{\mathrm{srs}}(\mathrm{W3}))/9}$$
$$=\frac{\mathrm{V}((\mathrm{W1}+\mathrm{W2}+\mathrm{W3})/3)}{\mathrm{V}_{\mathrm{rnd}}((\mathrm{W1}+\mathrm{W2}+\mathrm{W3})/3)}\cdot\frac{\mathrm{V}_{\mathrm{rnd}}(\mathrm{W1})+\mathrm{V}_{\mathrm{rnd}}(\mathrm{W2})+\mathrm{V}_{\mathrm{rnd}}(\mathrm{W3})}{\frac{\mathrm{V}_{\mathrm{rnd}}(\mathrm{W1})}{\mathrm{K}(\mathrm{W1})}+\frac{\mathrm{V}_{\mathrm{rnd}}(\mathrm{W2})}{\mathrm{K}(\mathrm{W2})}+\frac{\mathrm{V}_{\mathrm{rnd}}(\mathrm{W3})}{\mathrm{K}(\mathrm{W3})}}\cdot$$
$$\cdot\frac{\mathrm{V}_{\mathrm{rnd}}((\mathrm{W1}+\mathrm{W2}+\mathrm{W3})/3)}{(\mathrm{V}_{\mathrm{rnd}}(\mathrm{W1})+\mathrm{V}_{\mathrm{rnd}}(\mathrm{W2})+\mathrm{V}_{\mathrm{rnd}}(\mathrm{W3}))/9} \tag{2.7.18}$$

Similar expression can be written for the average of W2, W3 and W4. Finally, for the average over four waves we have:

$$\frac{V((W1+W2+W3+W4)/4)}{(V(W1)+V(W2)+V_{srs}(W3)+V_{srs}(W4))/16}$$
$$=\frac{V((W1+W2+W3+W4)/4)}{V_{rnd}((W1+W2+W3+W4)/4)}$$
$$\cdot\frac{V_{rnd}(W1)+V_{rnd}(W2)+V_{rnd}(W3)+V_{rnd}(W4)}{\frac{V_{rnd}(W1)}{K(W1)}+\frac{V_{rnd}(W2)}{K(W2)}+\frac{V_{rnd}(W3)}{K(W3)}+\frac{V_{rnd}(W4)}{K(W4)}}$$
$$\cdot\frac{V_{rnd}((W1+W2+W3+W4)/4)}{(V_{rnd}(W1)+V_{rnd}(W2)+V_{rnd}(W3)+V_{rnd}(W4))/16} \tag{2.7.19}$$

## 2.8 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, 29–30.

Australian Bureau of Statistics (1993), *The Australian Population Monitor*, Canberra, ABS.

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., Dourmashkin N., Rossi M.C., Verma V. and 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., Cheli B., Lemmi A. and Verma V. (2006), On the construction of fuzzy measures for the analysis of poverty and social exclusion. *Statistica & Applicazioni*, 4(1), 77-97.

Betti G., Gagliardi F., Nandi T. (2007), Jackknife variance estimation of differences and averages of poverty measures, Working Paper n°68/2007, Department of Quantitative Methods, University of Siena

Caplan D., Haworth M. and 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.

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.

Kish L., (1987), *Statistical Research Design*, New York: John Wiley & Sons.

Kish L., (1990), Rolling samples and censuses, *Survey Methodology*, 16, 1, 63-71.

Kish L., (1994), Multi-population survey designs,: five types with seven shared aspects, *International Statistical Review*, 62, 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, 31-46.

Kish L., (1998b), Combining multipopulation statistics, *Journal of Statistical Planning and Inference*, 102, 109-118.

Kish L., (1999a), Cumulating/ Combining Population Surveys, *Survey Methodology*, 25, 2, 129-138.

Kish L., (1999b), Combining Surveys: A Framework, *Proceedings of the 52nd Session of the International Statistical Institute*, Helsinki.

Laird N.M. and Mosteller F. (1990). Some Statistical Methods for Combining Experimental Results. *International Journal of Technology Assessment*, 6, 5-30.

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*, 584-A2.

Verma V. (1991). *Sampling Methods: Training Handbook*. Tokyo: Statistical Institute for Asia and the Pacific (SIAP).

Verma, V. (1992). Household surveys in Europe: some issues in comparative methodologies. *Seminar on International Comparison of Survey Methodologies*, Athens. Luxembourg: Training of European Statisticians (TES).

Verma, V. (1993). Comparative surveys in Europe: problems and possibilities. *Bulletin of the International Statistical Institute*, vol. 55.

Verma V. and A. Clemenceau, (1993), Methodology of the European Community Household Panel, *Statistics in Transition,* 2(7), 1023-1062.

Verma, V., (1995a), Comparative surveys in Europe: problems and possibilities*, Bulletin of the International Statistical Institute* 49(2), 527-8.

Verma, V. (1995b). European Community Household Panel and other comparative social surveys in the EU. *Euroconference on Social Policy in an Environment of Insecurity*, Lisbon. European Association for the Advancement of Social Sciences.

Verma, V. (1995c). Structuring and integration of household surveys in the European Community. In *The Future of European Social Statistics: Guidelines and Strategies*. Luxembourg: Office for Official Publications of the European Communities, Eurostat Series 0D, (ISBN 92-827-4969-X).

Verma, V. (1997). Comparability in multi-country survey programmes. *American Statistical Association Joint Statistical Meetings*, Special Session in Memory of Professor P.V. Sukhatme, Anaheim, California, USA.

Verma, V. (1998). Robustness and comparability in income distribution statistics. *Invited Paper*, European Union High Level Think-Tank on Poverty Statistics, Stockholm.

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. (2002a). Comparability in International Survey Statistics. *Keynote Address*, International Conference on Improving Surveys, Copenhagen, 25-28 August.

Verma, V. (2002b). Comparability in Multi-country Survey Programmes. *Journal of Statistical Planning and Inference*, vol. 102(1), pp. 189-210.

Verma, V. (2004). Comparability of statistics at the international level: concepts, approaches, methods. *Invited Lecture* at Comparabilité, harmonisation et intégration de données dans la construction de systèmes statistiques. Neuchatel: Swiss Statistical Society, Section of Official Statistics.

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. (2006). Issues in data quality and comparability in EU-SILC. Paper presented at *Comparative EU Statistics on Income and Living Conditions: Issues and Challenges*. Eurostat and Statistics Finland Conference, Helsinki, Finland.

Verma V. and Betti G. (2006). EU statistics on income and living conditions (EU-SILC): choosing survey structure and sample design. *Statistics in Transition*, 7(5), pp. 935-970.

Verma V., Gagliardi F, Ciampalini G (forthcoming 2009). Methodology of labour force survey in the EU: sample rotation patterns, Working Paper, Department of Quantitative Methods, University of Siena.

# Chapter 3

# Poverty and inequality measures for Regional and Local Governments (coordinator Tomasz Panek)

## 3.1 Introduction

Indicators of poverty and social exclusion are an essential tool for monitoring progress in the reduction of these problems. In the EU-wide context, these indicators need to be comparable across countries and time. For this purpose, the European Commission has adopted a common set of indicators, referred to as the Laeken Indicators (see Section 1.5). Most of these indicators are defined at the national level. However, indicators of poverty and social exclusion have also an important territorial dimension which is connected with the need to take into account regional and local differences in construction the system of these indicators. The important reason why regional and local levels should be considered in building system of poverty and social exclusion indicators it that many member states of the EU are decentralising decision-making, resources and responsibilities to lower levels of government. As a result, regional and local governments have greater opportunity to address problems of poverty and social exclusion. It causes also that in the construction of the National Actions Plans (national action plans for combating poverty and social exclusion) not only central governments are involved but also regional, or local governments. Moreover, to ensure a good allocation of public funds system of poverty and social exclusion indicators at regional, and local levels is necessary.

Regional and local governments could implement more effective poverty and social exclusion alleviation if local decision makers had better tools and strategies for prioritising actions and evaluating their outcomes. Just, an appropriate system of regional and local indicators help local governments develop, implement and evaluate programs to combat poverty and social exclusion. In order to construct an appropriate system of indicators at regional and local levels three main aspects have to be considered: choice of appropriate indicators at regional and local levels; making the best use of available data; using different sources of information to produce the best possible estimates for regions and local units using appropriate small area estimation (SAE) techniques.

For monitoring progress in the reduction of poverty and social exclusion the EC phrased the principles which the single indicator and the system (portfolio) of indicators as a whole have to follows (Atkinson *et.*

*al*., 2002). These principles refer to construction of the indicators at national level as well as at regional and local levels.

Principles applied to single indicators cover:

- an indicator should identify the essence of the problem and have a clear and accepted normative interpretation;
- an indicator should be robust and statistically validated;
- an indicator should be responsive to affective policy interventions but subject to manipulation;
- an indicator should be measurable in a sufficiently comparable way across members states and comparable as far as practicable with the standards applied internationally;
- an indicator should be timely and susceptible to revision;
- the measurement of an indicator should not impose too large a burden on member states, on enterprises, or on the Union's citizens.

The system of indicators should satisfy the following principles:

- the portfolio of indicators should be balanced across different dimensions;
- the indicators should be mutually consistent and the weight of single indicators in the portfolio should be proportionate;
- the portfolio of indicators should be as transparent and accessible as possible to the citizens of the European Union.

We may distinguish three main approaches to system of indicators construction, which are connected with three ways of the analysis of poverty and social exclusion (UNDP-Poland, 2006). In the first approach, which could be described as "one bag" approach, all the indicators describing the poverty and social exclusion make up one set, without differentiate between factors and symptoms indicators and indicators of strategic response to poverty and social exclusion.

Such approach is just typical for set of Laeken indicators to monitor poverty and social exclusion. The characteristic feature of this approach is the grouping of indicators into dimensions of poverty and social exclusion, without bringing up the cause-and-effect relations. This set of indicators, in which indicators of causes and symptoms are parts of the same set, are supplemented, by countries in their National Action Plans, with indicators specific to a given country.

The second approach uses certain elements of the cause-and-effect analysis. As an example the French system of social indicators may be indicated. In this system indicators are divided into three groups: indicators of context/environment, indicators of means and indicators of performance. Also OECD uses indicators designed to measure the scale of social exclusion and possible effectiveness of social response to this phenomenon (OECD, 2006).

The third approach basis on complete cause-effect sequence of indicators. This approach leads to construction a complete chain of cause-and-effect interactions in the analysis of poverty and social exclusion. Indicators are divided into three main functional group:

- indicators of causes (factors, determinants),

- indicators of states (syndromes),
- indicators of reaction (responses).

This approach was applied for example, by UN agencies (UNCSD) in all spheres, that is environmental, social, economic and institutional polices. A preliminary proposal for building the system of indicators for analysis poverty and social exclusion on the basis of cause-effect sequence of indicators was presented and the profit carried out by UNDP-Poland (2006). It was stressed that only this approach allow designing and evaluating policies in a consistent manner on all levels; national, regional and local.

Poverty indicators can be of different type and can have different properties. A poverty measure can be "relative", i.e. defined in relation to mean or median incomes or "absolute", i.e. based on ability to afford a given bundle of goods and services. Some indicators are based on a person or household's current status, but they can be also dynamic.

Current living conditions of households not depend only on their current income. This means that the portfolio of poverty measures have to be multidimensional covering not sole the monetary nature of phenomenon but also a deprivation in terms of life-style concerning a range of fields.

These should be complementary sources of information. Moreover, an indicator can be "objective", i.e. the status of individuals or households can be verified by documentary evidence or "subjective", i.e. based on a subjective judgement by the respondent. Finally, indicators may have an important territorial dimension reflect regional and local differences.

Using indicators for comparisons some problems can arise, like differences among countries and within countries differences.

### Choice of appropriate indicators at regional and local levels

As a point of departure for construction a system of poverty indicators the methodological framework endorsed at Laeken will be used. In order to choice a set of appropriate indicators of poverty for use by regional and local governments it is necessary to identify special features and requirements at regional and local levels. Specifically, the requirement is to identify whether, and if so in what manner, indicators appropriate for the regional and local level may differ from the Leaken indicators designed primarily for the national level.

Indicators of poverty of course have an important territorial dimension, pointing to the need to take into account regional and local differences. In an ideal context, one may seek to give regional and local breakdown on all indicators common for the all UE members. That is, one may introduce regional and local analyses within each of the indicator fields, for instance producing poverty rates at regional and local levels, urban-rural classification, etc. However, simply the introduction of more extensive breakdown is neither possible because of data limitations, nor sufficient in itself.

Some of the Laeken indicators may be suitable for regional and local application; others may be suitable after modification; while some may not be appropriate for the purpose. In addition, it is also necessary to consider whether there is need for addition to the existing indicators developed primarily for application at national level-region and local specific indicators able to capture aspects which are essentially regional or local. It is possible that a more diverse-"portfolio of indicators" is required for the purpose of addressing concerns of regional and local policies.

### Cross-sectional measures of income poverty at regional and local levels

Henceforth the Laeken indicators have been applied primarily at the national level. It is necessary to adapt them for regional and local application, taking into account differences in the requirements and the data situations. As a general rule, it is necessary to focus on the more basic among the indicators. This is because the data requirements are substantially increased when the results are to be geographically disaggregated.

Detailed disaggregation of the indicators not only by age, gender and other characteristics – but even only by geographical region – has to be severely restricted, especially when the information comes from sample surveys of limited size. Broad classification, such as distinguishing children, youth and elderly persons, may be possible, but even that has to be subsidiary to the need for adequate regional breakdown.

For the purpose of regional and local indicators, the focus has to be primarily on ordinary poverty rates for the total population, possibly with some major breakdowns. Certain more complex poverty and inequality measures - measures which are more sensitive to details and irregularities of the empirical income distribution - are less suited for disaggregation to small populations and small samples. Examples are Gini coefficient, relative median at-risk-of-poverty gap, and at-risk-of-poverty rate before social transfers.

On the other hand, poverty rates have to be supplemented by other indicators not considered explicitly in the Laeken list. Perhaps the most important of these is simply the mean income levels of the regions and local units, the dispersion among which provides a measure of regional and local disparities. General entropy measures may also be useful because they can be decomposed into within and between region and local components.

### Indicators of non-monetary deprivation at regional and local levels

In addition to the level of monetary income, the standard of living of households and persons can be described by a host of indicators, such as housing conditions or access to other goods and services (so called non-monetary indicators). Some of the deprivation elements are connected with the access to goods and services at local level, while others at region or national levels. This situation makes it necessary to harmonise system of poverty and social exclusion indicators for local governments with system of these indicators at regional and national levels. The data required for the construction of non-monetary indicators are generally simpler to collect than detailed data on monetary incomes. This makes such indicators more convenient and suitable for regional and local analysis. The set of Laeken indicators non-monetary type, mandatory for EU members, may be supplemented on the basis of the EU Statistics on Income and Living Conditions (EU-SILC) survey.

It is also useful to combine monetary and non-monetary measures in order to study the extent to which they overlap. If individuals are subject both to income poverty and non-monetary deprivation simultaneously, their overall deprivation is more intense. Similarly, if they are subject to only one of the two, their deprivation is, in relative terms, less intense. On the same lines as the monetary poverty rate, we can construct non-monetary deprivation rates, and also rates of what we have termed manifest deprivation (representing the presence of both income poverty and non-monetary deprivation simultaneously), and latent deprivation (representing the individual being subject to at least one of the two, income poverty and/or non-monetary deprivation).

### Longitudinal indicators of income poverty and non-monetary deprivation at regional and local levels

Longitudinal indicators are less frequently used in social inclusion and other reports than cross-sectional indicators of poverty and exclusion. These indicators are more demanding on the data. In constructing regional and local indicators, the emphasis has to be shifted away from the study of trends over time and longitudinal measures to essentially cross-sectional measures. Furthermore, it is more appropriate to aggregate such measures over suitable time periods, so as to illuminate the more stable aspects of the patterns of variation across regions and local units. Simpler indicators will be more robust and less demanding on the data available. As to longitudinal indicators, it is preferable to focus on indicators defined over a short time periods. Furthermore, such measures should be aggregated over suitable time periods, so as to illuminate the more stable aspects of the patterns of variation across regions and local units. Simpler indicators will be more robust and less demanding on the data available. Where the available statistical data cover longer time periods, short-duration longitudinal indicators can themselves be averaged over time to obtain more robust measures. In specific terms, we define and construct in the following illustrations indicators based on the persistence of poverty over pairs of adjacent years:

- Persons are persistently poor over two consecutive years if, in relation to the poverty line specific to each of the years, they are classified as poor in both the years,
- Persons are in any-time poverty over two consecutive years if, in relation to the poverty line specific to each of the years, they are classified as poor in either of the years.

With a longer reference period of T years, assuming that the necessary time series of data are available, the (T-1) pair-wise persistent or any-time rates can be averaged over time to obtain more stable measures for regional and local comparisons. The choice of the appropriate reference period T for averaging depends, apart from data availability, on substantive and policy considerations. It is matter of trade-off between temporal and spatial detail. Perhaps a moving average over a 4 or 5-year period may be generally appropriate.

The longitudinal measures of income poverty can be generalised to multi-dimensional measures of deprivation of the type noted above[5]: any-time, persistent and continuous deprivation, in monetary and non-monetary dimensions, and also in the two dimensions in combination (the above defined latent and manifest forms). The basic cross-sectional rates of monetary and non-monetary deprivation can be combined with each other and then also over time using fuzzy set operations.

### Indicators of regional and local cohesion or disparities

Laeken indicator SI-C3, ‘regional cohesion (dispersion of regional employment rates)’ was proposed in an attempt to measure regional disparities in employment rates. However, alternatives are required to this indicator because of its statistical and substantive shortcomings[6]. This indicator has been criticised for not providing statistically valid information for comparison across countries because its magnitude depends on the size and number of regions present in the country. There also has been some criticism of the indicator

[5] For more details see Part 1.

[6] The same remarks may refer to local disparities.

from a substantive/policy angle (for instance, Atkinson *et al*., 2002). While clearly the proposed indicator needs to be improved from a statistical point of view, we do need similar indicators to synthesis the wealth of information contained in the regional and local breakdowns of the common indicators of social inclusion. This applies not only to employment rates, but also to regional and local disparities in the rates of unemployment, poverty and deprivation etc.

### The third level indicators of poverty and deprivation

The first and the second level Laeken indicators, common at UE level, represent only the starting point to constructing a system of indicators at regional and local levels. EU member states are encouraged to develop third level indicators which would reflect specifity of individual countries. These indicators would constitute supplement of indicators for all UE countries as part of the monitoring system connected with implementation of National Action Plans. Within NAPS, draw up in 2-3 years cycles, the social indusion goals (adopted on the EC level) defined in operational terms on the individual countries are provided. Below, an review of selected countries NAPS was conducted in the context of searching for examples of regional and local indicators system for monitoring poverty and the links and inter-relations of systems at different levels. Based on National Actions Plans – information about indicators for each country are presented in different form and order.

**Table 3.1.1.** Selected Indicators for monitoring of the implementation of the NAP/Inclusion in Poland[7]

| **Name of priority**[8] | **Name of the action** | **Name of the Indicator**[9] | **Source**[10] |
|---|---|---|---|
| Priority 1. Support families with children Indicator | | At risk of poverty rate among children (0-19 years) (threshold fixed at the level of the minimum of existence) (NAT) | CSO |
| | | At risk of poverty rate among families with 4 or more children (threshold fixed at the level of the minimum of existence) (NAT) | CSO |
| | Development of the integrated family support system | Expenditures for programme supporting parents before and after childbirth (NAT) | Registries of MPiPS, MZ, MEN |
| | | Expenditures for non-insurance based family benefits connected with childbirth (NAT) | Registries of MPiPS |
| | | Number of children covered by actions „National Disabled Children Support Programme" (NAT) | Registries of MZ, MEN |
| | | Number of centers of daily support care within regulation of social welfare [run by municipalities or by any other entities] (NAT) | Registries of MPIPS |
| | | Number of children and youth covered by actions "Recreation room – internship – sociotherapy in the rural environment" (NAT) | Registries of MPIPS |
| | | NAT: Number of the municipalities who participate in the programme | Registries of MPIPS |

[7] Ministry of Labour and Social Policy (2008).

[8] Table contains indicators for monitoring of the implementation of the NAP/Inclusion included in the three sections connected to three priorities. Every section presents appropriate indicators assigned to the one priority.

[9] EU - commonly agreed EU indicator , NAT - commonly agreed national indicator

[10] Abbreviations used: MPiPS: Ministry of Labour and Social Policy, MZ: Ministry of Health, MEN: Ministry of National Education, CSO: Central Statistical Office, PFRON: National Disabled People's Rehabilitation Fund

| | | | |
|---|---|---|---|
| | | „Recreation room – internship – sociotherapy in the rural environment" (NAT) | |
| | | Number of people covered by actions provided by centers of daily support care within regulation of social welfare [run by municipalities or by any other entities] (NAT) | Registries of MPIPS |
| | | Number of completed social houses (NAT) | Registries of the Ministry of Construction |
| | | Number of night shelters, hostels and homeless shelters [run by powiats or by other entities] (NAT) | Registries of MPiPS |
| | | Number of people who live in night shelters, hostels and homeless shelters [run by powiats or by other entities] (NAT) | Registries of MPiPS |
| | | Number of the Citizens Advice Bureaux (NAT) | Data of Union Citizens Advice Bureaux |
| | Development of the income support system | Expenditures for the system of family benefits (in PLN) (NAT) | Finances of MPiPS |
| | | Expenditures for the family benefits as a share of the total social expenditures in relation to the she of the people aged 0-18 years in the total population (EU) | Finances of MPiPS |
| | | At-risk-of-poverty rate among children (0-15 years) before all social transfers except old-age/survivors' pensions (EU) | CSO |
| | | The amount of paid scholarships (in PLN) | Registries of MEN |
| | | Number of children covered by the social and science scholarship | Registries of MEN |
| | | The percentage of pupils who receive scholarship | Registries of MEN |
| | | Number of children covered by programs of counteracting the malnutrition | Registries of MPiPS |
| | | Number of families using tax relieves | Registries of Ministry of Finance |
| | Supporting employees in | Number of people employed in nursery schools | CSO |

|  |  |  |  |
|---|---|---|---|
|  | reconciliation of work and family life |  |  |
|  |  | Number of people employed in care services for elderly, disabled people and long-term ill | CSO |
|  |  | Number of children attending nurseries in relation to 100 children aged 0-3 years | CSO, Statistical Yearbook |
|  |  | Number of children in nurseries and nursery wards | CSO, Statistical Yearbook |
|  |  | Number of children attending nursery schools in relation to 1000 children aged 3-5 years (NAT) | CSO, Statistical Yearbook |
|  |  | Number of places in nursery schools By the place of residents (urban/rural areas) | CSO, Statistical Yearbook |
|  |  | The percentage of people employed in the part-time job in the total working population (NAT) | CSO, BAEL |
|  |  | Employment rate of women aged 25-49 years (EU) | Eurostat, LFS |
| Priority 2. Inclusion by activation | Reform of tools and instruments for active inclusion | Number of social welfare beneficiares who are covered by social contracts (NAT) | Registries of MPiPS |
|  |  | Number of reimbursement of contributions to mandatory social insurance for employers who employ the disabled (NAT) | PFRON |
|  | Development of the public-social partnership | Number of projects which are carried out by the public-social partnerships (NAT) | Registries of MPiPS |
|  | Development of the social economy institution | Number of Centres of Social Integration | Registries of MPiPS |
|  |  | Number of people covered by activities provided by Centres of Social Integration (NAT) | Registries of MPiPS |
|  |  | Number of people covered by activities provided by Centres of Social Integration who obtained a job (NAT) | Registries of MPiPS |
|  |  | Number of Clubs of Social Integration (NAT) | Registries of MPiPS |

| | | | |
|---|---|---|---|
| | | Number of social co-operatives (NAT) | Registries of MPiPS |
| Priority 3: Mobilisation and partnership | Programming of social inclusion policy | Percentage of municipalities and powiats which implemented local strategies counteracting social problems | Registries of MPiPS |
| | | Percentage of municipalities and powiats covered by system of monitoring and evaluation of the proces of social integration | Registries of MPiPS |
| | Integration and development of social services | Number of social worker in relation to 2000 residents of municipality (NAT) | Registries of MPiPS |
| | | Number of employment agents working in powiat's labour offices (NAT) | Registries of MPiPS |
| | | Number of employment counsellors working In powiat's labour offices (NAT) | Registries of MPiPS |
| | | Percentage of social workers covered by educational programme relating to implementation of professional standards (NAT) | Registries of MPiPS |

**Table 3.1.2**. Social Exclusion Indicators in Finish NAP[11]

| Key dimension | Name of the group | Name of the Indicator | Definition |
|---|---|---|---|
| Economic Exclusion | Relative risk of low income / poverty | Number of low-income persons | Persons in households below the poverty level |
| | | Poverty risk – relative poverty level | % of population among different groups (i.e.: children, middle-aged, elderly, unemployed, etc.) (poverty risk is calculated on the basis of a household's disposable income, the poverty risk limit being 60% of median income each year) |
| | | Individual poverty rate before income transfers | - Percentage of persons living in households below the poverty level, based on net factor income + pensions (%)<br>- Percentage of persons living in households below the poverty level, based on net factor income alone (%) |
| | | Persistent poverty | - Percentage of persons below the 60% poverty level in at least three out of four years (%) |
| | Last-resort social welfare benefits | Income support | - Number of persons receiving income support during the year<br>- Percentage of persons receiving income support, % of the population (%)<br>- Number of households receiving income support during the year<br>- Percentage of single-parent households, % of households receiving income support (%)<br>- Number of households receiving income support for 10 to 12 months of the year (%)<br>- Percentage of households receiving income support for 10 to 12 months of the year (%) |
| | Indebtedness | Debt recovery | Persons subject to debt recovery, % of the population |

[11] Ministry of Social Affairs and Health (2006).

| | | | |
|---|---|---|---|
| Health Problems | Perceived state of health | Persons assessing their state of health as bad or fairly bad | Percentage of persons assessing their state of health as bad or fairly bad |
| | Functional capacity of pensioners | Age-adjusted persons aged 65 to 84 with problems in their ability to move | Percentage of persons aged 65 to 84 with problems in their ability to move (measured by climbing stairs) |
| | Social-based health differences | Life expectancy of 35-year-olds by social group | (managers = 100) |
| Exclusion From the Labour Market | Unemployment | Unemployment rate, % | among different groups (i.e.: foreigners, young people, etc.) |
| | | Long-term unemployment rate | Percentage of workforce unemployed for more than a year<br>among different groups (i.e.: foreigners, young people, etc.) |
| | | Long-term unemployed jobseekers registered with employment office | - Number of unemployed for over one year<br>- Number of  unemployed for over two years |
| | | Unemployed jobseekers with disabilities | - Number during the year<br>- % of unemployed jobseekers |
| | | Distribution of regional employment rates (NUTS2 level) | |
| | Measures to promote employment | Persons employed as a result of such measures | - Number of persons employed through wage-based measures, end-of-month average<br>- Number of persons participating in labour market training, end-of-month average |
| | Non-participation in work | Rate of non-participation during the year | - According to Income distribution statistics (Total persons living in households including at least one person of working age (18 to 59) but with no one employed during the year, percentage of all persons living in households including at least one person of working age (excluding student households))<br>- According to Labour force survey (in an unemployed  household no |

| | | | |
|---|---|---|---|
| | | | household member has been employed during the survey week) |
| Exclusion from the Housing Market | Accommodation problems | Households living in very inadequately equipped accommodation | Percentage of households living in very inadequately equipped accommodation |
| | | Persons living in cramped quarters | Percentage of persons living in cramped quarters |
| | | Households in queue for ARAVA rental housing | Number of households in queue for ARAVA rental housing |
| | Homelessness | Homeless persons | Total no. of unattached homeless persons |
| | | Homeless families | Total no. of unattached homeless families |
| Exclusion from Education | Inadequate schooling | Not completed comprehensive education | - Number of dropouts, those receiving a leaving certificate and those who left without a certificate |
| | | Young people with deficient education | Percentage of persons aged 18 to 24 who have only completed comprehensive education and are not in training, percentage of the age group |
| Other Exclusion | Children and young people threatened by exclusion | Young people who have finished school but are not at work, in education, military service, non-military service or in pension | - Percentage of persons aged 15 to 19 who have finished school but are not at work, in education, military service, non-military service or in pension<br>- Percentage of persons aged 20 to 24 who have finished school but are not at work, in education, military service, non-military service or in pension |
| | | Children and young people in open care | - Number of children and young people in open care |
| | | Children and young people placed outside the home | - Number of children and young people placed outside the home<br>- Children placed outside home,% of total aged 0 to 17<br>- Number of the above taken into custodial care |
| | | Prisoners | - Average per day<br>- Percentage of women |
| | | Violent crime | - Number of cases |

| | | | |
|---|---|---|---|
| | | | - Violent crime rate per 1,000 inhabitants, of crimes reported to the police. |
| | | Youth crime | - Number of persons aged 15 to 20 suspected of crimes investigated by police |
| | | Drugrelated crime | - Drug-related crime rate per 1,000 inhabitants, of crimes reported to the police<br>- Number of drugrelated crime<br>- Number of women |
| | | Suicides | - Number of persons<br>- Suicides per 10,000 inhabitants (average population) |
| | | Alcohol | - Number of alcohol-related deaths<br>- Number of alcohol ailment or similar as primary cause of death<br>- Number of deaths by accident or violence while intoxicated<br>- Total number of deaths directly or indirectly caused by alcohol |
| | | Persons treated in hospital for alcohol-related ailments | Alcohol-related ailment as main or subsidiary diagnosis |
| | | Drugs | - Number of deaths with forensic drug-related findings<br>- Number of persons treated in hospital for drug-related ailments (Drug-related ailment as main or subsidiary diagnosis: total, % woman) |
| | | Clients in open intoxicant care during the year | - Alcohol outpatient centres ('A clinics')<br>- Short-term treatment centres for young people |

**Table 3.1.3**. Tertiary Social Exclusion indicators in United Kingdom NAP [12] [13]

| **Key dimension** | **Name of the Indicator** | **Source** |
|---|---|---|
| Low Income | Persistent at-risk-of-poverty rate for children<br>Number of children living in absolute low income<br>Proportion of pensioners living in low income households<br>The number of pensioners on relative low incomes<br>Lone parent employment rate<br>The employment rate of ethnic minority<br>The employment rate of disabled people<br>Employment rate for the age group 50 to 69 | ----- |
| | Relative low income rate | Households Below Average Income, Department for Work and Pensions |
| | Absolute low income rate | Households Below Average Income, Department for Work and Pensions. |
| | Persistence of low income | British Household Panel Survey |
| Labour Market | Long term benefit dependency | LFS |
| | The employment rates of disadvantaged groups | LFS |
| | The Employment Rate for the Most Deprived Wards | LFS |
| | The proportion of people living in workless households | Not updated |
| Housing | The percentage of households in fuel poverty | English House Condition Survey and the Energy Follow-up Survey<br>The Northern Ireland House Condition Survey<br>Scottish House Condition Survey |

[12] DWP (2006)

[13] The tertiary indicators have been selected by the UK, so are not harmonized at the EU level. In addition to providing more timely and detailed information as outlined above, some of the indicators highlight areas which are strategic priorities for the UK.

| | | |
|---|---|---|
| | | Living in Wales Survey 2004 |
| | Households in non decent homes | English Housing Condition Survey |
| | The proportion of people who have ever accessed the internet | Not updated |
| | The Number of People Sleeping Rough | Local Authority "Housing Investment Programme" (HIP) Returns. Data cover England |
| | Homeless families with dependent children | Housing & Communities Analysis Division of the Department for Communities and Local Government |
| Education | The proportion of 16-year-olds without any GCSE or equivalent | GCSE/GNVQ examination results, England; Summary of Annual Examination Results,<br>Northern Ireland; Welsh Examinations database - Welsh Assembly Government |
| | The proportion of S4 pupils without any Standard Grades or equivalent | Scottish Executive Education Department. Data cover Scotland |
| | The proportion of 16-year-olds with five GCSEs grade A* to C or equivalent | Summary of Annual Examination Results, Northern Ireland; GCSE/GNVQ examination<br>results, England; Welsh Examinations database - Welsh Assembly Government. |
| | The proportion of S4 pupils with five Standard Grades 1 to 3 or equivalent | Scottish Executive Education Department. |
| | Schools achievement Key Stage 2 for literacy and numeracy - England | National Curriculum end of Key Stage 2 assessment tests - Department for Education<br>and Skills. |
| | achievement Key Stage 2 for literacy and numeracy - Wales | Welsh Assembly Government |
| | Proportion of publicly funded primary schools in Scotland where less than 65% of P7 pupils attained 5-14 level D or above by end of school year | No longer available |
| | The proportion of working-age people with at least a qualification at NVQ/SVQ level 2 qualification or equivalent | LFS |
| | The proportion of economically active adults in England with at least an NVQ/SVQ level 2 qualification or equivalent | LFS |

| | | |
|---|---|---|
| | The educational attainment of young people leaving care | Department for Education and Skills, England<br>Children Looked After, Scotland<br>Data collected by DHSSPS using the Outcome Indicator return (OC1, Northern Ireland).<br>The Local Government Data Unit, Wales. |
| Health | The rate of conceptions for those aged under 18 | Office for National Statistics; ISD Scotland, SMR01 and SMR02 returns |
| | The rate of births to mothers aged under 18 | Northern Ireland Statistics and Research Agency |
| | Infant mortality rates by social groups | Office for National Statistics (England and Wales).<br>Linked file – linking information on birth and death registrations;<br>Northern Ireland Statistics and Research agency;<br>GROS (General Register Office Scotland). |
| | The gap in life expectancy at birth between the "fifth of local authorities with the worst health and deprivation indicators" and the population as a whole | Office for National Statistics (life expectancy data based on population estimates and mortality statistics from death registrations). Data cover England. |
| | The proportion of people with a long term limiting illness | General Household Survey.<br>Northern Ireland Continuous Households Survey. |
| | Adult smoking prevalence by social group | Office for National Statistics<br>General Household Survey, England.<br>Continuous Household survey, Northern Ireland. Scottish Health Survey, Scotland. |
| Community | The proportion of people whose lives are greatly affected by fear of crime | British Crime Survey;<br>Northern Ireland Crime Survey;<br>Scottish Crime and Victimisation Survey |

Selected indicators of Social Exclusion in French NAP[14]

- Long-term employment: years spent in employment,
- Percentage of men and women working in stable jobs,
- The evolution of the percentage of men and women working in stable jobs,
- The percentage of men and women having a part time job and wishing to work more,
- The evolution of the percentage of men and women having a part time job and wishing to work more,
- Rate of access to employment for the active young who have exited the educational system for a period between 1 and 5 years,
- Unemployment rate for the active young who have exited the educational system for a period between 1 and 5 years,
- Percentage of people asking/looking for employment who have followed the classes of a professional training course during the last 12th months,
- Rate of long-term unemployment (as ratio of the unemployed longer than 1 year from the total population),
- Share of the long-term unemployed (longer than 1 year) from the unemployed in the sense intended by BIT (this indicator is calculated as a proportion between the number of the unemployed for a period longer than 12 months to the total number of unemployed),
- The proportion of those who exist long-term unemployment in the successive cohorts of those who enter (long-term unemployment),
- Percentage of young people who exist unemployment (DEFM1) before the 6th month (of unemployment) – annual mean,
- Percentage of adults who exist unemployment (DEFM1) before the 12th month (of unemployment) – annual mean
- Rate of the very long-term unemployed (more than 2 years) from the total number of the active population,
- Maintenance of children: number of places offered by the collective structures (apart from the maternelles system).

[14] UNDP - Poland (2006).

Selected indicators of Social Exclusion in Estonian NAP[15]

- In-work poverty risk,
- Dispersion in regional employment rates,
- People living in jobless households in %,
- Population with low levels of education in %,
- Early school leavers not in education or training in %,
- Low reading literacy performance of pupils,
- Old age dependency ratio,
- Theoretical pension replacement ratios,
- Change in projected public pension expenditure,
- Unemployment trap in %,
- Inactivity trap in %,
- Low-wage trap, %,
- Net income of social assistance recipients as a % of the at-risk-of-poverty threshold,
- Infant mortality rate,
- Life expectancy,
- Healthy life expectancy,
- The proportion of the population covered by health insurance,
- Self-reported limitations in daily activities in %,
- Self-reported unmet need for medical examination in %,
- Self-reported unmet need for dental care in %,
- Acute care beds (per 100 000 inhabitants),
- Practicing physicians or doctors (per 100 000 inhabitants),
- Practicing nursing and midwives staff (per 100 000 inhabitants),
- Self-perceived health in %,
- Prevention measures: infant vaccination in %,

[15] Ministry of Social Affairs of Estonia (2006), *National Report on Strategies for Social Protection and Social Inclusion 2006-2008 under the Open Method of Coordination (for Estonia).*

- Total health expenditure per capita,
- Sources of health care financing in %,
- Social protection expenditure.

Selected SE indicators in Portuguese NAP[16]
Poverty and Deprivation (Source: ECHP, SILC, Eurostat):

- Risk of monetary poverty (total, female, male, children and elderly) (EU)[17],
- Risk of persistent monetary poverty (total, female, male) (EU),
- Risk of monetary poverty before social transfer, female, male) (EU),
- Risk of regional monetary poverty (NUTS II) (NAT),
- Risk of deprivation (total, family typology) (NAT),
- Risk of consistency Income (NAT),
- Weight of non-monetary income in families, total income,
- Rate of low salaried workers (NAT),
- Inequality in the distribution of income - S80/S20,
- Index de Gini (EU),
- Salary Disparity between genders (EU).

Social Expense:

- Total expense in social protection (EU),
- Total expense in pensions (EU).

Macro-economic:

- GDP per capita in PPC; GDP Real Growth Tax (EU),
- Work productivity by people occupied, work productivity of the work done an hour (EU).

[16] ANED (2008), *National Report on Strategies for Social Protection and Social Inclusion Portugal 2006-2008, A New Integrated Strategy on Social Policies*, University of Leeds.

[17] EU - commonly agreed EU indicator, NAT - commonly agreed national indicator

Employment/Unemployment:

- Employment rate 15-64 years (total, female, male) (EU),
- Employment rate of the 55 to 64 year old workers (total, female, male) (EU),
- Population in aggregate families unemployed 18-59 years old (EU),
- Long term unemployment (total, female, male) (EU),
- Regional Cohesion – dispersion of the regional employment rates NUTS II (total, female, male) (EU).

Education/Qualification:

- Percentage of workers working for others with low qualifications (EU),
- Percentage of the population 25-64 years old that participate in education and training throughout the four weeks before the enquiry (EU),
- Percentage of the population 25-64 years com with schooling below Secondary School (NAT),
- Percentage of the population 18-24 years com with schooling below Secondary School and does not study (EU).

Housing:

- Growth rate of Homes (NAT),
- Percentage of occupied homes by the owner with common residence in the total of occupied homes (NAT),
- Percentage of homes uninhabited (NAT),
- Percentage of homes that need great repairs (NAT),
- Percentage of private domestic aggregates that have the minimum housing (electricity, piped water, sewage, sanitary installations) (NAT).

Demography/Health:

- Average life expectancy at birth (NAT),
- Infant mortality rate (NAT),
- Synthetic index of fecundity (NAT),
- Fecundity rate of the 15-19 year old,

- Number of doctors for each 1000 inhabitants (NAT),
- Births following the location and aid (NAT),
- Percentage of Drug addicts in the ensemble of cases diagnosed with AIDS (NAT),
- AIDS incidence rate (NAT),
- Deaths related to drug use (NAT).

Social inclusion indicators in Italian NAP[18]
Economic Poverty:

- Risk of poverty rate,
- Persistence of risk of poverty,
- Severity of risk of poverty,
- Dispersion of risk of poverty,
- Risk of Poverty with fixed poverty threshold,
- Absolute risk of poverty,
- Risk of poverty before social transfers,
- Perceived poverty.

Mostly relative risks of poverty calculated on EHP data and national longitudinal data. Indicators are calculated on country level in different splits (by age, employment etc). Subjective measures of perceived poverty are also included.

Participation in Employment and Social Exclusion:

- General labour market conditions,
- Long-term unemployment,
- Jobless households,
- Risk of poverty of the employed,
- Risk of poverty and household participation in the labour market.

Basic labour market indicators calculated with LFS data and risk of poverty based on EHP data. Individual characteristics split.

[18] UNDP - Poland (2006).

Economic Inequality:

- Income inequality – Relationship between top and bottom quintiles,
- Income distribution inequality – Gini coefficient,
- Horizontal inequality.

Relative measures calculated with EHP data.

Living Conditions:

- Material hardship – Housing,
- Material hardship – Commodities,
- Living conditions – Area of residence,
- Access to services.

Housing and living conditions based on Household Survey data. Declared access to various services based on survey longitudinal results.

Education and Training:

- Young people with low educational attainment,
- Adults with low educational attainment,
- Early school-leavers,
- Lifelong training.

Measures related to labour market and schooling behaviour (drop-outs, early school leavers etc.) Ministry of Education and LFS data.

Poverty and Social Exclusion of Minors:

- Minors (under 18),
- Poverty of minors – risk (before and after social transfers), persistence and severity of poverty,
- Minors in jobless households,
- Hardship of minors.

Poverty of children and youth – calculated with EHP and HBS as well as LFS data.

State of Health and Social Exclusion:

- Life expectancy,
- Perceived state of health and economic condition,
- Multi-chronic persons,
- Disabled persons.

Subjective indicators form EHP and objective from national health statistics.

Social Participation:

- Socially isolated persons,
- Social, cultural and political participation,
- Social and family support networks.

Subjective measures from survey data (Multipurpose Survey of Households).

## 3.2. National systems of social indicators

In some EU members countries systems of social indicators are constructed by the research centres. Their propositions of indicators may be used for building the system of poverty and deprivation indicators at regional and local levels. The examples of these systems are the German System of Social Indicators (GESIS) and system of indicators used to construct Index of Multiple Deprivation for local government in the UK[19].

### The German System of Social Indicators

The GESIS is a project aimed at monitoring broadly conceived living conditions of German citizens (GESIS-ZUMA, 2007)[20]. It is conducted by the Centre for Survey Research and Methodology in Mannheim (Zentrum für Umfragen, Methoden und Analysen - ZUMA). The study has been launched in 1950 for Western Germany and, starting from 1991, captures the whole Federal Republic. All indicators are also displayed separately for western and eastern lands. Recently about 400 indicators are available. Out of this number, 89 indicators have been selected as core ones and are being published. They are taken from the following 14 life domains:

- Population,

[19] See also system of poverty and social exclusion indicators worked out for the United Kingdom by Joseph Rowntree Foundation (http://www.jrf.org.uk).

[20] See also Social Indicators Research Centre website: http://www.gesis.org/en/services/data/social-indicators/the-german-system-of-social-indicators/

- Socio-economic status and subjective class identification,
- The labour market and working conditions,
- Income and income distribution,
- Consumption and supply,
- Transportation,
- Housing,
- Health,
- Education,
- Participation,
- The environment,
- Public safety and crime,
- Leisure and media consumption,
- Global welfare measures.

Five of the above-mentioned domains interact with monetary and non-monetary poverty or deprivation, therefore they are presented below in more details by listing all sub-indicators.

1. The labour market and working conditions:

- Adjusted labour force participation rate,
- Part-time employment rate,
- Occupational qualification of persons in gainful employment,
- Share of gainfully employed persons working in the tertiary sector,
- Employees subject to social insurance contribution as a percentage of all,
- Gainfully employed persons,
- Unemployment rate,
- Positive subjective assessment of labour-market opportunities: employees,
- Rate of long-term unemployment,
- Average working week (according to collective agreements),
- Index of real wages,
- General job satisfaction.

2. Income and Income Distribution:

- Per Capita Net National Income in Constant Prices (in Euro),
- Ratio of Household Incomes in the Old and New Federal States,
- Concentration of Net Income (Gini index),
- Relative Poverty Rate (Poverty Line at 50% of Mean Income),
- Satisfaction With One's Household Income.

3. Consumption and Supply:

- Private Per Capita Consumption in Constant Prices,
- Costs of Welfare Expenditures,
- Rate of Savings of Private Households,
- Satisfaction With One's Own Standard of Living.

4. Housing:

- Residential Space per Person,
- Housing Without Standard Amenities,
- Average Rental Burden,
- Households Owning Their Own Housing,
- Satisfaction With One's Housing Conditions.

5. Health:

- Life Expectancy at Birth,
- Perinatal Mortality Rate per 1,000 Births,
- Persons With a Permanent Disability or Illness as a Percentage of the Total Population,
- Index of Subjective Evaluation of Personal Health,
- Number of Physicians per 100,000 Inhabitants,
- Health Care Expenditures as a Percentage of the GDP,
- Utilisation of Early Cancer Diagnosis Examinations,

- Daily Alcohol Consumption,
- Percentage of Smokers,
- Percentage of Overweight Persons.

The following data sources are being used: published data (statistical yearbooks, reports of the ministries), own welfare survey, German Socio-Economic Panel, Income and Consumption Survey, some other specialist surveys.

**The British Household Panel Survey (BHPS)**

The British Household Panel Survey (BHPS) has been started in 1991. BHPS is a multi-purpose annual panel study intended to capture changes in broadly conceived well-being of British citizens (and the UK citizens, since 2001, i. e. 11th wave)[21]. Starting from 2008, the BHPS has been incorporated into Understanding Society: the UK Household Longitudinal Study[22]. Being a component of the European Community Household Panel in the past, it is based on the same methodological principles. Except being a source of the data, the BHPS is also utilised to identify, model and forecast changes in well-being, their causes and consequences in relation to a range of socio-economic variables. Particularly, the BHPS provides information on household organisation, employment, accommodation, tenancy, income and wealth, housing, health, socio-economic values, residential mobility, marital and relationship history, social support, and individual and household demographics.

The BHPS is conducted by the ESRC UK Longitudinal Studies Centre, together with the Institute for Social and Economic Research at the University of Essex. The data are being collected by GfK NOP (a market research agency), Office for National Statistics and Northern Ireland Statistics and Research Agency. The main sponsor of the study is the Economic and Social Research Council.

The wave 1 consisted of approximately 5,500 households and 10,300 adult individuals drawn from 250 areas of Great Britain. Additional samples of 1,500 households from Scotland and Wales were added to the main sample in 1999. It permitted independent analyses of these two countries and facilitated analysis compared to England. In 2001 a sample of 2,000 households was added in Northern Ireland to increase the representativity of the whole of the UK. Therefore, the BHPS data are made up of five subsamples:

- Original BHPS from 1991 till recently comprising approximately 5,500 households

[21] BHPS Study Description:
http://www.esds.ac.uk/findingData/snDescription.asp?sn=5151
See also: The BHPS User Documentation and Questionnaires
http://www.iser.essex.ac.uk/survey/bhps/documentation
Understanding Society :
http://www.understandingsociety.org.uk/

[22] The UK-wide study taking a sample of 40,000 households, aimed at measuring various aspects of life.

- Former European Community Household Panel survey low-income sub-sample from 1997 to 2001 (Waves 7 to 11) comprising approximately 1,000 households
- Welsh extension from 1999 (Wave 9) comprising approximately 1500 households
- Scottish extension from 1999 (Wave 9); comprising approximately 1500 households
- Northern Ireland extension from 2001 (Wave 11) comprising approximately 1900 households

Altogether it makes approximately 10 000 HH. All members of the household aged 16 or over are interviewed. In addition children aged 11 – 15 complete a self-completion questionnaire – the Youth Questionnaire introduced in 1994. Recently data of 17 Waves (ending at the beginning of 2008) are available.

BHPS provides extensive support to the users. Three BHPS samplers have been added to the Teaching Datasets section of the ESDS Nesstar Catalogue[23]. The registered users have access to the BHPS datasets via the instant download service or can analyse, visualise, subset and download teaching datasets from BHPS via the online Nesstar software tool.

The core questionnaire covers a broad range of social science and policy interests including:

- household composition
- housing conditions
- residential mobility
- education and training
- health and the usage of health services
- labour market behaviour
- socio-economic values
- income from employment, benefits and pensions

There is also a variable component containing supplementary questions, asked less frequently than annually, new questions engendered by changing policy and research issues, and questions to elicit retrospective data on panel members' life histories before the first interview. These have included a lifetime history of marriage, cohabitation and fertility; lifetime job history; questions on wealth and assets, additional health measures, ageing, retirement and quality of life, children and parenting, neighbourhood and social networks.

[23] The Economic and Social Data Service (ESDS) is a national data service providing access and support for key economic and social data. ESDS provides an integrated service offering enhanced support for the secondary use of data across the research, learning and teaching communities.

### Measures of poverty and social exclusion in France for regional governments

A very good example of collecting and promoting regional data on poverty and social exclusion is provided by the local authorities in the region of ILE-de France. The Information Mission on Poverty and Social Exclusion (La Mission d'Information sur la Pauverte et l'Exclusion Sociale en Ile-de-France) issued 9th edition of the report covering both the results of the statistical surveys and the administration data on poverty and social exclusion in the region, completed by analytical and methodological comments. The publication (MIPES, 2008) is also accessible in the internet[24].The publication includes a wide scope of data, among others on:

- household incomes (Median, D1,D9, D9/D1),
- monetary poverty (poverty indicator by age, type of household),
- dwelling conditions (housing resources, unfavorable housing conditions causing a threat to human health, beneficiaries of various forms of assistance, eviction procedures)
- various forms of social assistance, among others granted to people gaining the minimum income ensured (RMI - Revenu Minimum d'Insertion),
- health (types of insurance, incidence of selected diseases such as tuberculosis),
- access to the labour market (unemployment duration, sociodemographic traits of job seekers, forms of vocational activation),
- education (pupils/students at different levels of education, including special education, learning difficulties),
- early social intervention (placement in hostels etc.).

Detailed statistics can be accessed on the Missions website:www.mipes.org.

### Measures of deprivation in the United Kingdom for local governments

Measures of poverty and deprivation at a small-area level in the United Kingdom have the longest tradition in Europe. They have been developed since early 1970s. The calculations of the appropriate indices for England and comparable indices for Northern Ireland, Wales, and Scotland have been supported by local authorities (Department of Social Policy..., 2003; Disadvantage Research Centre; 2007; Northern Ireland...,

[24] Among others at:
http://www.ile-de france.pref.gouv.fr/mipes/documents/Mipess_donnees_31_12_2007.pdf

2006; Statistical Directorate..., 2008)[25]. Those studies represent the highest advancement both in terms of scope of measures of deprivation and of disaggregation of national measures.

Multiple deprivation is defined as an aggregate measure of discrete dimensions or ‘domains’ of deprivation. Recently, similar indices of deprivation are published for all four constituent countries of the United Kingdom. They are estimated independently, however their construction is based on common principles. All country indices include the following domains:

1. Income deprivation,
2. Employment deprivation,
3. Health deprivation and disability,
4. Education, skills and training deprivation,
5. Access to services.

Moreover, the following domains are included into selected country studies:

6. Housing deprivation: England, Ireland, Wales,
7. Living environment deprivation: Ireland, Wales,
8. Crime: Ireland, England.

Usually, particular dimension of deprivation is not measured directly. This is true also for income deprivation. There is no income variable, therefore the proportion of monetary poor is measured through counts of people receiving various types of social support. When it is not possible to obtain number of the people at particular deprivation (as in the case of health and disability deprivation), factor analysis is employed to produce a single score. Below there are listed individual indicators for five domains that are common for all constituent countries and for housing which seems to be very important domain of deprivation. They are taken from the English study and indicators may vary in details between constituent countries.

1. Income deprivation:

- Adults and children in Income Support households,

[25] See also:

http://www.communities.gov.uk/publications/communities/indicesdeprivation07

http://www.nisra.gov.uk/aboutus/default.asp2.htm

http://www.scotland.gov.uk/Publications/2003/02/16377/18195

http://www.childreninwales.org.uk/policy/documents/statistics/10018.html

- Adults and children in Income Based Job Seekers Allowance households,
- Adults and children in Working Families Tax Credit households whose equivalent income (excluding housing benefits) is below 60% of median before housing costs,
- Adults and children in Disabled Person's Tax Credit households whose equivalent income (excluding housing benefits) is below 60% of median before housing costs,
- National Asylum Support Service supported asylum seekers in England in receipt of subsistence only and accommodation support

2. Employment deprivation:

- Unemployment claimant count of women aged 18-59 and men aged 18-64 averaged over 4 quarters,
- Incapacity Benefit claimants women aged 18-59 and men aged 18-64,
- Severe Disablement Allowance claimants women aged 18-59 and men aged 18-64,
- Participants in New Deal for the 18-24s who are not included in the claimant count,
- Participants in New Deal for 25+ who are not included in the claimant count,
- Participants in New Deal for Lone Parents aged 18 and over.

3. Health deprivation and disability:

- Years of Potential Life Lost,
- Comparative Illness and Disability Ratio,
- Measures of emergency admissions to hospital,
- Adults under 60 suffering from mood or anxiety disorders.

4. Education, skills and training deprivation:

- Average points score of children at Key Stage 2,
- Average points score of children at Key Stage 3,
- Average points score of children at Key Stage 4,
- Proportion of young people not staying on in school or school level education above 16,
- Proportion of those aged under 21 not entering Higher Education,
- Secondary school absence rate,

- Proportions of working age adults (aged 25-54) in the area with no or low qualifications.

5. Access to services:

- Road distance to GP premises,
- Road distance to a supermarket or convenience store,
- Road distance to a primary school,
- Road distance to a Post Office.

6. Housing deprivation:

- Household overcrowding,
- percentage of households for whom a decision on their application for assistance under the homeless provisions of housing legislation has been made,
- Difficulty of Access to owner-occupation.

Aggregation of sub-indices into single Multiple Deprivation Measure is achieved by means of the arbitrary weights (that may vary between countries, even for common domains). The highest weights are attached to income and employment. Moreover, the domains with the most statistically robust indicators receive greater weights.

Indices of multiple deprivation are calculated up to the levels that are equivalent to NUTS5. For England, Scotland and Northern Ireland the estimates are produced for wards[26], for Wales Lower Layer Super Output Areas (SOAs)[27]. The presented studies are based mainly on administrative registers and the census data. In some studies information from surveys (for instance, Northern Ireland House Conditions Survey or Scottish Local Labour Force Surveys) is occasionally used as a source of supplementary information. Therefore, generally survey data is not imputed into large datasets for construction of deprivation maps. However, some statistical problems arise in the case of some small wards, i. e. in calculation of rates when denominators are small and indicators are likely to be misestimated. In such a case ‘shrunken’ estimates of ward proportions are used. For a particular ward they are estimated as weighted combinations of data from that ward and data from other neighbouring or similar wards (e.g. all others in the same district). The ‘similar set’ of wards may

[26] Ward is an electoral district. As of 2004 there are 10,661 electoral wards (including Welsh and Wight electoral divisions) in the UK, with an average population of 5,500.

[27] There are 1,896 Lower Layer SOAs in Wales each having about 1,500 people. The advantages of the Lower Layer SOAs over the electoral divisions (which were used for the 2000 Index) is that they are more stable and they are roughly equal in size: there are big variations in the size of electoral divisions and regular boundary changes.

be defined by means of the national mean, the local authority district mean, the means of areas of similar characteristics or the mean of adjacent wards.

Except for mapping deprivation the indices can be used for:

- giving an overall deprivation score for each region,
- giving scores for separate deprivation domains for each region,
- comparing the deprivation scores for two or more region,
- ranking the scores for the regions.

The most recent poverty maps were produced for the years: 2003 (Scotland), 2005 (Northern England) 2007 (England) and 2008 (Wales). The estimates for England and Scotland were developed by the teams from Department of Social Policy and Social Work of the University of Oxford. Those for Northern Ireland and Wales were created by national statistical agencies (respectively, Northern Ireland Statistics and Research Agency and Statistical Directorate and the Local Government Data Unit of the Welsh Assembly Government).

## 3.3. Dimensions of poverty indicators

The departure point to construct system of poverty indicators is to identify the underlying dimensions (and sub-dimensions) of these phenomena and to group indicators accordingly. Particular indicators are treated as symptoms of poverty in the distinguished dimensions. Grouping indicators into these dimension may by carried out on the basis of heuristic methods (*brain storming, delphi method*) or statistical methods (*taxonomic methods, factor analysis*).

The Laeken indicators may be divided into four main dimension:

- Economic poverty,
- Labour market,
- Health,
- Education.

The indicators used for monitoring National Action Plans as well as systems of indicators employed in UE countries are usually also grouped into different poverty and social exclusion areas (see: Sections 3.1. and 3.2.).

Grouping of indicators into poverty dimensions was also employed by researchers of these phenomena. Whelan *at al.,* (2001) proposed to group indicators into five dimensions of non-monetary deprivation, through a factor analysis as follows:

1. Basic non-monetary deprivation – 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,
- Replacing any worn-out furniture,
- Buying new, rather than second hand clothes,
- Eating meat chicken or fish every second day, if the household wanted to,
- Having friends or family for a drink or meal at least once a month,
- Inability to meet payment of scheduled mortgage payments, utility bills or hire purchase instalments.

2. Secondary non-monetary deprivation – 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 video recorder,
- A micro wave,
- A dishwasher.
- A telephone.

3. Lacking housing facilities – 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,
- Hot running water.

4. Housing deterioration – these concern serious problems with accommodation:

- Leaky roof.
- Damp walls, floors, foundation etc.,
- Rot in window frames or floors.

5. Environmental problems – these concern problems with the neighbourhood and the environment:

- Shortage of space,
- Noise from neighbours or outside,
- Dwelling too dark/not enough light.
- Pollution, grime or other environmental problems caused by traffic or industry,
- Vandalism or crime in the area.

The indicators within each dimension can be aggregated into group indexes and finally into a single synthetic index of the non-monetary deprivation.

Other example of identifying the dimensions of poverty and social exclusion was presented by Czapiński and Panek (UNDP-Poland, 2006). Poverty and social exclusion has been measured using two approaches: objective and subjective. Indicators of poverty and social exclusion in the objective approach were divided into two main dimensions, i.e. exclusion from the labour market and exclusion from the goods and services market. Within particular dimensions sub-dimensions were distinguished. Sub-dimensions of exclusion from the labour market were created mainly with regards to groups of excluded persons, and in the area of exclusion from goods and services market with regards to types of goods and services.

The following sub-dimensions and indicators were selected:

1. Exclusion from the labour market

1.1. Unemployment:

- Unemployment rate,
- Long-term unemployment rate,
- Long-term unemployment intensity,
- Very long-term unemployment rate,
- Intensity of persons living-in non-working households,
- Intensity of persons with short tenure,
- Flow from unemployment to employment,
- Flow from unemployment to economic inactivity,
- Intensity of part-time employment.

1.2. Occupational inactivity:

- Occupational activity.

1.3. Exclusion due to discrimination:

- Intensity of unemployment among single mothers,
- Intensity of women seeking work.

1.4. Exclusion due to low-level education or lack of professional experience:

- Intensity of non-workers among new alumni,
- Intensity of non-workers among older alumni,
- Intensity of non-workers with low education level,
- Intensity of unemployment among persons with low education level,
- Flow from unemployment to employment of persons with low educational attainment,
- Flow from unemployment to inactivity of persons with low educational attainment.

1.5. Exclusion due to disability:

- Intensity of unemployment among the disabled.

2. Exclusion from the goods and services market-partial indicators
2.1. Financial poverty:

- Relative poverty rate,
- Absolute poverty rate,
- Relative poverty gap,
- Absolute poverty gap.

2.2. Material poverty (concerns the lack of ability to afford due to financial reasons):

- Intensity of lack of refrigerator,
- Intensity of lack of cooker,
- Intensity of lack of automatic washing machine.

2.3. Deficit in apartment equipment:

- Intensity of lack of WC facilities,
- Intensity of lack of bathroom,
- Intensity of lack of running water,
- Intensity of lack of central heating.

2.4. Deficit in access to health services (due to financial reasons):

- Intensity of resignation from dental treatment,
- Intensity of resignation from medical visits,
- Intensity of resignation from medical examinations,
- Intensity of resignation from rehabilitation treatments.

2.5. Deficit in access to leisure and cultural services (due to financial reasons):

- Intensity of resignation from travelling among adults,
- Intensity of resignation from travelling among children,
- Intensity of resignation from theatre, opera, operetta,
- Intensity of resignation from buying a book,
- Intensity of resignation from buying press (newspapers, magazines).

2.6. Deficit in access to communication and social communication services:

- Intensity of households lacking a telephone land line,
- Intensity of households lacking a mobile phone,
- Intensity of households lacking a computer,
- Intensity of households lacking an Internet connection,
- Intensity of households lacking a car.

In the subjective approach the following three main dimensions and indicators were identified:
1. Subjective (perceived) material exclusion:

- Assessment of the wealth-related living standard,
- Satisfaction with the financial status of the family,
- Satisfaction with the current family income,
- Satisfaction with the housing conditions,
- Satisfaction with the level of accessible goods and services.

2. Subjective (perceived) social exclusion:

- Number of friends,
- Feeling of being loved and trusted,
- Feeling of loneliness,
- Feeling of being discriminated for any reason.

3. Psychological ill-being:

- Whole life assessment,
- Feeling of happiness,
- Suicidal tendencies,
- Will to live,
- Mental depression.

In both approaches indicators (symptoms of exclusion) were later aggregated in sub-groups, groups and in synthetic indicators.

## 3.4. Strategy for defining regional and local indicators

According to *Nomenclature of Territorial Units for Statistics,* worked out by Eurostat, regional level covers NUTS1, NUTS2 and NUTS3 (small, medium and large territorial units), while local level (subregional) concerns NUTS4 and NUTS5 (Eurostat, 2004a). The definition and choice of appropriate units to serve as 'regions' an 'subregions' (local units) for the construction of poverty and related indicators is a fundamental issue which should be considered in our context.

We can consider three classes of relevant units:

- Units based on administrative or political criteria, specifically NUTS regions,
- Units defined in terms of the urban-rural classification. The classification often has to be more elaborate than a simple dichotomy28,
- Geographical units based on or defined according to some functional criteria29.

[28] For example, in Poland three types of NUTS 5 (gminas) are distinguished: urban, rural and urban-rural.

The other fundamental aspect in the definition and building of indicators at regional and local levels is related to the available data. The construction of indicators at these levels is necessarily a compromise between the theoretical definition and the empirically possible. The strategy recommended for the construction of such indicators has four fundamental aspects:

- making the best use of available sample survey data at national level, which usually do not readily provide accurate regional level estimates, but contain detailed income and other poverty related data which may be useful for regional and local estimates,
- exploiting to the maximum existing regional ('meso') data sources, which cover main economic and social variables – for the purpose of constructing regional indicators and for estimation of indicator at local level,
- using information collected at local level, which comes from local databases, registers ('administrative data') and sample surveys at local level,
- and using all the mentioned sources of data in combination to produce more precise estimates for regions and local units using appropriate small area estimation (SAE) techniques.

**Choice of units**

The first issue in developing regional and local indicators concerns the choice of the type of units to serve as "regions" and "subregions". For a number of substantive and practical reasons, we consider geographical-administrative regions and administrative local units, specifically NUTS regions and subregions (and LAUs) at various level of classification, as the most appropriate choice for EU countries. The reasons for this choice include the following. NUTS regions and subregions are the most commonly used units for the formulation and implementation of social policy: the units are well-defined and identifiable, and are already widely accepted and used by different users and producers of statistical information. Despite the fact that NUTS units are not defined in exactly the same way in different countries and can differ greatly in size and homogeneity, this territorial system of classification provides a common framework which enhances comparability of the resulting statistical information. Inter-country, EU-wide research also benefits from the use of units based on the same system of classification. The classification covers each country exhaustively, providing a hierarchical set of units for which data can be linked across different levels. A lot of information already exists for this type of units from many different sources. Above all, data availability for the purpose of constructing the required indicators is the major reason for the choice of NUTS regions and subregions for the purpose.

This by no means precludes the above being supplemented by other dimensions. For instance, it is possible to consider 'functional regions and local units', such as regions and subregions defined in terms of the labour market, production, trade or other economic indicators, or in terms of density and other characteristics of the population distribution (e.g., urban-rural distinction). Alternatively, may be

[29] Examples are Labour Market Regions - such as Sistema Economico Locale (*SEL)* in Tuscany, which are largely but not entirely confined to be within Provinces (NUTS3 regions), but may not take account of administrative divisions below that level.

disaggregated according to population subgroups, i.e., groups identified by characteristics of individual households and persons: children, elderly persons, national minorities, immigrants, etc. Indeed, the analysis can refer to different types of indicator disaggregation simultaneously. For instance, NUTS at a sufficiently low level can be classified according to whether their character is primarily urban or primarily rural. In fact, indicators can be constructed for geographical-administrative units precisely for the purpose of such classification. Furthermore, NUTS-based indicators can be enriched by subpopulation analysis to the extent the available data permit their further disaggregation.

## The available data

Poverty indicators may be derived from diverse data sources. Data may come from sample surveys at national level, meso databases at regional level, local databases, registers and sample surveys conducted at regional and local levels.

### *Sample surveys*

The complexity of the information on which indicators of poverty and deprivation are based (such as detailed income distribution of households and persons in the population) causes that most of them have to be obtained from intensive surveys such as EU-SILC, LFS, HBS, LIS or ECHP.

*The EU Statistics on Income and Living Conditions (EU-SILC)*

The EU-SILC replaced the European Community Households Panel (ECHP) carried out over 1994-2001 (Guio, 2005). EU-SILC organisation and methodology is governed by the European Parliament's and European Council's regulation No. 1177/2003 of June 16, 2003 (with amendments included in regulation No. 1553/2005) concerning Community Statistics on Income and Living Conditions (EU-SILC) along with regulations of the European Commission corresponding to that legal act.

The EU-SILC was launched under a gentleman's agreement with six EU-15 countries plus Norway in 2003 and re-launched under a Regulation with twelve EU-15 countries (Belgium, Denmark, Greece, Spain, France, Ireland, Italy, Luxembourg, Austria, Portugal, Finland and Sweden) and in Estonia, Norway and Iceland in 2004. In 2005 the rest of the EU-25 countries joined the EU-SILC. Bulgaria, Romania, Turkey and Switzerland have launched SILC in 2006.

The set of mandatory EU-SILC variables covers the basic information on the demographic traits respondents, their involvement in the education process, the evaluation of health status, selected data on deprivation of basic necessities, data on housing conditions, detailed information on activity in economic life, and above all, an extensive range of information on the level and sources of income (Eurostat, 2007)[30]. Information on social exclusion and housing conditions are obtained for households. Education and health data are collected for persons aged 16 and over.

The EU-SILC survey provides data for monitoring poverty and social exclusion, collected in a uniform and standarized manner, at national and regional levels, namely:

[30] Additional information on EU-SILC can be found at: http://forum.europa.eu.int/Public/irc/dsis/eusilc/library

- cross-secional data pertaining to a given time or a certain time period,
- longitudinal data pertaining to individual level changes over time, observed over a four year period.

It allows to combine of data from local databases and registers with information gained from the individuals and households levels, obtained from representative panel household surveys.

As a rule Eurostat presents EU-SILC results both at the EU level and at the level of individual countries (available on-line on Eurostat website). For the scientific purposes it is possible to use the anonymous individual data files.

The generally accessible EU_SILC Eurostat's database contains among others:

- Predefined tables (relate to: inequality of income distribution, at-risk-of-poverty rates),
- Multi-dimensional tables (relate to: income distribution, monetary poverty, non-monetary poverty and social exclusion).

The main indicators are monetary indicators used in the context of the Open Method of Coordination (OMC) on social inclusion and social protection. They are divided into: overarching indicators, social inclusion indicators, pensions indicators. Of these OMC indicators included as multidimensional tables on the Eurostat website, some are extremely visible as they are used for the social cohesion domain of the Structural Indicators while others are Sustainable Development Indicators.

In recognition of the need to increase the analysis of households' material condition differentiation in EU and having in mind the multidimensional nature of poverty, Eurostat, together with the Indicators Sub-Group (ISG), work on compiling the commonly agreed list of material deprivation indicators. On the basis of the EU-SILC 2006 results the list 'of candidates' indicators' was suggested, covering the three main domains: economic strain, durables and housing (European Commission, 2008c and 2008f):

1. Economic strain:

- Afford paying for a week's annual holiday away from home,
- Afford to keep home adequately warm,
- To pay as scheduled rent or mortgage,
- To pay as schedules utility bills,
- To pays as schedule hire purchase instalments,
- Afford a meal with meat, chicken or fish every second day if wanted,
- Capacity to face unexpected expenses.

2. Durables:

- Enforced lack of a colour T,
- Enforced lack of a telephone,
- Enforced lack of a car or van for private use,

- Enforced lack of washing machine.

3. Housing:

- Bath or shower,
- Indoor flushing toilet,
- Accomodation too dark,
- Leaky roof/rot in window frames or floors/ damp walls, floors, foundations.
- EU-SILC determines also to carry out module studies focused on issues of special interest to European Union. The following module surveys were conducted EU-SILC in the previous years:
- 2005 - Intergenerational transmission of poverty (Regulation (EC) N° 16/2004),
- 2006 - Social participation (Regulation (EC) N° 13/2005),
- 2007 - Housing conditions (Regulation (EC) N° 315/2006),
- Over-indebtedness and financial exclusion (Regulation (EC) N° 215/2007),
- 2009 - Material deprivation (Regulation (EC) N° 362/2008).

Particularly useful to find indicators to extend the common set of Laeken non-monetary indicators, has module devoted to material deprivation which will be included in EU-SILC survey in 2009. This is in relation to the ISG intensions to develop indicators in the 3 priority domains:' economic strain and durables', 'housing and environment' and 'child deprivation'.

The list of variables on material deprivation is as follows (European Commission, 2008c and Council Regulation (EC), 2008):

1. Household items asked at household level

1.1. Housing items:

- Place to live with hot running water,
- Expectation of household to change dwelling,
- Main reason for the expectation to change dwelling,
- Shortage of space in dwelling,
- Size of dwelling in square meters (optional).

1.2. Environment items:

- Littering around in the neighbourhood,
- Damaged public amenities (bus stop, lamp posts, pavements, etc.) in the neighbourhood,
- Accessibility of public transport,

- Accessibility of postal or banking services.

1.3. Financial stress:

- Replacing worn out furniture.

1.4. Durables:

- Internet connection.

2. Items asked at individual level

2.1. Durables:

- Mobile phone.

2.2. Basic needs:

- Replace worn out clothes by some new (not second –hand) ones,
- Two pairs of properly –fitting shoes (including a pair of all-weather shoes).

2.3. Unmet needs:

- Number of visits to GP's and specialists, excluding dentists and ophthalmologists.

2.4. Leisure and social activities:

- Get together with friends/family(relatives) for a drink/meal at least one a month,
- Regulary participate in a leisure activity such as sport, cinema, concert,
- Spend a small amount of money each week on yourself.

3. Children items asked at household level

3.1. Basic needs:

- Some new (not second –hand) clothes,
- Two pairs of properly –fitting shoes (including a pair of all-weather shoes),
- Fresh fruit and vegetables once day,
- Three meals a day,
- One meal with meat, chicken or fish (or vegetarian equivalent) at least once day.

3.2. Educational or leisure needs:

- Books at home suitable for their age,

- Outdoor leisure equipment (bicycle, rollerskates, etc.),
- Indoor games (educational baby toys, building blockes, board games, computer games, etc.),
- Regular leisure activity (swimming, playing an instrument, youth organisations, etc.),
- Celebrations on special occasions (birthday, name day, religious events, etc.),
- Invite friends round to play and eat from time to time,
- Participate in school trips and school events that cost money,
- Suitable place to study or do homework,
- Outdoor space in the neighbourhood where children can play safely,
- Go on holiday away from home at least 1 week per year.

3.3. Medical needs:

- Unmet need for consulting a GP or specialist, excluding dentists and ophthalmologists – optional,
- Main reasons for unmet need for consulting a GP or specialist, excluding dentists and ophthalmologists – optional,
- Unmet need for consulting dentist-optional,
- Main reason for unmet need for consulting dentist-optional.

*The Household Budget Survey (HBS)*

Household Budget Surveys (HBSs) are national surveys mainly focusing on consumption expenditure (European Commission, 2003a). They are conducted in all EU Member States and their primary aim (especially at national level) is to calculate weights for the Consumer Price Index. However it may also be used for many other purposes either at national or European level (economic studies, social analyses, market research), among others for monitoring poverty (Eurostat, 2004c).

It should be noted that before the implementation of EU-SILC the basic source of data used by Eurostat for measuring income and poverty indicators in the new member states or candidate countries was provided by HBS (Eurostat, 2004d and 2005)[31].

The HBS project is run under a Gentlemen's agreement. Every few years Eurostat collects (recently for 2005 and previously for 1999) EU comparable household budget survey results.

The generally accessible HBS Eurostat's database contains among others:

- Predefined tables (relate to consumption expenditure of private households),

[31] It should be remembered that by the time ECHP was introduced, it was the household budget survey that provided the basis for the Eurostat's analysis of material poverty. The objective poverty range was estimated on the basis of the expenditure. See: Hagenaars *at al.*, 1992 and European Commission, 1990.

- Multi-dimensional tables (relate to mean consumption expenditure, structure of consumption expenditure, household characteristics).

In terms of publications for the HBS field, *Statistics in Focus* (SiFs), which can be downloaded from the Eurostat website, as well as detailed methodological documents available on CIRCA are produced (Eurostat).

*The European Union Labour Force Survey (EU LFS)*

The European Union Labour Force Survey (EU LFS) is the EU's harmonised survey on labour market developments (European Commission, 2003c and 2007[32]). It is a rotating random sample survey of persons in private households. The sampling units are dwellings, households or individuals depending on the sampling frame.

The LFS provides population estimates for the main labour market characteristics, such as employment, unemployment, inactivity, hours of work, occupation, economic activity and much else as well as important socio-demographic characteristics, such as sex, age, education, households and regions of residence.

[32] See also: http://circa.europa.eu/irc/dsis/employment/info/data/eulfs/LFS_MAIN/LFSuserguide/EULFS_database

**Table 3.4.1**. Modules of LFS.

| Year | Topic | Commission regulation |
|---|---|---|
| 1999 | Accidents at work and occupational diseases | (EC) No 1571/1998 |
| 2000 | Transition from school to working life | (EC) No 1925/1999 |
| 2001 | Length and patterns of working time | (EC) No 1578/2000 |
| 2002 | Employment of disabled people | (EC) No 1566/2001 |
| 2003 | Lifelong learning | (EC) No 1313/2002 |
| 2004 | Work organisation and working time arrangements | (EC) No 247/2003 |
| 2005 | Reconciliation between work and family life | (EC) No 29/2004 |
| 2006 | Transition from work into retirement | (EC) No 388/2005 |
| 2007 | Accidents at work and work-related health problems | (EC) No 341/2006 |
| 2008 | Labour market situation of migrants and their immediate descendants | (EC) No 102/2007 |
| 2009 | Entry of young people into the labour market | (EC) No 207/2008 |

The division of the population into employed persons, unemployed persons and inactive persons follows the International Labour Organisation definition. Other concepts also follow as close as possible the recommendations of ILO.

The survey is based on European legislation. The principal legal act is the Council Regulation (EC) No. 577/98. The implementation rules are specified in the successive Commission regulations.

Since 1999 an inherent part of the European Union labour force survey (LFS) are the so called 'ad hoc modules', which are presented in the table below.

The national statistical institutes are responsible for selecting the sample, preparing the questionnaires, conducting the direct interviews among households, and forwarding the results to Eurostat in accordance with the common coding scheme.

The Eurostat publications related to the EU LFS ( available on-line on Eurostat website) are organised in five series:

- Methods and definitions,
- Characteristics of the national surveys,
- Quality reports,
- Statistics in Focus and data in Focus (presenting the main results for variety of topics),
- Detailed results.

As a rule Eurostat presents LFS results both at the EU level and at the level of individual countries. However, the basic information on the economic activity of the population are also accessible at the regional level (available on-line on Eurostat website, theme: *General and regional statistics*). For the scientific purposes it is possible to use the anonymous individual data files.

The generally accessible Eurostat's database contains among others a set of LFS main indicators on employment and social cohesion divided into following groups:

- Population, activity and inactivity indicators (including also the structural indicators Average exit age and Population in jobless households),
- Employment (including main characteristics, employment rates, employment growth and activity branches),
- Unemployment (including also Harmonised long-term unemployment),
- Education and Training (including the structural indicators: Lifelong Learning, Education Attainment Level and Early School Leavers).

Additionally, the EU LFS results are used in the series of analyses published by the European Commission. The EU LFS is one of the main sources used in the Annual Employment Report. This publication provides analytical and statistical background to the European Employment Strategy.

The EU Labour Force Survey has been recognised as the data source for the construction of all the employment- related commonly agreed indicators. The so called overarching and the context indicators based on LFS are grouped into dimensions as follows (European Commission, 2008a and 2008g).

1. The overarching indicators:

1.1. Educational outcome and human capital formation:

- Early school leavers (by gender).

1.2. Access to labour market:

- People living in jobless households (by age:0-17,18-59, by gender – 18+only).

1.3. Employment of older workers:

- Employment rate of older workers (by age;55-59;60-64' by gender), possibly replaced or supplemented by „average exit age from the labour market".

1.4. Participation in labour market:

- Activity rate (by gender and age: 15-24,25-54,55-59,60-64,total).

1.5. Regional cohesion:

- Regional disparities – coefficient of variation of employment rates.

2. The context indicators:

- Employment rate ( by gender),
- Unemployment rate (by key age groups and by gender),
- Long term unemployment rate (by key age groups and by gender),
- Jobless households by main household types.

*Luxembourg Income Study*

The Luxembourg Income Study (LIS) is a cooperative research project that covers more than 30 member countries[33]. The LIS database includes household microdata, usually at multiple points in time. The surveys are based on probability samples collected in member countries. They are intended to capture all main types of households, though the coverage of the population varies between countries and years. The contents of the surveys are generally consistent with standard household surveys like European Community Household Panel or Polish Household Budget Survey with two important exceptions: there are no panel datasets and information on consumer expenditures is available for selected countries only. All country databases include information on various types of income, labour market and demographic data as well. The datasets are being

[33] See also: Luxembourg Income Study website: http://www.lisproject.org
http://www.lisproject.org/publications/liswps/303.pdf

collected by the national statistical agencies in the LIS member countries and then all the variables are harmonised by the LIS staff. Therefore they can be directly compared across countries.

The data can be accessed for research purposes via the internet mailing system by submitting SAS, SPSS or STATA programs. No fee is charged for researches affiliated with LIS member countries. The LIS database captures most of developed countries as well as several transition and developing countries, in this number Italy, Poland, Spain and UK. The project was launched in 1983 and the first data are of 1967 (for Sweden). The most recent datasets are of 2005 (Sweden and Taiwan). For most of the countries the data are available for more than one year.

The LIS database is intended to ensure free access to harmonised and standardises micro-data from the different national surveys in order to facilitate comparative research on individual incomes and related variables. Another purpose is dissemination of scientific results capturing a broad issue of well-being. The LIS working papers written by the researchers using the LIS data are available from the website.

The LIS is the joint project of the government of the Grand Duchy of Luxembourg and the Centre for Population, Poverty and Policy Studies (CEPS).

Main modules in LIS variables capture:

1. Income variables:

- gross and net household income,
- labour income,
- self-employed income,
- pensions,
- social transfers,
- private transfers,
- several other types of income.

2. Labour market and human capital variables:

- employment status,
- occupation,
- education,
- occupational training.

3. Demographic variables and other household attributes.

4. Expenditures (nor available for all countries).

LIS focuses, by definition, on observing individual incomes, therefore only income poverty/inequality indicators may by calculated for all countries. Social exclusion may be observed through the unemployment measures only. Breakdown by age, labour status, household tenure status and gender is possible for most of national surveys. Most of datasets have a geographic location indicator. The lowest possible level of NUTS depends on the sample sizes (that differ significantly between countries) as well as on availability of appropriate regional identifiers. For virtually all countries (in this number, all EU countries) averages may be calculated by NUTS1 region and for several of them by NUTS2 region (see Stewart, 2002 for an empirical study on several Eu-15 countries).

For all countries and available years selected aggregate measures of well-being are published on the LIS website. Except median equivalent income they include:

1. Inequality indices:

- Gini Coefficient,
- Atkinson Coefficient (epsilon=0.5),
- Atkinson Coefficient (epsilon=1),
- Percentile Ratio (90/10),
- Percentile Ratio (90/50),
- Percentile Ratio (80/20).

2. Relative poverty indices:

- Relative Poverty Rates - Total Population (40%),
- Relative Poverty Rates - Total Population (50%),
- Relative Poverty Rates - Total Population (60%),
- Relative Poverty Rates - Children (40%),
- Relative Poverty Rates - Children (50%),
- Relative Poverty Rates - Children (60%),
- Relative Poverty Rates - Elderly (40%),
- Relative Poverty Rates - Elderly (50%),
- Relative Poverty Rates - Elderly (60%).

3. Information on children:

- Distribution of Children by income group (50-75%),
- Distribution of Children by income group (75-150%),
- Distribution of Children by income group (above 150%),
- Children Poverty Rates - Two Parents Family (50%),
- Children Poverty Rates - Single Mother Family (50%),
- % Children living in Single Mother Family.

### *Employment of sample survey for poverty estimation beyond national level*

The major problem in the production of indicators for regions or other small domains is the smallness of sample size available in such surveys. Generally, adequate sampling precision is available at the national level, as demonstrated by the extensive use of intensive surveys data for the purpose. The same may apply to indicators at NUTS1 level in some cases, but generally sampling errors may be too large for the results to be useful even at that level. For instance, in a study on measures of well-being and exclusion in Europe's NUTS1 regions using LIS data (Stewart, 2002, 2003), it is clear that the large sampling errors involved often make it difficult to draw clear conclusion in many instances. This is also true in a study on patterns of poverty across European regions (Berthoud, 2004), to a lesser extent only because the results are aggregated over small NUTS1 regions where sample size problems are most critical. As a final example at the EU level, a study on the impact of relative poverty lines (Kangas and Ritakallio, 2004) goes below the national level only to "mega regions", meaning the division of each country into at most two – the richer and the poorer – parts. There are a few studies of regional variations for individual countries, but mostly confined to countries with large samples. For instance, Rodrigues (1999) reports estimates of mean income (per 'equivalent adult') and head-count ratios (poverty rates) by NUTS2 regions in Portugal. It has been possible to do so at this level of detail because Portugal had an exceptionally large sample size in the ECHP, compared to other countries in the project. Chakravarty and D'Ambrosio (2003) report similar results for Italy, but this time only at NUTS1 level. At a finer level of detail such as NUTS2 there are interesting results in Verma *et al*. (2005). This is despite the fact that among the countries participating in the ECHP, Italy had the largest sample size. In UNDP Poland (2006) report a large part of labour market and consumption exclusion indicators were calculated for Polish regions (NUTS2) on the basis of data obtained through LFS and HBS surveys carried out in Poland. These surveys base on large sample size, but cannot been employed directly for indicator estimation beyond NUTS2.

Our requirement is considerably more demanding than the above: ideally to be able to produce useful regional indicators for NUTS2 regions (or other geographical subpopulations of similar size) in most cases, and even to NUTS3 level in some cases.

The problem of sample size requires a more sophisticated statistical approach than simply using direct estimates from single rounds of sample surveys of moderate size. In order to overcome that problem, the first aspects to consider is of 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 (see: Part 2).

### *Meso data*

In order to construct regional indicators is fundamental exploiting to the maximum 'meso' data (such as highly disaggegated tabulation data). If the target of the research is the EU context, the NewCronos (now termed *Eurostat Free Dissemination Database*) provides a valuable data resource for the construction of regional indicators. In itself it is not a source of original data, but a compilation of information from a diversity of sources presented in the form of very detailed tabulations. NewCronos REGIO domain covers the principal aspects of the economic and social life of the European Union: demography, economic accounts, labour force, health, education, etc., by region. The concepts and definitions used are as close as possible to those used by Eurostat for the production or compilation of statistics at national level. The standard model for compiling regional aggregates at various levels has been as follows: first, data from various national sources are compiled in the National Statistical Offices, and then provided to Eurostat for validation. This data set is then loaded into NewCronos by the thematic unit in question.

We believe that this resource, NewCronos, has hitherto been under-utilised, and that there is a great potential for more thorough exploitation of the information which already exists. While direct indicators of regional poverty and living conditions are generally not available with sufficient regional breakdown in NewCronos, several exceptionally positive aspects of the resource need to be appreciated. Some of these become even more important as we move down from the national to the regional level.

There are three main forms in which variables derived from NewCronos can be utilised for the construction of poverty and deprivation indicators, though the first two of them concern only regional level (Verma *at al.*, 2006):

1. *Direct deprivation indicators*: some statistics in NewCronos can serve, in their own right, as direct indicators pertaining to poverty and living conditions. In fact, the scope for such use is likely to be greater in the context of regional indicators, compared to that in the national context. This is because measures of levels – which are more abundantly available in NewCronos than the generally more complex distributional measures - can themselves serve as indicators of disparity when compared across regions.
2. *Intermediate output indicators: NewCronos provides a very large number of measures, giving what has been termed as "intermediate output" indicators. Such indicators express on the one hand the policy effort in favour of those at risk of poverty and social exclusion, and on the other hand the impact of social policies as well as of the economic context. NewCronos is a unique source of such indicators.*
3. *Predictors: a large number of measures correlated with direct indicators of poverty and deprivation can be constructed. In conjunction with direct indicators obtained from more intensive surveys, these measures can be used as "covariates" or "regressors" to produce more precise indicators using small area estimation (SAE) procedures described in the Section 5.*

### *Sources of data at local level*

Administration data (regional or local databases and registers) can serve as source of information to calculate directly some indicators referring to poverty at local level (NUTS4 and NUTS5), first of all non-monetary indicators. Raw data and indicators obtained from sample surveys carried out only in some administrative units (NUTS4 and NUTS5) may be used for estimation of indicators for other administrative units of similar type (with similar attributes). It of course requires the development of small estimates that take into account spatial correlation between administrative units with similar characteristics.

## 3.5. Specific methods to estimate poverty and inequality measures at local level

The problem of sample size requires a more sophisticated statistical approach[34] than simply using direct estimates from a sample surveys of moderate size. We have considered the issue regarding the data, now we concentrate the attention on estimation methods required to overcome the inefficiency due to the small sample size.

### Small Area Estimation (SAE)

#### *Introduction*

It was assumed that the system of poverty indicators should also cover the local level, i.e. that of municipalities, regions, districts, where social services are delivered and social needs are more directly perceived. There is an inherent tension between costs, the statistical significance of sample surveys, the longitudinal aspects and data disaggregation. An appropriate mix of registers, administrative data and sample surveys, and the application of small area sampling, should permit to find the most appropriate balance in the trade-offs. In any case, the local, regional, national and international systems should be fully consistent among one another, by applying the relevant statistical standards. This should also promote an integrated approach to monitoring. However, geographically-based domains, like regions, states, counties, wards and metropolitan areas are typically of most interest. A traditional approach to estimation for such domains is based on application of classical design-based survey sampling methods. Estimates based on this approach are often called direct estimates in the literature. However, sample sizes are typically small or even zero within the domains/areas of interest. This results in the direct estimators having large variances. When there are no sample observations in some of the small areas of interest, direct estimators cannot even be calculated. Small area estimation theory is concerned with resolving these problems.

Often there is auxiliary information that can be used to define estimators for small areas. In some cases these are values of the variable of interest in other, similar, areas, or past values of this variable in the same area or values of other variables that are related to the variable of interest. Estimation and inference approaches based on using this auxiliary information are called indirect or model-based. Methods based on the use of auxiliary information have been characterized in the statistical literature as ″borrowing strength" from the relationship between the values of the response variables and the auxiliary information. Model-based methods have long history, but have only received attention in the last few decades as defining an approach to estimating small area characteristics. In this context, two main ideas have been used in developing models for use in small area estimation. These either assume that the inter-domain variability in the response variable can be explained entirely in terms of corresponding variability in the auxiliary information, leading to so-called fixed effect models, or require the assumption that "unexplained" domain specific variability remains even after accounting for the auxiliary information, leading to mixed models incorporating domain specific random effects.

[34] Citro and Kalton (2000) wrote:"...none of the existing or planned surveys or administrative records sources can, by itself, provide direct estimates of sufficient reliability, timeliness, and quality of responses for all of the SAIPE income and poverty estimates. Therefore, the panel concludes that the SAIPE program must to rely primarily on models that combine data from more than one source to produce indirect estimates".

Fixed effect models explain inter-domain variation in the response variable of interest entirely in terms of variation in known factors. Estimates of small area characteristics based on fixed effect models are referred to as synthetic estimators, composite estimators, and prediction estimators. Mixed models also have a long history, but have received special interest only in the last few decades. This is partly due to the heavy computational burden of estimation methods used with such models. Recent developments in computing hardware, software and estimation methods have however led to increased attention being paid to the use of mixed models for data analysis. Linear mixed models have a wide range of applications. In particular, the ability to predict linear combination of fixed and random effects is one the more attractive properties of such models. In a series of papers, Henderson (1948 - 1975) developed the best linear unbiased prediction (BLUP) method for mixed models. In this case "best" stands for minimum mean square error among all linear unbiased predictors, "linear" means that the predictor is a linear combination of the response variable values and "unbiased" means that expected value of the prediction error (predicted value of variable - actual value of variable) is zero. The BLUP method has become a powerful and widely used procedure for fitting models for genetic trends in animal populations based on traits measured on both the continuous and the categorical scale. However, the BLUP methods described in Henderson assumed that the variances associated with random effects in the mixed model (the variance components) are known. In practice of course such variance components are unknown and have to be estimated from the data. There are several methods for estimating variance components. Harville (1977) reviews these methods, including maximum likelihood and residual maximum likelihood and three other methods suggested by Henderson. The predictor obtained from the BLUP when unknown variance components are replaced by associated estimators is called the empirical best linear unbiased predictor (EBLUP) and is described in Harville (1990), Robinson (1990) and Harville (1991). Over the last decade several important approaches have been developed for estimating/predicting the value of a linear combination of fixed and random effects in discrete response data. In virtually all of these random effects are assumed to be normally distributed. Several authors have extended the empirical best linear predictor (EBLUP) to generalized linear models.

Mixed models have been used to improve estimation of small area characteristics of small area based on survey sampling or census data by Fay and Herriot (1979), Ghosh and Rao (1994), Rao (1999) and Pfeffermann (1999). In these applications, the mixed model derives from the concept that the vector of finite population values is a realisation of a super-population. In this context, estimation of a small area mean is equivalent to prediction of the realization of the unobservable random area effect in a linear mixed model for the super-population distribution of the variable defining this mean. In addition to EBLUP, empirical Bayes (EB) and hierarchical Bayes (HB) estimation and inference methods have been also applied to small area estimation. Under the EB approach, Bayes estimation and inferential approaches are used in which posterior distributions are estimated from data. Under the HB approach, unknown model parameters (including variance components) are treated as random, with values drawn from specified prior distributions. Posterior distributions for the small area characteristics of interest are then obtained by integrating over these priors, with inferences based on these posterior distributions. Ghosh and Rao (1994) review the application of these estimation methods in small area estimation. Maiti (1998) has used non-informative priors for hyper-parameters when applying HB methods. You and Rao (2000) have used HB methods to estimate small area means under random effect models. Many surveys or censuses are repeated over time, and this means that auxiliary information from past values of a variable of interest can be used to improve current estimates of this variable. This "borrowing of strength over time" has been used to improve estimators for small areas.

Starting with the work of Scott and Smith (1974), Pfeffermann and Burck (1990), Tiller (1991) Rao and You (1992), Singh et al (1994), and Ghosh et al (1996) have used time series models and associated estimation methods to improve estimators for small areas. In particular, Datta *et al.* (1999) have used times series models in estimating State-level unemployment rates for the US. Often the survey or census data from which small area estimates are needed are discrete or categorical. A general approach for small area estimation based on generalized linear models is described in Ghosh *et al*. (1998). Malec *et al*. (1999) have extended the models described in Malec *et al.* (1997) by including an oversampling component in the likelihood. Farrell et al (1997) extend the mixed logistic model of Mac Gibbon and Tomberlin (1989). Moura and Migon (2001) further extend this model, introducing a component to account for spatially correlated structure in the binary response data. The naïve mean square error estimator suggested by Henderson (1975) underestimates the true mean square error of small area estimators based on mixed effect models. Kacker and Harville (1984) introduced an estimator for the mean square error of an estimator of a small area mean based on an approximation to its true mean squared error under such models. Prasad and Rao (1990) developed an approximation to the Kacker and Harville estimator. Harville and Jeske (1992) also developed approximations to the mean square error of predictors. Rivest and Belmonte (2000) have developed the conditional mean square error of a small area estimator.

***The complementary roles of design and model***

Small area estimation is an example of a generic problem in which design-based methods, used on their own, are not satisfactory because making provisions for sufficient sample sizes within every one of a large number of areas is not feasible. This does not justify discarding design-based methods altogether and placing all our faith in model-based methods. Given perfect implementation of a design, the properties of the design-based estimators are either known exactly or their levels of approximation are well understood. In contrast, the properties of model-based methods are usually well understood only when the model holds. A model can rarely be confirmed logically; in practice, we regard a model as appropriate if, after a reasonable analytical-diagnostic effort, we fail to find any contradiction with the adopted model. The model draws on information about variables other than those involved in design-based estimation. Such information is called auxiliary. A key issue is to identify auxiliary information (variables) that would contribute to small area estimation, and to construct estimators that take full advantage of this information. When we are not certain about the quality of the auxiliary information it is essential to ensure that using such information would not be counterproductive – incorporating it must not make the resulting estimators less efficient (on average) than if the information is not used at all. Design-based (direct) small area estimation can be motivated as estimation for each area, in total isolation from the information external to the area and the survey variable concerned. Such estimators are usually formed from the estimator for the whole of the population by restricting it to the area of interest. With known probabilities of marginal and joint inclusion, most direct estimators are design-unbiased. Modified estimators use some external information but still aim for approximate design-unbiasedness. Indirect estimators combine information that originates from the design and a model. Their aim is to have as small a mean squared error, on average, as possible. The mean squared error is defined over replications of the sampling design, but the averaging is with respect to the model. Synthetic estimators are based on a model, making little use of the design information. Composite estimators combine a design-based (direct) and a model-based estimator. The model-based estimator is derived from a fitted model. This fit is often derived using standard software for linear regression or maximum likelihood estimation, ignoring the sampling design. However, improvement can be achieved by incorporating the sampling design in the estimation. In composite estimation, the principal role of the model is to describe the similarity of the areas. The model refers to a super-population, of which the studied population is meant to be a single realisation. Extensions of this model replace the grand mean μ with (univariate or multivariate) linear regression, or with Marker (1999) reviews models used for small area estimation from a different perspective. He presents the models in a natural order of complexity, from direct estimates through symptomatic regression and variance component models to the generalised linear models. In the context of small area estimation, these models are not meant to be precise descriptions of the observed processes, merely models that promote more precise estimation of the area-level quantities of interest. The goal of the modelling should be exploitation of all the available information, as opposed to the search for a 'correct' model.

Implementation of a typical large-scale survey encounters a host of contingencies (deviations from the planned design) at every stage. First, the frame is not perfect, containing duplicates, units from outside the population, and missing some units that may be systematically different from the units in the frame. Next, the survey relies on voluntary participation, and the subjects may withdraw their co-operation at any point. Further, the process of recording and transferring the data may not be perfect. As a consequence, the original inclusion probabilities do not describe the sampling and data collection process precisely. This problem is commonly addressed by adjusting the sampling weights so that they would accurately reflect the participation among the responding subjects. The resulting weights are estimates; they are subject to sampling variation.

Auxiliary information is crucial for efficient small area estimation. Such information may be collected in the survey, may be available from a different survey or from a census or administrative records. The information in existing sources is generally acknowledged as an important source for survey design. The realisation that such information should be exploited also in the analysis is relatively recent. This is most relevant to organisations that conduct or have access to many surveys, and to small area estimation and inferences that involve a large number of parameters.

### *Development of SAE methods*

Significant impact on poverty estimates for small had the EURAREA project founded by Eurostat to investigate SAE methods and their application. The EURAREA (2004, Heady *et al.*, 2004) project experiences permitted undertake the search on the most important surveys and the potential covariates bases. Among the surveys were tested Labour Force Survey (in Polish BAEL), Survey on smallest enterprises (SP3) and Household Budget Survey (HBS). The following administrative registers as auxiliary information were used: in Unemployed register (PULS) for BAEL, personal tax data (POLTAX) for SP3, social assistance register (POMOST).

Important impact on development of poverty estimates for small area had also USA publications, and mainly Citro, Kalton *et al.*, (1997, 2000). Citro and Kalton (1997, 2000) give details of Small Area Income and Poverty Estimates (SAIPE) an extensive programme of estimation related to poverty in small areas in the U.S.A. The programme uses data from several sources: the latest Decennial Census, information from the Internal Revenue Service, Food Stamps Program records, and the Current Population Survey (CPS). Composite estimation is used for state- and county-level counts of children in poverty. The outcome variable, poverty status, is recorded by CPS. Many counties are not represented in CPS, so composite estimation reduces to its model component. The model for the counties is defined at the county level. Individual models are not feasible because the complete set of model covariates is recorded for very few subjects; the overlap of the surveys is very sparse. The covariates are highly associated with the outcome variable: number of child exemptions of low-income families in the tax returns; number of people receiving food stamps, estimated population of the county; and number of poor school-age children in the county in the previous census. The outcome variable is the average of the direct estimates for the current, preceding and year following the year in question. The purpose of this averaging is to improve the stability of the estimates. A similar model is used for state-level composite estimation, although the outcomes are not averaged (the direct estimate for the current year is used). For county-level model, the logarithms of all the variables are used; for state-level model the variables are not transformed. The county-level estimates are adjusted by raking to agree with the state-level estimates. The sampling designs of the surveys used are not incorporated in the model, although the direct estimators use the sampling weights. The model adopted is not regarded as final, and future rounds of estimation may be based on sources that become available in the meantime, or may use different variables from the sources currently available. Model checking and improvement is an integral part of the programme.

### *Experiences in poverty estimation for small area in Poland*

The problems of the poverty rates estimation are discussed in Blaszkiewicz' papers (2007a and 2007b). She used first time in Poland the social aid register POMOST and Household Budgets to estimate the poverty rates on the NUTS4 level. The GUS yearly publishes poverty rates in Poland only for all country and in NUTS2 territorial units, but social politics need information about diversity of poverty in NUTS4 and even in NUTS5. These assessments are possible combining information form various sources (special sample surveys and administrative registers) and by using methods of small area estimations. Because of the confidentiality of statistical data, the household survey obtained for research from the GUS was without NUTS4 and NUTS5 codes. Blaszkiewicz (2007a) proposed randomized method to assign record codes to proper NUTS4 on the basis of various characteristics of the household. From numerous methods of poverty measurement she has chosen relative poverty defined (according to Eurostat's recommendation) as the share of persons with an equivalised disposable income, below the risk-of-poverty threshold, which is set at 60% of the national median equivalised disposable income. Beside direct estimator there were used: synthetic ratio and synthetic regression estimators based on linear model with area-level covariates and empirical Bayes estimators, assessing their usefulness to estimate poverty in Polish NUTS 4.

In the paper Kordos and Kubacki (1999) the possibility of poverty estimation for small area in Poland was discussed. The authors suggested to use Household Budget Survey (HBS) and administrative registers, in particular the tax register (in Poland named POLTAX). Also using the empirical Bayes procedure for that purpose was proposed.

Cz. Bracha, B. Lednicki and R. Wieczorkowski (Bracha *et al.*, 2003, 2004) applied SAE methods for estimating some unemployment characteristics by region, sub-region and poviat (county) for 1995-2002 using the Polish Labour Force Survey (PLFS) and the 2002 Population Census and Housing (PCH) results. The authors used *direct, synthetic* and *composite estimators*. Direct estimates were obtained from the PLFS in 1995—2002, and appropriate data from the 2002 CPH were used as auxiliary information. Efficiency analysis of direct, synthetic and composite estimators was conducted. The composite estimator was a combination of the direct and synthetic estimators with equal weights. In the second paper (Bracha *et al.*, 2004), the composite estimator uses data from the PLFS 2003 together with data from administrative sources that are available on Polish Public Statistics web pages. They compare similar estimates that are based on Census 2002 data which may reveal usefulness census vs. administrative data.

Interesting results are given in Kubacki (2004, 2006) papers. The author presents a synthetic review concerning methodology and results considered in Bracha *et al.* (2003, 2004) and some own results. The author discusses various methods of estimation together with evaluation of quality of such estimation related in particular with the type of auxiliary data used for "borrowing strength" and efficiency of initial estimates used in models (Kubacki, 2004). The author presents the application of Empirical Bayes (EB) and Hierarchical Bayes (HB) methods to the estimates of unemployment size for small areas using the Polish Labour Force Survey (PLFS) and auxiliary information. The constructed model includes the data obtained from published results of PLFS for regions in Poland and 2002 Census data.

E. Gołata (2004) proved the usefulness of PULS to estimate of unemployment on the NUTS 4 level. J. Paradysz (2007) reviewed the most important administrative registers as a potential source covariates in Polish regional estimation. He proved the usefulness some of them in the process of estimation at regional and local level. J. Paradysz and M. Szymkowiak (2006) present some aspects of imputation and calibration as standard methods of dealing with nonresponse in statistical surveys in Poland. The authors focused on Polish household budget surveys where the nonresponse is a significant problem. Because the nonresponse occurs on a large scale, it influences the quality of the results. The authors showed how calibration estimators which were proposed by Sixten Lundström and Carl–Erik Särndal (2005), can be used in order to improve estimations in Polish household budget surveys.

**Poverty Mapping**

If the objective of the study is beyond poverty measure at area level, that is, we are interested in producing poverty and inequality maps, large data sets are required which include reasonable measures of income or consumption expenditure and which are representative or of sufficient size at low levels of aggregation to yield statistically reliable estimates.

The Poverty Mapping is the World Bank project intended to promote usage of small area methods in recognition of poverty and programming poverty reducing social policies at local levels. The final output of the project appears in the form of a website presenting educational materials, poverty maps for more than 20 countries as well as theoretical analyses. It is intention of the project to construct poverty maps for middle-income and poor countries. Moreover, A team of the World Bank researchers provides training courses, technical assistance, capacity building, and software tools to institutions in developing countries. The tools include Geographic Information Systems (GISs) which are database management systems that use geographic location as a reference for each database record. They are capable, inter alia, to yield village-level estimates of income poverty or exclusion, however it is possible to obtain geographical distribution of any sort of data.

The methodology utilised in the Poverty Mapping Project has been developed by the World Bank experts. The details are available in the papers by Elbers, Lanjouw and Lanjouw (2002) and Elbers, Lanjouw and Lanjouw (2003). The main sources of information on socio-economic indicators are typically censuses and/on administrative data on the one side and household surveys on the other side. The general idea of the method is to combine advantages of both sources. Whereas the datasets of the first type are large enough to allow satisfactory disaggregation, they include very restricted information on individual's well-being and practically no income information. The latter type of information is available in household surveys, however they are too small to ensure adequate precision of estimates at local levels. The small area estimation techniques involve imputing measures of well-being from household survey data into large sample data. For that purpose survey income (or expenditure) variables are regressed on explanatory variables that are common to both datasets. These estimates use survey data at the lowest acceptable geographical level. In the second step the estimated coefficients are used to predict income for every household in the census. It is possible to obtain income distribution, including poverty estimates for small areas, as well as standard errors of the estimates[35].

The level of disaggregation in the Poverty Mapping Project varies between countries. For Albania maps were constructed for 374 municipalities and for 262 for Bulgaria[36]. This level is approximately equivalent to NUTS4. For all countries included into the Project income, consumption and monetary poverty rates were estimated at lowest obtainable geographical levels (i. e., for instance, for NUTS4 levels for Albania and Bulgaria) while other measures may be available at higher levels.

[35] Poverty maps are also constructed on the basis of multiple deprivation indexes in the United Kingdom (see: Section 3.2.2).

[36] Albania and Bulgaria are two only European countries included into the project.

## 3.6 References

ANED (2008), *National Report on Strategies for Social Protection and Social Inclusion, Portugal 2006-2008, A New Integrated Strategy on Social Policies*, University of Leeds.

Atkinson A., Cantillon B., Marlier E., Nolan B. (2002), *Social Indicators: The EU and social inclusion*, Oxford University Press, Oxford.

Atkinson A.B., Marlier E. and Nolan B. (2004), Indicators and Targets for Social Inclusion in the EU, *Journal of Common Market Studies*, **42**(1), pp. 47-75.

Battese G.E., Harter R.M., Fuller W.A. (1988), An error-component model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, **83**, pp. 28-36.

Bedi, T, Coudouel, A. and Simler, K. (2007), More than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions, The World Bank, Washington, DC.

Berthoud R. (2004), *Patterns of poverty across Europe*, The Policy Press, Bristol.

Betti G., Verma V. (2002), *Non-monetary or Lifestyle Deprivation*, in Giorgi L. and Verma V. (eds.), *European Social Statistics: Income, Poverty and Social Exclusion: 2nd Report*. Luxembourg: Office for Official Publications of the European Communities, pp. 76-92.

Blaszkiewicz A. (2007a) *Confidentiality and access to unit data in statistical research*, in Paradysz J. (eds.) *Regional statistics in uniting Europe*, Centre of Regional Statistics Publishing, Poznan pp. 65-80. (in Polish).

Blaszkiewicz A. (2007b) *Administrative register POMOST as an information source in life condition research*, in Paradysz J. (eds.) *Regional statistics in uniting Europe*, Centre of Regional Statistics Publishing, Poznan pp. 249-264. (in Polish).

Bracha Cz., Lednicki B. and Wieczorkowski R. (2004), *Estimation of Data from the Polish Labour Force Surveys by poviats (counties) in 1995—2002* (in Polish), Central Statistical Office of Poland, Warsaw.

Chakravarty S. A. And D’Ambrosio C. (2003), *The Measurement of Social Exclusion*, Discussion Paper, DIW, Berlin.

Citro C. F., Cohen M. L., Kalton G. and Kirsten K. (1997), *Small Area Estimates of School-Age Children in Poverty*, Committee on Statistics, National Research Council, Washington, D.C.

Citro C. F., Kalton G. (eds); (2000), *Small Area Income and poverty estimates: priorities for 200 and beyond*, Panel on Estimates of Poverty for Small Geographic Area, Committee on National Statistics, National Research Council.

Council Regulation (EC) No 362/2008 of 14 April 2008 as regards the 2009 list of target secondary variables on material deprivation.

Dehnel E., Golata E. and Klimanek T. (2004), Consideration on Optimal Design for Small Area Estimation, *Statistics in Transition*, (6)5, pp. 725-754.

Department for Work and Pensions (2006), *UK National Report on Strategies for Social Protection and Social Inclusion 2006-2008*.

Department of Social Policy and Social Work (2003), *Scottish Indices of Deprivation 2003*, University of Oxford, Oxford.

Elbers C., Lanjouw J. O. and Lanjouw P. (2002), Micro-Level Estimation of Welfare, *Research Working Paper*, 2911, World Bank, Development Research Group, Washington, DC.

Elbers C., Lanjouw J. O., and Lanjouw P. (2003). Micro-level Estimation of Poverty and Inequality, *Econometrica*, **71**, 355-364.

Eurarea (2004), *Enhancing small area estimation techniques to meet European needs*, Project IST-2000-26290. Final report. http://www.statistics.gov.uk/eurarea/.

European Commission, Joint Report on Social Protection and Social Inclusion, 2005-2008 editions, available: http://ec.europa.eu/employment_social/spsi/joint_reports_en.

European Commission (1990), *Methodological Issues, Poverty in Figures in the Early 1980s*, Office for Official Publications of European Communities, Luxembourg.

European Commission (2003a), *Definition of Quality in Statistics*, Eurostat Working Group on Assessment of Quality in Statistics, Luxembourg, 2—3 October 2003.

European Commission (2003b), *Joint Report on Social Inclusion*, COM 773 final.

European Commission (2003c), *The European Union labour force survey, Methods and definitions – 2001*. Office for Official Publications of European Communities, Luxembourg.

European Commission (2003d), *Household Budget Surveys in the EU, Methodology and recommendations for harmonisation*, Office for Official Publications of European Communities, Luxembourg.

European Commission (2004), *Report on Social Inclusion 2004. An Analysis of the National Action Plans on Social Inclusion (2004-2006)*, submitted by the 10 new Member States, Commission Start Working Paper, SEC (2004) 256.

European Commission, (2007), *EU Labour Force Survey Database User Guide*; http://circa.europa.eu/irc/dsis/employment/info/data/eulfs/LFS_MAIN/LFSuserguide/EULFS_database.

European Commission (2008a), *Algorithms to compute overarching indicators, based on EU-SILC and adopted under the Open Method of Coordination (OMC)*, DOC LC-ILC/11/08/EN-REV.

European Commission (2008b), *Child Poverty and Well-Being in the EU – Current status and way forward*, New Report on Child Poverty and Well-Being adopted by Social Protection Commitee.

European Commission, *Eurostat* (2008c), *EU-SILC 2009 Module, Description of secondary target variables and corresponding questionnaire*, Doc. LC-ILC-DEPIV/17/08/EN.

European Commission (2008d), *Joint Report on Social Protection and Social Inclusion 2008, Social inclusion, pensions, healthcare and long-term care*, Office for Official Publications of European Communities, Luxembourg.

European Commission, Eurostat (2008e), *Material deprivation-recent development*, Doc. LC-ILC/18/08/EN/rev.

European Commission (2008f), *Portfolio of overarching indicators and Streamlined Social Inclusion, Pensions and Health Portfolios*, April 2008 Update, Office for Official Publications of European Communities, Luxembourg.

Eurostat, Data transmission for the HBS round of the reference year 2005, http://circa.europa.eu/Public/hbs

Eurostat (2004a), *European Regional Statistics: Reference Guide,* Office for Official Publications of European Communities, Luxembourg.

Eurostat (2004b), *Regions: Statistical Yearbook 2004*, Office for Official Publications of European Communities, Luxembourg.

Eurostat (2004c), Monetary Poverty in New Member States and Candidate Countries, *Statistics in Focus*, Theme 3-12/2004.

Eurostat (2004d), Monetary Poverty in EU Acceding and Candidate Countries, *Statistics in Focus*, Theme 3-21/2003.

Eurostat (2005), The European Consumer in the enlarged Union, *Statistics in Focus*, No 2.

Eurostat (2007), *Comparative EU Statistics on Income and Living Conditions: Issues and Challenges*, Office for Official Publications of European Communities, Luxembourg.

Fay R. E., Herriot R. A. (1979), Estimates of income for small places: an application of James-Stein procedure to census data, *Journal of the American Statistical Association*, **74**, pp. 269-277.

GESIS – ZUMA (2007), German System of Social Indicators: Key Indicators 1950-2005, Mannheim.

Ghosh M. (2001), *Model-dependent small area estimation: Theory and practice*, in: Lehtonen R. and Djerf K., (eds.), *Lecture Notes on Estimation for Population Domains and Small Areas,* Statistics Finland, Reviews 2001/5, Helsinki, pp. 51-108.

Giorgi L, Verma V. (2002), European Social Statistics: Income, Poverty and Social Exclusion: 2nd Report, Luxembourg: Office for Official Publications of the European Communities.

Glennerster H., Lupton R., Norden P., Power A. (1999), *Poverty, social exclusion and neighbourhood: studying the area bases of social exclusion*, CASEpaper/22. Centre for Analysis of Social Exclusion, London School of Economics, London.

Golata E. (2004), Problems of Estimate Unemployment for Small Domains in Poland, *Statistics in Transition*, (**6**)5, pp. 755-776.

Gosh M., Rao J. N. K. (1994), Small Area Estimation: An Appraisal (with discussion), *Statistical Science*, **9** (1), pp. 55-93.

Guio A. (2005), Income poverty and social exclusion in the EU25, *Statistics in Focus*, **13**.

Hagenaars AJ. M., de Vos K., Asghar Zaidi M.(1992), *Poverty Based on Micro-Data*, Erasmus University, Rotterdam.

Handerson C. R. (1950), Estimation of Genetic Parameters, *Annals of Mathematical Statistics*, **21**, pp. 309-310.

Hansen M., Hurwitz W., Madow W. (1953), *Sample Survey Methods and Theory*. Wiley, New York.

Heady P. and Ralphs M. (2004), Some Findings of the Eurarea Project - and their Implications for Statistical Policy, *Statistics in Transition*, (6)5, 641-654.

Kangas O., Ritakallio V. M. (2004), Relative to what? Cross-national picture of European poverty measured by regional, national and European standards, *Working Paper*, No. 384. Luxembourg Income Study.

Kalton, G., Kordos, J. and Platek, R. (1993), *Small Area Statistics and Survey Designs*, Vol. I: Invited Papers, Vol. II: Contributed Papers and Panel Discussion, Central Statistical Office, Warsaw.

Kordos, J. (1990), Research on Income Distribution by Size in Poland, in Dagum C. and Zenga M. (eds.), *Studies in Contemporary Economics, Income and Wealth Distribution, Inequalities and Poverty*. Springer-Verlag Berlin Heidelberg, pp. 335-351.

Kordos J. (1994), Small Area Statistics in Poland (Historical Review), *Statistics in Transition*, (1)6, pp. 783-796.

Kordos J. (2005), Some Aspects of Small Area Statistics and Data Quality, *Statistics in Transition*, (7)1, pp.102-131.

Kordos J. (2006), Impact of Different Factors on Research in Small Area Estimation in Poland, *Statistics in Transition*, (7)4, pp. 863-879.

Kordos J. and Kubacki, J. (1999), Possibilities of Poverty Estimation for Small Area in Poland, *Wiadomości Statystyczne*, No 3, pp. 4 –16 (in Polish).

Kordos J. and Paradysz, J. (2000), Some Experiments in Small Area Estimation in Poland, *Statistics in Transition*, (4)4, pp. 679-697.

Kubacki J. (2004), Application of the Hierarchical Bayes Estimation to the Polish Labour Force Survey, *Statistics in Transition*, (6)5, pp.785-796.

Kubacki J. (2006), Remarks on Using the Polish LFS Data for Unemployment Estimation by County, *Statistics in Transition*, (7)4, pp. 901-916.

Ministry of Labour and Social Policy (2008), *National Report on Strategies for Social Protection and Social Inclusion for 2006-2008 (for Poland)*, Warsaw.

Ministry of Social Affairs and Health (2006), *National Reports on Strategies for Social Protection and Social Inclusion (for Finland)*, Helsinki.

Ministry of Social Affairs of Estonia (2006), *National Report on Strategies for Social Protection and Social Inclusion 2006-2008 under the Open Method of Coordination* (for Estonia), Tallin.

MIPES (2008), Recueil Statistique Relattif à la Pauvreté et la Precarité en Ile –de –France.

Northern Ireland Statistics and Research Agency (2005), Northern Ireland Multiple Deprivation Measure 2005, Belfast.

OECD (2006), *Societé at a Glance: OECD Social Indicators*, Paris.

Paradysz J. (1998), Small Area Statistics in Poland - First Experiences and Application Possibilities, *Statistics in Transition*, (**3**)5, pp. 1003-1015.

Platek R., Rao J. N. K. Särndal C. E. and Singh M. P. (eds.) (1987), *Small Area Statistics*, John Wiley & Sons, New York.

Pfeffermann D. (2002), Small area estimation - new developments and directions, *International Statistical Review* **70**, pp. 125-143.

Rao J. N. K. (2003), *Small Area Estimation*, John Wiley & Sons, New Jersey.

Rao J. N. K. (2003), *Small Area Estimation*. Wiley, London.

Rodrigues F. C. (1999), *Income Distribution and Poverty in Portugal (1994/95), A Comparison between the European Community Household Panel and the Household Budget Survey*, CISEP, ISEG/Universidade Técnica de Lisboa.

Sarndal C-E., Swensson B., Wretman J. (1992), *Model assisted survey sampling*, Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo.

Schaible W. L. and Casady, R. J. (1994): The Development, Application, and Evaluation of Small Area Estimators*, Statistics in Transition*, (**1**)6, pp. 727-46.

Social Disadvantage Research Centre (2007), The English *Indices of Deprivation 2007*, University of Oxford, Oxford.

Statistical Policy Office (1993), *Indirect Estimators in Federal Programs*, Subcommittee on Small Area Estimation, Statistical Policy Working Papers, 21.

Statistics Canada (1998), *Statistics Canada Quality Guidelines*. Third Edition. October 1998.

Statistical Directorate and the Local Government Data Unit of the Welsh Assembly Government (2005), *Welsh Index of Multiple Deprivation*.

Stewart K. (2002), Measuring Well-Being and Exclusion in Europe's Regions, *Luxembourg Income Study Working Paper*, No 303.

Stewart K. (2003), Monitoring social exclusion in Europe's regions. Journal of European *Social Policy*, **13**(4), pp. 335-356.

UNDP – Poland (2006), *Social Exclusion and Integration in Poland: an indicators-based approach*, Warsaw.

UDDSD (2005), *Indicators of Sustainable Development*, Expert Group Meeting on Indicators Sustainable Development, New York.

Whelan C. T., Layte R., Maitre B. and Nolan B. (2001): Income, deprivation and economic strain: an analysis of the European Community Household Panel, *European Sociological Review*, **17**, pp. 357-372.

Verma V., Betti G., Lemmi A., Mulas A., Natilli M., Neri L., Salvati N. (2005), *Regional indicators to reflect social exclusion and poverty*, Final report. Project VT/2003/45, European Commission, Employment and Social Affairs D.G.

Verma V., Betti G., Natilli M., Achille L. (2006), *Indicators of Social Exclusion and Poverty in Europe's Regions*, Dipartimento di Metodi Quantitativi, University of Siena, Working paper 59.

Internet sources and links

Luxembourg Income Study website: http://www.lisproject.org

http://www.lisproject.org/publications/liswps/303.pdf

http://go.worldbank.org/TXDGYYAN00

Poverty Mapping website: http://go.worldbank.org/PSB4P6AMX0

http://www.communities.gov.uk/publications/communities/indicesdeprivation07

http://www.ile-de-france.pref.gouv.fr/mipes/documents/Mipess_donnees_31_12_2007.pdf
http://www.nisra.gov.uk/aboutus/default.asp2.htm

http://www.scotland.gov.uk/Publications/2003/02/16377/18195

http://www.childreninwales.org.uk/policy/documents/statistics/10018.html

Social Indicators Research Centre website:

http://www.gesis.org/en/services/data/social-indicators/the-german-system-of-social-indicators/

Eurostat website: http://epp.eurostat.ec.europa.eu

http://circa.europa.eu/irc/dsis/employment/info/data/eulfs/LFS_MAIN/LFSuserguide/EULFS_database http://forum.europa.eu.int/Public/irc/dsis/eusilc/library

BHPS Study Description:
http://www.esds.ac.uk/findingData/snDescription.asp?sn=5151

The BHPS User Documentation and Questionnaires
http://www.iser.essex.ac.uk/survey/bhps/documentation

Understanding Society:
http://www.understandingsociety.org.uk/

# LITERATURE REVIEW

# VOLUME II

| | |
|---|---|
| Project Acronym: | SAMPLE |
| Project Full title: | Small Area Methods for Poverty and Living Conditions Estimates |
| Project/Contract No: | EU-FP7-SSH-2007-1 |
| Grant agreement No: | 217565 |
| Work Package 2: | Small area estimation of poverty and inequality indicators |
| Document title: | WP2 Literature review |
| Date: | 30th March 2009 |
| Type of document: | Deliverable D4 |
| Status: | Final |
| Editors: | I. Molina, D. Morales, M. Pratesi and N. Tzavidis |
| Authors: | Monica Pratesi (m.pratesi@ec.unipi.it), UNIPI-DSMAE<br>Caterina Giusti (caterina.giusti@ec.unipi.it), UNIPI-DSMAE<br>Stefano Marchetti (s.marchetti@ds.unifi.it), UNIPI-DSMAE<br>Nicola Salvati (salvati@ec.unipi.it), UNIPI-DSMAE |
| | Nikos Tzavidis (nikos.tzavidis@manchester.ac.uk), CCSR |
| | Isabel Molina (imolina@est-econ.uc3m.es), UC3M<br>Aurea Grané (agrane@est-econ.uc3m.es), UC3M<br>Juan Miguel Marín (jmmarin@est-econ.uc3m.es), UC3M<br>Maria Helena Veiga (mhveiga@est-econ.uc3m.es), UC3M |
| | Domingo Morales (d.morales@umh.es), UMH<br>María Dolores Esteban (md.esteban@umh.es), UMH |
| | J.N.K. Rao (jrao@math.carleton.ca), Carleton University<br>Balbogin Nandram (balnan@wpi.edu), Carleton University |

# Contents

**Prologue** **1**

**1 Mixed models** **3**

1.1 Introduction . . . 3
1.2 Linear mixed models . . . 5
1.3 The ANOVA model . . . 7
1.4 Maximum likelihood estimation . . . 8
1.5 Residual maximum likelihood estimation . . . 8
1.6 Henderson 3 method . . . 9
1.7 Generalized linear mixed models . . . 10
1.8 References . . . 14

**2 Small Area EBLUP** **17**

2.1 Introduction . . . 17
2.2 The EBLUP and its mean squared error . . . 19
2.3 The GEBLUP and its mean squared error . . . 20
2.4 Bootstrap estimation of the MSE . . . 21
2.5 References . . . 24

**3 Time Models** **27**

3.1 introduction . . . 27
3.2 Individual-level linear model with correlated time effects . . . 29
3.3 Individual-level linear model with independent time effects . . . 30
3.4 Area-level linear model with correlated time effects . . . 31
3.5 Area-level linear model with independent time effects . . . 31
3.6 Logistic model with correlated time effects . . . 32
3.7 Logistic model with independent time effects . . . 33
3.8 References . . . 33

**4 Spatial models** **35**

4.1 Introduction . . . 35
4.2 Area level model with spatially correlated area effects . . . 36
4.3 Unit level model with spatially correlated area effects . . . 37
4.4 Unit level logit model with spatially correlated area effects . . . 37

4.5 References . . . . . . . . . . 39

**5 Small Area Estimation of Poverty indicators 41**

5.1 Introduction . . . . . . . . . . 41
5.2 FGT poverty measures for small areas . . . . . . . . . . 42
5.3 Direct estimators of poverty measures . . . . . . . . . . 42
5.4 Model-based estimation of the poverty incidence . . . . . . . . . . 43
5.4.1 Unit level model . . . . . . . . . . 43
5.4.2 Area level model . . . . . . . . . . 44
5.5 Model-based estimation of the poverty gap . . . . . . . . . . 45
5.5.1 Two parts unit level model – Approach 1 . . . . . . . . . . 45
5.5.2 Area level model . . . . . . . . . . 46
5.6 Model-based estimation of the poverty severity . . . . . . . . . . 47
5.7 Empirical Best prediction of FGT poverty measures at unit level – Approach 2 . . . . . . 48
5.7.1 Best prediction under a finite population . . . . . . . . . . 48
5.7.2 Best prediction of FGT poverty measures . . . . . . . . . . 49
5.8 Hierarchical Bayes estimation of FGT poverty measures – Approach 3 . . . . . . . . . . 51
5.9 References . . . . . . . . . . 52

**6 Quantile / M-quantile Models for Small Area Estimation 55**

6.1 Introduction . . . . . . . . . . 55
6.2 Quantile regression . . . . . . . . . . 55
6.3 M-quantile regression . . . . . . . . . . 56
6.4 Nonparametric M-quantile regression . . . . . . . . . . 56
6.5 Small area estimation with M-quantile models . . . . . . . . . . 58
6.6 Mean squared error estimation for M-quantile predictors for domains . . . . . . . . . . 61
6.7 Borrowing strength over space with M-quantile small area models . . . . . . . . . . 62
6.8 Mean squared error estimation for M-quantile GWR predictors for domains . . . . . . . 64
6.9 Nonparametric M-quantile small area estimation . . . . . . . . . . 64
6.10 References . . . . . . . . . . 67

**7 Estimation of cumulative distribution functions 71**

7.1 Introduction . . . . . . . . . . 71
7.2 Estimating cumulative distribution function from survey data . . . . . . . . . . 72
7.3 Estimating cumulative distribution function without auxiliary information . . . . . . . . 72
7.4 Estimating cumulative distribution function using auxiliary information . . . . . . . . . 73
7.5 Quantile estimation . . . . . . . . . . 76
7.6 Small area estimation of the cumulative distribution function . . . . . . . . . . 77
7.7 References . . . . . . . . . . 78

# Prologue

This literature review is the result of a collaborative work of the partners of the WP2 in the SAMPLE project. The target of the review is somewhat more ambitious than a simple bibliography recording. We have also given an introduction to the main statistical tools that will be used in the SAMPLE project to estimate poverty measures.

The manuscript is organized in seven chapters. Chapter 1 introduces the basic theory of linear mixed models (LMMs) and of generalized linear mixed models (GLMMs). Special attention is given to model fitting methods and algorithms. Chapter 2 deals with estimating linear parameters of finite populations when the underlying distribution arises from a LMM or a GLMM. The estimation of mean squared errors with either explicit formulas or bootstrap resampling methods is also addressed. Chapters 3 and 4 describe the time and spatial models to be developed in the SAMPLE project. Chapter 5 gives several procedures for small area estimation of selected poverty indicators.

The small area estimation of quantiles of the distribution of a welfare variable is also treated in the SAMPLE project. For this sake, two approaches are considered. The first one is the quantile regression, where the goal is to predict the value of a quantile (instead of the mean) of the distribution of the response variable by using auxiliary information. This is introduced and commented in Chapter 6. Finally, Chapter 7 is devoted to the estimation of cumulative distribution functions and its further use to produce quantile estimates.

This literature review has been coordinated by Domingo Morales (UMH). He has also been in charge of writing Chapters 1-3. Isabel Molina (UC3M) has been responsible for the elaboration of Chapters 4-5. Finally, Nikos Tzavidis (CCSR) and Monica Pratesi (UNIPI-DSMAE) have coordinated the production of the contents of Chapters 6-7.

# Chapter 1

# Mixed models

## 1.1 Introduction

Linear models (LMs) model the relationship between a dependent or response variable $y$ and a vector of auxiliary variables $\mathbf{x}$. Three basic hypotheses are assumed for these models: linearity, normality and independence. The first assumption states that the mean of $y$ is a linear function of the components of $\mathbf{x}$. The second assumption specifies a multivariate normal distribution for the vector of observed $y$-values. The last one is the stochastic independence of the measurements of $y$. The study of LMs is a classical matter within applied statistics and there are many books dealing at length with them. We name but a few: Graybill (1976), Seber (1977), Arnold (1981), Hocking (1985), Searle (1997) and Rencher (2000).

Generalized linear models (GLMs) extend the applicability of the LMs in two directions. The hypothesis of linearity is relaxed in the sense that a function (called *link*) of the mean of $y$ is linear in the components of $\mathbf{x}$. The hypothesis of normality is relaxed to the assumption that the distribution of $y$ belongs to the exponential family. This family of distributions includes the Gaussian or normal, binomial, Poisson, gamma, inverse gamma, geometric and negative binomial. Nelder and Wedderburn (1972) introduced the unified theory of GLMs. They discovered the underlying unity of a wide class of regression models. GLMs have become very popular among applied statisticians, overall in the field of categorical data analysis. Some books dealing with GLMs, under a broad perspective, are McCullogh and Nelder (1989), Dobson (1997), Lindsey, J.K. (2000) and Fahrmeir and Tutz (2001). For books dealing with categorical data analysis, including logistic regression and log-linear models, see Agresti (1990), Christensen R. (1990), Andersen, E.B. (1997), Lloyd (1999) and Hosmer and Lemeshow (2000).

LMs and GLMs assume that observations are drawn from the same population and are independent. Mixed models have a more complex multilevel or hierarchical structure. Observations in different levels or clusters are assumed to be independent, but observations within the same level or cluster are considered as dependent because they share common properties. For these data, we can speak about two sources of variation: between and within clusters. The possibility of modeling those sources of variation, commonly present in real data, gives a high flexibility, and therefore applicability, to mixed models.

Linear mixed models (LMMs) and generalized linear mixed models (GLMMs) handle data in which observations are not independent. That is, LMMs and GLMMs correctly model correlated errors, whereas the procedures in the LM or GLM family usually do not. Mixed models are generalizations of LMs and GLMs to better support the analysis of a dependent variable. These models allow to incorporate:

1. *Random effects:* sometimes the number of levels of a categorical explanatory variable is so large (with respect to sample size) that introduction of fixed effects for its levels would lead to poor estimates of the model parameters. If this is the case, the explanatory variable should not be introduced in a LM or GLM. Mixed models solve this problem by treating the levels of the categorical variable as random, and then predicting their values.

2. *Hierarchical effects:* response variables are often measured at more than one level; for example in nested territories in small area estimation problems. This situation can be modeled by mixed models and it is thus an appealing property of them.

3. *Repeated measures:* when several observations are collected on the same individual then the corresponding measurements are likely to be correlated rather than independent. This happens in longitudinal studies, time series data or matched-pairs designs.

4. *Spatial correlations:* when there is correlation among clusters due to their location; for example, the correlation between nearby domains may give useful information to improve predictions.

5. *Small area estimation:* where the flexibility in effectively combining different sources of information and explaining different sources of errors is of great help. Mixed models typically incorporate area-specific random effects that explain the additional between area variation in the data that is not explained by the fixed part of the model.

Books dealing with LMMs and GLMMs include Searle, Casella and McCullogh (1992), Longford (1995), McCullogh and Searle (2001), Goldstein (2003), Demidenko (2004) and Jiang (2007).

Several fitting methods are available for LMMs. Here we revise three common ones, namely maximum likelihood, residual maximum likelihood and Henderson 3. In the case of GLMMs the estimation of the model parameters is more difficult because the marginal loglikelihood is represented by an integral that cannot be explicitly evaluated. Several methods have been proposed to overcome this problem, most of them relying on Taylor linearizations. Goldstein (1991) considers a Taylor linearization of the inverse link function and then applies standard estimation procedures for linear multilevel models. Longford (1994) and Breslow and Clayton (1993) apply Laplace's method for integral approximations. Wolfinger and O'Connell (1993) approximate the conditional distribution of the difference between the response variable and its prediction, given the fixed and random effects, by a Gaussian distribution with the same first two moments. The procedure is implemented via iterated fitting of a weighted Gaussian linear mixed model to the modified dependent variable, which is obtained by a Taylor series approximation of the linked response. McCulloch (1994, 1997) and Booth and Hobert (1999) use EM-type algorithms assisted by Monte Carlo methods. Schall (1991) and McGilchrist (1994) use the PQL algorithm introduced by Breslow and Clayton (1993) in combination with a Gaussian approximation of the marginal density that provides approximate maximum likelihood estimators of variance components, obtaining a double iteration scheme. Although it is known that in some cases the method proposed by Schall (1991) may lead to inconsistent and biased estimators, this method works well in some situations (see e.g. González-Manteiga et al., 2007). Further, Schall's method is conceptually very simple and allows the adaptation of the theory for prediction and mean squared error estimation under linear mixed models to the case of generalized linear models.

The aim of this chapter is to summarize the basic mathematical theory of LMMs and GLMMs, focussing on key points that are useful for the SAMPLE project. In Section 1.2 a general description of LMMs is given. This section also shows how to estimate the regression parameters and how to predict the random effects when the variance components are known. Section 1.3 describes a widely used LMM in small area estimation, i.e. the so called ANOVA model. Sufficient conditions guaranteing that the model parameters are estimable are also listed. Sections 1.4-1.6 deals with three alternative fitting methods for LMMs; namely, maximum likelihood, residual maximum likelihood and a moment estimation procedures. Finally, Section 1.7 gives an introduction to GLMMs and describes a simple fitting algorithm.

## 1.2 Linear mixed models

Let us consider the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{1.1}$$

where $\mathbf{y}_{n\times 1}$ is the vector of observations, $\boldsymbol{\beta}_{p\times 1}$ is the vector of fixed effects, $\mathbf{u}_{q\times 1}$ is the vector of random effects, $\mathbf{X}_{n\times p}$ and $\mathbf{Z}_{n\times q}$ are the incidence matrices and $\mathbf{e}_{n\times 1}$ is the vector of random perturbations. We assume that the random effects and the perturbations are independent and normally distributed with zero means and known covariance matrices

$$var[\mathbf{u}] = E[\mathbf{u}\mathbf{u}'] = \mathbf{V}_u \qquad \text{and} \qquad var[\mathbf{e}] = E[\mathbf{e}\mathbf{e}'] = \mathbf{V}_e,$$

depending on some parameters that are called variance components. From (1.1) we obtain

$$\mathbf{V} = var[\mathbf{y}] = \mathbf{Z}\mathbf{V}_u\mathbf{Z}' + \mathbf{V}_e$$

and we assume that $\mathbf{V}$ is not singular.

### BLUE of $\boldsymbol{\beta}$

Let us first assume that the variance components of model (1.1) are known, i.e. $\mathbf{V}$ is known. The regression parameter $\boldsymbol{\beta}$ can be estimated by applying the weighted least squares method

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}}(\mathbf{e}^{*t}\mathbf{e}^*), \quad \text{with } \mathbf{e}^* = \mathbf{V}^{-1/2}(\mathbf{Z}\mathbf{u} + \mathbf{e}),$$

to obtain

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \tag{1.2}$$

which is also the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and coincides with the maximum likelihood estimator under normality, i.e.

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} \left\{ -\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}.$$

### BLUP of $\mathbf{u}$

The BLUP of $\mathbf{u}$ is

$$\hat{\mathbf{u}} = \mathbf{V}_u\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right). \tag{1.3}$$

See e.g. Searle (1971), 458-462, for details on the derivation of the BLUP of $\mathbf{u}$.

### BLUP of a mixed effect

A mixed effect is a linear combination of both fixed and random effects. Here we consider particular mixed effects that will be useful under the small area estimation framework. Concretely, we consider target quantities of the form $\tau = \mathbf{a}_r'(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{Z}_r\mathbf{u})$, where $\mathbf{a}_r$ $(k \times 1)$, $\mathbf{X}_r$ $(k \times p)$ and $\mathbf{Z}_r$ $(k \times q)$ are known. Let $\hat{\tau} = \mathbf{g}'\mathbf{y} + g_0$ be a linear estimator of $\tau$, where $\mathbf{g}$ $(n \times 1)$ and $g_0$ $(1 \times 1)$ are selected in such a way that:

1. $\hat{\tau}$ is unbiased, i.e.

   $$E[\tau] = \mathbf{a}_r'\mathbf{X}_r\boldsymbol{\beta} \quad \text{and} \quad E[\hat{\tau}] = \mathbf{g}'\mathbf{X}\boldsymbol{\beta} + g_0$$

   coincide. Then, it holds $g_0 = 0$ and $\mathbf{a}_r'\mathbf{X}_r = \mathbf{g}'\mathbf{X}$.

2. $\hat{\tau}$ minimizes the prediction error

   $$\begin{aligned} E[(\hat{\tau} - \tau)^2] &= \text{var}(\hat{\tau} - \tau) = \text{var}(\mathbf{g}'\mathbf{y} - \mathbf{a}_r'\mathbf{X}_r\boldsymbol{\beta} - \mathbf{a}_r'\mathbf{Z}_r\mathbf{u}) = \text{var}(\mathbf{g}'\mathbf{y} - \mathbf{a}_r'\mathbf{Z}_r\mathbf{u}) \\ &= \mathbf{g}'\mathbf{V}\mathbf{g} + \mathbf{a}_r'\mathbf{Z}_r\mathbf{V}_u\mathbf{Z}_r'\mathbf{a}_r - 2\mathbf{g}'\mathbf{C}\mathbf{Z}_r'\mathbf{a}_r, \end{aligned}$$

   where $\mathbf{C} = \text{cov}(\mathbf{y}, \mathbf{u}) = \mathbf{Z}\mathbf{V}_u$.

The problem is

$$\text{Minimize } \text{var}(\hat{\tau} - \tau), \quad \text{subject to } \mathbf{a}_r'\mathbf{X}_r = \mathbf{g}'\mathbf{X}.$$

The solution to this problem is the Best Linear Unbiased Predictor (BLUP), given by

$$\hat{\tau}_B = \mathbf{a}_r'\left[\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\mathbf{C}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right], \tag{1.4}$$

where $\hat{\boldsymbol{\beta}}$ is the weighted least squared estimator of $\boldsymbol{\beta}$ given in (1.2).

## 1.3 The ANOVA model

Let us consider the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \ldots + \mathbf{Z}_m\mathbf{u}_m + \mathbf{e}\,, \tag{1.5}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$ is the vector of observations of the response variable, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the vector of fixed effects, $\mathbf{u}_i = (u_{i1}, \ldots, u_{iq_i})'$ is a vector with the $q_i$ levels of the $i$-th random factor, $\mathbf{e} = (e_1, \ldots, e_n)'$ is the vector of random errors, and $\mathbf{X}$, $\mathbf{Z}_1, \ldots, \mathbf{Z}_m$ are design matrices of dimensions $n \times p$, $n \times q_1, \ldots, n \times q_m$ respectively. Model (1.5) can be expressed as (1.1) by taking

$$\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_m], \quad \mathbf{u} = [\mathbf{u}_1', \ldots, \mathbf{u}_m']' \quad \text{and} \quad q = \sum_{i=1}^{m} q_i.$$

The following hypotheses are sufficient to guarantee that the model parameters are estimable:

(F1) $\mathbf{u}_1, \ldots, \mathbf{u}_m$, $\mathbf{e}$ are independent and

$$\mathbf{e} \sim N_n(\mathbf{0}, \sigma_0^2\boldsymbol{\Sigma}_e), \quad \mathbf{u}_i \sim N_{q_i}(\mathbf{0}, \sigma_i^2\boldsymbol{\Sigma}_{u_i}),\ i = 1, \ldots, m,$$

with $\boldsymbol{\Sigma}_e$ and $\boldsymbol{\Sigma}_{u_i}$, $i = 1, \ldots, m$, known.

(F2) rg$(\mathbf{X}) = p\,.$

(F3) $n \geq p + m + 1\ .$

(F4) rg$(\mathbf{X} : \mathbf{Z}_i) > p,\ i = 1, \ldots, m.$

(F5) The matrices $\mathbf{G}_0 = \boldsymbol{\Sigma}_e, \mathbf{G}_1 = \mathbf{Z}_1\boldsymbol{\Sigma}_{u_1}\mathbf{Z}_1', \ldots, \mathbf{G}_m = \mathbf{Z}_m\boldsymbol{\Sigma}_{u_m}\mathbf{Z}_m'$ are linearly independent, i.e. $\sum_{i=0}^{m} \alpha_i\mathbf{G}_i = \mathbf{0}$ implies $\alpha_i = 0$, $i = 0, 1, \ldots, m$.

(F6) $\mathbf{Z}_i$ contains zeros and ones, in such a way that there is exactly one 1 in each row, and at least one 1 in each column, $i = 1, \ldots, m$.

Let us consider the parameters $\sigma^2 = \sigma_0^2$, $\varphi_i = \sigma_i^2/\sigma_0^2$, $i = 1, \ldots, m$, in the model (1.5). It holds that

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}),\ \text{with}\ \mathbf{V} = \sigma^2\boldsymbol{\Sigma}_e + \sigma^2\sum_{i=1}^{m}\varphi_i\mathbf{G}_i = \sigma^2\boldsymbol{\Sigma}.$$

Let us define $\boldsymbol{\varphi}' = (\sigma^2, \varphi_1, \ldots, \varphi_m)$ and $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \boldsymbol{\varphi}')$. The parameter space is

$$\Theta = \{\boldsymbol{\theta}' = (\boldsymbol{\beta}', \boldsymbol{\varphi}') :\ \boldsymbol{\beta} \in R^p,\ \sigma^2 > 0,\ \varphi_i \geq 0,\ i = 1, \ldots, m\}\,. \tag{1.6}$$

and the density of $\mathbf{y}$, given $\boldsymbol{\theta}$, is

$$f_{\boldsymbol{\theta}}(\mathbf{y}) = (2\pi)^{-n/2}|\mathbf{V}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}\,. \tag{1.7}$$

## 1.4 Maximum likelihood estimation

Consider model (1.5) with parameters $\sigma^2 = \sigma_0^2$, $\varphi_i = \sigma_i^2/\sigma_0^2$, $i = 1, \dots, m$. Assume that hypotheses (F1)-(F6) hold. The maximum likelihood (ML) estimator, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\varphi}}')'$, of $\boldsymbol{\theta}$ is defined as

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta} \in \Theta} f_{\boldsymbol{\theta}}(\mathbf{y}) = \text{argmax}_{\boldsymbol{\theta} \in \Theta} \log f_{\boldsymbol{\theta}}(\mathbf{y}) \,.$$

For a given vector of observations $\mathbf{y}$, the likelihood of $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \,.$$

The loglikelihood function is

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

The components of the vector of scores (first-order partial derivatives of $l(\boldsymbol{\theta})$) are

$$S_{\boldsymbol{\beta}} = \frac{1}{\sigma^2} \mathbf{X}' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{1.8}$$

$$S_{\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{1.9}$$

$$S_{\varphi_i} = -\frac{1}{2} \text{trace}(\boldsymbol{\Sigma}^{-1} \mathbf{G}_i) + \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} \mathbf{G}_i \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \; i = 1, \dots, m. \tag{1.10}$$

The second-order partial derivatives of $l(\boldsymbol{\theta})$ are

$$\begin{array}{ll}
H_{\boldsymbol{\beta\beta}} = -\frac{1}{\sigma^2} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}, & H_{\boldsymbol{\beta}\sigma^2} = -\frac{1}{\sigma^4} \mathbf{X}' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\
H_{\boldsymbol{\beta}\varphi_i} = -\frac{1}{\sigma^2} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{G}_i \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), & H_{\sigma^2\sigma^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\
H_{\sigma^2\varphi_i} = -\frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} \mathbf{G}_i \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), & \\
\multicolumn{2}{l}{H_{\varphi_i\varphi_j} = \frac{1}{2} tr(\boldsymbol{\Sigma}^{-1} \mathbf{G}_j \boldsymbol{\Sigma}^{-1} \mathbf{G}_i) - \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} \mathbf{G}_j \boldsymbol{\Sigma}^{-1} \mathbf{G}_i \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).}
\end{array}$$

Taking negative expectations, we obtain the components of the Fisher information matrix

$$\begin{array}{lll}
F_{\boldsymbol{\beta\beta}} = \frac{1}{\sigma^2} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}, & F_{\boldsymbol{\beta}\sigma^2} = \mathbf{0}, & F_{\boldsymbol{\beta}\varphi_i} = \mathbf{0}, \\
F_{\sigma^2\sigma^2} = \frac{n}{2\sigma^4}, & F_{\sigma^2\varphi_i} = \frac{1}{2\sigma^2} tr(\boldsymbol{\Sigma}^{-1} \mathbf{G}_i), & F_{\varphi_i\varphi_j} = \frac{1}{2} tr(\boldsymbol{\Sigma}^{-1} \mathbf{V}_j \boldsymbol{\Sigma}^{-1} \mathbf{G}_i).
\end{array}$$

The Fisher scoring algorithm can be used to calculate the numerical values of the ML estimators. This method updates the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ with the equations

$$\boldsymbol{\beta}^{i+1} = \boldsymbol{\beta}^i + F(\boldsymbol{\beta}^i)^{-1} S(\boldsymbol{\beta}^i), \qquad \boldsymbol{\varphi}^{i+1} = \boldsymbol{\varphi}^i + F(\boldsymbol{\varphi}^i)^{-1} S(\boldsymbol{\varphi}^i).$$

## 1.5 Residual maximum likelihood estimation

Consider again model (1.5) with parameters $\sigma^2 = \sigma_0^2$, $\varphi_i = \sigma_i^2/\sigma_0^2$, $i = 1, \dots, m$, and assume that hypotheses (F1)-(F6) hold. The residual maximum likelihood (REML) estimation method reduces the bias of the ML estimators of the variance components by transforming the data vector $\mathbf{y}$ to $\mathbf{y}^\star = (\mathbf{y}_1^\star, \mathbf{y}_2^\star) = (\mathbf{K}_1\mathbf{y}, \mathbf{K}_2\mathbf{y})$, where matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ are such that $\mathbf{K}_1\mathbf{X} = \mathbf{0}$ and $\mathbf{y}_1^\star$ is independent of $\mathbf{y}_2^\star$. The resulting loglikelihood of $\boldsymbol{\theta}$, given $\mathbf{y}^*$, can be decomposed in the sum of the following two terms

$$\begin{aligned} l(\boldsymbol{\beta}) &= -\frac{p\log 2\pi}{2} - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \\ l(\boldsymbol{\varphi}) &= -\frac{1}{2}(n-p)\log 2\pi - \frac{1}{2}(n-p)\log\sigma^2 - \frac{1}{2}\log|\mathbf{K}'\boldsymbol{\Sigma}\mathbf{K}| - \frac{1}{2\sigma^2}\mathbf{y}'\mathbf{P}\mathbf{y}, \end{aligned}$$

where $\mathbf{P} = \mathbf{K}(\mathbf{K}'\boldsymbol{\Sigma}\mathbf{K})^{-1}\mathbf{K}'$ and $\mathbf{K} = \boldsymbol{\Sigma}_e^{-1} - \boldsymbol{\Sigma}_e^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_e^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_e^{-1}$. Taking partial derivatives of $l(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and equating to zero, we get the REML estimator of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}}_{REML} = \left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y},$$

where $\hat{\mathbf{V}} = \sum_{i=0}^m \hat{\sigma}_i^2\mathbf{G}_i$ and $\hat{\sigma}_0^2, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_m^2$ are the REML estimators of $\sigma_0^2, \sigma_1^2, \ldots, \sigma_m^2$ obtained by maximizing $l(\boldsymbol{\varphi})$. Now, taking partial derivatives of $l(\boldsymbol{\varphi})$ with respect to the elements of $\boldsymbol{\varphi}$ we obtain

$$\begin{aligned} S_{\sigma^2} &= -\frac{n-p}{2\sigma^2} + \frac{1}{2\sigma^4}\mathbf{y}'\mathbf{P}\mathbf{y}, \\ S_{\varphi_i} &= -\frac{1}{2}\text{trace}(\mathbf{P}\mathbf{G}_i) + \frac{1}{2\sigma^2}\mathbf{y}'\mathbf{P}\mathbf{G}_i\mathbf{P}\mathbf{y}, \quad i = 1, \ldots, m. \end{aligned}$$

Taking partial derivatives and negative expectations, the components of the Fisher information matrix are obtained. They are

$$F_{\sigma^2\sigma^2} = -\frac{n-p}{2\sigma^4} + \frac{1}{\sigma^4}tr(\mathbf{P}\boldsymbol{\Sigma}), \quad F_{\sigma^2\varphi_i} = \frac{1}{2\sigma^2}tr(\mathbf{P}\mathbf{G}_i), \quad F_{\varphi_i\varphi_j} = \frac{1}{2}tr(\mathbf{P}\mathbf{V}_j\mathbf{P}\mathbf{G}_i).$$

Again, the Fisher scoring algorithm can be used to calculate the numerical values of the REML estimators of the variance components. In this algorithm, the updating equation is

$$\boldsymbol{\varphi}^{i+1} = \boldsymbol{\varphi}^i + F(\boldsymbol{\varphi}^i)^{-1}S(\boldsymbol{\varphi}^i).$$

## 1.6 Henderson 3 method

The Henderson 3 (H3) method, also called “fitting constants method” since it was introduced by Henderson (1953), treats the effects $\mathbf{u}_1, \ldots, \mathbf{u}_m$ of model (1.5) as fixed and uses the method of moments to fit the model. The basic idea is to obtain a collection of equations containing the variance components and the expectations of some selected quadratic forms. This method does not require the assumption of normality. Here we use the notation $\mathbf{y} \sim (\boldsymbol{\mu}_y, \mathbf{V}_y)_n$ for a random vector $\mathbf{y}$ of dimension $n$ with mean vector $\boldsymbol{\mu}_y$ and covariance matrix $\mathbf{V}_y$. We consider the new hypotheses

(F0) $\mathbf{u}_1, \ldots, \mathbf{u}_m$, $\mathbf{e}$ are independent and $\mathbf{e} \sim (\mathbf{0}, \sigma_0^2\mathbf{W}^{-1})_n$, $\mathbf{u}_i \sim (\mathbf{0}, \sigma_i^2\mathbf{I}_{q_i})_{q_i}$, $i = 1, \ldots, m$, with $\mathbf{W}$ known, and we assume that (F0) and (F2)-(F6) hold. Let us write model (1.5) in the form

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}, \tag{1.11}$$

where both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are assumed to be fixed effects, and define its reduced version

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{e}, \tag{1.12}$$

For $i = 1, \dots, m$, we consider the case

$$\mathbf{X}_1 = \mathbf{X}_1^{(i)} = (\mathbf{X}, \mathbf{Z}_1, \dots, \mathbf{Z}_{i-1}),\ \boldsymbol{\beta}_1 = \boldsymbol{\beta}^{(i)} \quad \text{and} \quad \mathbf{X}_2 = \mathbf{X}_2^{(i)} = (\mathbf{Z}_i, \dots, \mathbf{Z}_m),\ \boldsymbol{\beta}_2 = \mathbf{u}^{(i)},$$

where $\boldsymbol{\beta}^{(i)} = (\boldsymbol{\beta}', \mathbf{u}_1', \dots, \mathbf{u}_{i-1}')'$ and $\mathbf{u}^{(i)} = (\mathbf{u}_i', \dots, \mathbf{u}_m')'$. Calculating the expectations of the sum of squares of residuals $SSR(\boldsymbol{\beta}_1)$ and $SSR(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ of models (1.12) and (1.11) for each $i = 1, \dots, m$, we obtain a system of linear equations in the variances components. Solving this system we obtain the H3 estimators of $\sigma_0^2, \sigma_1^2, \dots, \sigma_m^2$, given by

$$\begin{aligned}
\hat\sigma_0^2 &= \frac{\mathbf{y}'\mathbf{M}_{m+1}\mathbf{y}}{n - \text{rank}(\mathbf{X}_1^{(m+1)})}, \\
\hat\sigma_m^2 &= \frac{\mathbf{y}'\mathbf{M}_m\mathbf{y} - \mathbf{y}'\mathbf{M}_{m+1}\mathbf{y} - \hat\sigma_0^2\left[\text{rank}(X_1^{(m+1)}) - \text{rank}(X_1^{(m)})\right]}{\text{trace}(\mathbf{L}_m)}, \\
\vdots \quad &\ \vdots \\
\hat\sigma_i^2 &= \frac{\mathbf{y}'\mathbf{M}_i\mathbf{y} - \mathbf{y}'\mathbf{M}_{m+1}\mathbf{y} - \hat\sigma_0^2\left[\text{rank}(X_1^{(m+1)}) - \text{rank}(X_1^{(i)})\right] - \sum_{j=i+1}^m \hat\sigma_j^2 \text{trace}(\mathbf{L}_j)}{\text{trace}(\mathbf{L}_i)}, \\
\vdots \quad &\ \vdots \\
\hat\sigma_1^2 &= \frac{\mathbf{y}'\mathbf{M}_1\mathbf{y} - \mathbf{y}'\mathbf{M}_{m+1}\mathbf{y} - \hat\sigma_0^2\left[\text{rank}(X_1^{(m+1)}) - \text{rank}(X_1^{(1)})\right] - \sum_{j=2}^m \hat\sigma_j^2\, \text{trace}(\mathbf{L}_j)}{\text{trace}(\mathbf{L}_i)},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{M}_i &= \mathbf{W} - \mathbf{W}\mathbf{X}_1^{(i)}(\mathbf{X}_1^{(i)t}\mathbf{X}_1^{(i)})^{-1}\mathbf{X}_1^{(i)t}\mathbf{W}, \\
\mathbf{L}_i &= \mathbf{Z}_i'\mathbf{W}[\mathbf{W}^{-1} - \mathbf{X}_1^{(i)}(\mathbf{X}_1^{(i)t}\mathbf{W}\mathbf{X}_1^{(i)})^{-1}\mathbf{X}_1^{(i)t}]\mathbf{W}\mathbf{Z}_i\,.
\end{aligned}$$

For more details see Searle at al. (1992), 202-208, or Searle (1971), 443-445. An estimator of $\boldsymbol{\beta}$ and predictors of $\mathbf{u}_1, \dots, \mathbf{u}_m$ can be obtained by replacing the variance components $\sigma_0^2, \sigma_1^2, \dots, \sigma_m^2$ by their estimators $\hat\sigma_0^2, \hat\sigma_1^2, \dots, \hat\sigma_m^2$ in (1.2) and (1.3).

## 1.7 Generalized linear mixed models

Let $\mathbf{u}_1, \dots, \mathbf{u}_m$ be independent random vectors satisfying

$$\mathbf{u}_i \sim N_{q_i}(\mathbf{0}, \varphi_i\boldsymbol{\Sigma}_{ui}), \quad i = 1, \dots, m,$$

where $\mathbf{\Sigma}_{u1},\dots,\mathbf{\Sigma}_{um}$ are known symmetric and positive definite matrices. Let us define $\mathbf{u}=(\mathbf{u}_1',\dots,\mathbf{u}_m')'$ and $q=\sum_{i=1}^m q_i$. Then

$$\mathbf{u}\sim N_q(\mathbf{0},\mathbf{V}_u),\qquad \text{with}\quad \mathbf{V}_u=\text{diag}(\varphi_1\mathbf{\Sigma}_{u1},\dots,\varphi_m\mathbf{\Sigma}_{um}). \tag{1.13}$$

Let $y_1,\dots,y_n$ be independent random variables whose densities, given $\mathbf{u}$, belong to the exponential family; i.e.

$$f(y_j|\mathbf{u})=c(y_j)\exp\left\{\theta_j y_j-b(\theta_j)\right\},\quad j=1,\dots,n, \tag{1.14}$$

where $\theta_j\in\Theta$, $j=1,\dots,n$, are called *natural parameters*. Let $\mu_j=\mu(\theta_j)$ and $\sigma_j^2=\sigma^2(\theta_j)$ be the mean and the variance of $y_j$ given $\mathbf{u}$. It holds that

$$\mu(\theta_j)=\frac{\partial b(\theta_j)}{\partial\theta_j},\quad \sigma^2(\theta_j)=\frac{\partial\mu_j}{\partial\theta_j}=\frac{\partial^2 b(\theta_j)}{\partial\theta_j^2},\quad j=1,\dots,n.$$

Let us define the quantities

$$\eta_j=\mathbf{x}_j\boldsymbol{\beta}+\mathbf{z}_j\mathbf{u},\quad j=1,\dots,n,$$

where $\mathbf{x}_j$ and $\mathbf{z}_j$ are known vectors of dimensions $1\times p$ and $1\times q$ respectively. A GLMM assumes that

$$g(\mu(\theta_j))=\eta_j,\quad j=1,\dots,n,$$

where $g:M\mapsto R$ is an injective function called *link function* and $M\subset R$ is the set of all possible values of $\mu(\theta_j)$. The natural and the mean parameters depend on the linear predictors $\eta_j$ through the functions $d=(g\circ\mu)^{-1}$ and $h=g^{-1}$; i.e.

$$\theta_j=d(\eta_j)\quad\text{and}\quad \mu_j=h_j(\eta_j),\quad j=1,\dots,n.$$

For the *natural link*, $g=\mu^{-1}$, the GLMM assumes

$$\theta_j=\mathbf{x}_j\boldsymbol{\beta}+\mathbf{z}_j\mathbf{u},\quad j=1,\dots,n.$$

The joint probability density function (p.d.f.) of $\mathbf{y}=(y_1,\dots,y_n)'$ given $\mathbf{u}$ is

$$f_1(\mathbf{y}|\mathbf{u})=\left[\prod_{j=1}^n c(y_j)\right]\exp\left\{\sum_{j=1}^n\left(\theta_j' y_j-b(\theta_j)\right)\right\}$$

and the loglikelihood is

$$l_1(\mathbf{y}|\mathbf{u})=c_1+\sum_{j=1}^n\left(\theta_j' y_j-b(\theta_j)\right),$$

where $c_1$ is a constant not depending on $\theta_j$. The p.d.f. of $\mathbf{u}$ is

$$f_2(\mathbf{u})=(2\pi)^{-\nu/2}\left|\mathbf{V}_u\right|^{-1/2}\exp\left\{\frac{-1}{2}\mathbf{u}'\mathbf{V}_u^{-1}\mathbf{u}\right\},$$

and the loglikelihood of $\mathbf{u}$ is

$$l_2(\mathbf{u}) = c_2 - \frac{1}{2}\left\{\log|\mathbf{V}_u| + \mathbf{u}'\mathbf{V}_u^{-1}\mathbf{u}\right\} = c_2 - \frac{1}{2}\left\{\sum_{i=1}^{m}\left(q_i \log\varphi_i + \log|\mathbf{\Sigma}_{ui}| + \varphi_i^{-1}\mathbf{u}_i'\mathbf{\Sigma}_{ui}^{-1}\mathbf{u}_i\right)\right\}.$$

The *penalized quasi-likelihood* (PQL) method is described by Breslow and Clayton (1993), and consist of maximizing

$$l(\mathbf{y},\mathbf{u}) = l_1(\mathbf{y}|\mathbf{u}) + l_2(\mathbf{u}),$$

with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$. The vector of scores and the Fisher information matrix associated to $l(\mathbf{y},\mathbf{u})$ are

$$\mathbf{S}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{S}_\beta(\boldsymbol{\theta}) \\ \mathbf{S}_u(\boldsymbol{\theta}) \end{bmatrix}, \quad \mathbf{F}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{F}_{\beta\beta}(\boldsymbol{\theta}) & \mathbf{F}_{\beta u}(\boldsymbol{\theta}) \\ \mathbf{F}_{u\beta}(\boldsymbol{\theta}) & \mathbf{F}_{uu}(\boldsymbol{\theta}) \end{bmatrix},$$

where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)'$,

$$\begin{array}{ll} \mathbf{S}_\beta(\boldsymbol{\theta}) = \sum_{j=1}^n \mathbf{x}_j'[y_j - \mu_j], & \mathbf{S}_u(\boldsymbol{\theta}) = \sum_{j=1}^n \mathbf{z}_j'[y_j - \mu_j] - \mathbf{V}_u^{-1}\mathbf{u}, \\ \mathbf{F}_{\beta\beta}(\boldsymbol{\theta}) = \sum_{j=1}^n \sigma_j^2\, \mathbf{x}_j'\mathbf{x}_j, & \mathbf{F}_{\beta u}(\boldsymbol{\theta}) = \sum_{j=1}^n \sigma_j^2\, \mathbf{x}_j'\mathbf{z}_j, \\ \mathbf{F}_{u\beta}(\boldsymbol{\theta}) = \sum_{j=1}^n \sigma_j^2\, \mathbf{z}_j'\mathbf{x}_j, & \mathbf{F}_{uu}(\boldsymbol{\theta}) = \sum_{j=1}^n \sigma_j^2\, \mathbf{z}_j'\mathbf{z}_j - \mathbf{V}_u^{-1}. \end{array}$$

If $\varphi_1,\ldots,\varphi_m$ are known, the Fisher-scoring algorithm for obtaining the PQL estimator of $\boldsymbol{\beta}$ and the predictor of $\mathbf{u}$ works as follows:

(A.1) *Step 0:* Set initial values $r = 0$, $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}^{initial}$ and $\mathbf{u}^{(0)} = \mathbf{u}^{initial}$.

(A.2) *Iteration* $r+1$*:* Calculate $\theta_j^{(r)} = \mathbf{x}_j\boldsymbol{\beta}^{(r)} + \mathbf{z}_j\mathbf{u}^{(r)}$, $\mu_j^{(r)} = \mu(\theta_j^{(r)})$, $\sigma_j^{2(r)} = \sigma^2(\theta_j^{(r)})$, $j = 1,\ldots,n$. Update $\boldsymbol{\beta}^{(r)}$ and $\mathbf{u}^{(r)}$ as

$$\begin{bmatrix} \boldsymbol{\beta}^{(r+1)} \\ \mathbf{u}^{(r+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}^{(r)} \\ \mathbf{u}^{(r)} \end{bmatrix} + \left[\mathbf{F}(\theta^{(r)})\right]^{-1}\mathbf{S}(\theta^{(r)}).$$

(A.3) *End:* Repeat step (A.2) until convergence of $\boldsymbol{\beta}^{(r)}$ and $\mathbf{u}_i^{(r)}$, $i = 1,\ldots,m$.

In the following, we describe a method for obtaining estimates of $\varphi_1,\ldots,\varphi_m$. For this, let us denote $l_1(\boldsymbol{\beta},\mathbf{u}) = l_1(\mathbf{y}|\mathbf{u})$ and let $\boldsymbol{\beta}^\circ$ and $\mathbf{u}^\circ$ be the values that maximize $l_1(\boldsymbol{\beta},\mathbf{u})$. Consider a Taylor series expansion of $l_1(\boldsymbol{\beta},\mathbf{u})$ around $\boldsymbol{\beta}^\circ$ and $\mathbf{u}^\circ$, i.e.

$$\begin{aligned} l_1(\boldsymbol{\beta},\mathbf{u}) &\approx l_1(\boldsymbol{\beta}^\circ,\mathbf{u}^\circ) + \left(\frac{\partial l_1(\boldsymbol{\beta}^\circ,\mathbf{u}^\circ)}{\partial\boldsymbol{\beta}}, \frac{\partial l_1(\boldsymbol{\beta}^\circ,\mathbf{u}^\circ)}{\partial\mathbf{u}}\right)\begin{pmatrix} \boldsymbol{\beta}-\boldsymbol{\beta}^\circ \\ \mathbf{u}-\mathbf{u}^\circ \end{pmatrix} \\ &+ \frac{1}{2}\left(\boldsymbol{\beta}' - \boldsymbol{\beta}^{\circ t}, \mathbf{u}' - \mathbf{u}^{\circ t}\right)\begin{pmatrix} \frac{\partial^2 l_1(\boldsymbol{\beta}^\circ,\mathbf{u}^\circ)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \frac{\partial^2 l_1(\boldsymbol{\beta}^\circ,\mathbf{u}^\circ)}{\partial\boldsymbol{\beta}\partial\mathbf{u}'} \\ \frac{\partial^2 l_1(\boldsymbol{\beta}^\circ,\mathbf{u}^\circ)}{\partial\mathbf{u}\partial\boldsymbol{\beta}'} & \frac{\partial^2 l_1(\boldsymbol{\beta}^\circ,\mathbf{u}^\circ)}{\partial\mathbf{u}\partial\mathbf{u}'} \end{pmatrix}\begin{pmatrix} \boldsymbol{\beta}-\boldsymbol{\beta}^\circ \\ \mathbf{u}-\mathbf{u}^\circ \end{pmatrix} \qquad (1.15) \end{aligned}$$

After some straightforward algebra, we get

$$\begin{aligned} l_1(\mathbf{y}|\mathbf{u}) &\approx c + \frac{1}{2}\left(\boldsymbol{\beta}' - \boldsymbol{\beta}^{\circ t}, \mathbf{u}' - \mathbf{u}^{\circ t}\right)\begin{pmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{pmatrix}\left(\frac{\partial^2 l_1}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'}\right)(\mathbf{X}, \mathbf{Z})\begin{pmatrix} \boldsymbol{\beta} - \boldsymbol{\beta}^{\circ} \\ \mathbf{u} - \mathbf{u}^{\circ} \end{pmatrix} \\ &\approx c - \frac{1}{2}\left(\boldsymbol{\eta}^{\circ} - \boldsymbol{\eta}\right)' \mathbf{W}\left(\boldsymbol{\eta}^{\circ} - \boldsymbol{\eta}\right) \doteq \ell_1(\boldsymbol{\eta}^{\circ}|\mathbf{u}), \end{aligned}$$

where $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$, $\boldsymbol{\eta}^{\circ} = \mathbf{X}\boldsymbol{\beta}^{\circ} + \mathbf{Z}\mathbf{u}^{\circ}$ and

$$\mathbf{W} = -E\left[\frac{\partial^2 l_1(\mathbf{y}|\mathbf{u})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'}\right]\Bigg|_{\eta=\eta^{\circ}}. \tag{1.16}$$

As we have seen, $l_1(\mathbf{y}|\mathbf{u}) \approx \ell_1(\boldsymbol{\eta}^{\circ}|\mathbf{u})$. Then the marginal p.d.f. are also approximately equal, i.e. $l_1(\mathbf{y}) \approx \ell_1(\boldsymbol{\eta}^{\circ})$. This is to say, the values of $\varphi_1, \dots, \varphi_m$ maximizing $l_1(\mathbf{y})$ are approximately equal to those maximizing $\ell_1(\boldsymbol{\eta}^{\circ})$. Further $\ell_1(\boldsymbol{\eta}^{\circ})$ is obtained from $\ell_1(\boldsymbol{\eta}^{\circ}|\mathbf{u})$ and $l_2(\mathbf{u})$ (both are log-likelihoods of normal distributions). Then,

- $\ell_1(\boldsymbol{\eta}^{\circ})$ is the loglikelihood of $\boldsymbol{\eta}^{\circ} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}_y)$, with $\mathbf{V}_y = \mathbf{Z}'\mathbf{V}_u\mathbf{Z} + \mathbf{W}^{-1}$,
- It is assumed that $\boldsymbol{\eta}^{\circ}$ follows the model $\boldsymbol{\eta}^{\circ} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$, where $\mathbf{e} \sim N(\mathbf{0}, \mathbf{W}^{-1})$ and $\mathbf{u} \sim N(\mathbf{0}, \mathbf{V}_u)$ are independent.
- The values of $\varphi_1, \dots, \varphi_m$ that maximize $\ell_1(\boldsymbol{\eta}^{\circ})$ and $\ell_1(\boldsymbol{\eta}^{\circ}|\mathbf{u}) + l_2(\mathbf{u})$ are equal.

McGilchrist (1994) proposed to estimate $\varphi_1, \dots, \varphi_m$ by maximizing $\ell_1(\boldsymbol{\eta}^{\circ})$. Depending on the approach (ML or REML) when doing this, one gets the PQL-ML or the PQL-REML estimates.

The PQL-ML approach maximizes $\ell_1(\boldsymbol{\eta}^{\circ})$ to obtain estimates of $\varphi_1, \dots, \varphi_m$. For this sake, we assume the model

$$\boldsymbol{\eta}^{\circ} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{1.17}$$

where $\mathbf{e} \sim N(\mathbf{0}, \mathbf{W}^{-1})$ and $\mathbf{u} \sim N(\mathbf{0}, \mathbf{V}_u)$ are independent. Therefore $\boldsymbol{\eta}^{\circ} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}_y)$, with $\mathbf{V}_y = \mathbf{Z}\mathbf{V}_u\mathbf{Z}' + \mathbf{W}^{-1}$ and $\mathbf{V}_u = \text{diag}(\varphi_1\boldsymbol{\Sigma}_{u_1}, \dots, \varphi_m\boldsymbol{\Sigma}_{u_m})$. The log-likelihood of $\boldsymbol{\eta}^{\circ}$ under model (1.17) is

$$\ell_{ML}(\boldsymbol{\eta}^{\circ}) = -\frac{nq}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{V}_y| - \frac{1}{2}(\boldsymbol{\eta}^{\circ} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}_y{}^{-1}(\boldsymbol{\eta}^{\circ} - \mathbf{X}\boldsymbol{\beta}). \tag{1.18}$$

In practice, the PQL-ML estimates of $\varphi_1, \dots, \varphi_m$ are obtained through the Fisher-scoring algorithm.

On the other hand, PQL-REML approach obtains estimates of $\varphi_1, \dots, \varphi_m$ by applying the Fisher-scoring algorithm to the REML log-likelihood

$$l_{REML}(\boldsymbol{\eta}^{\circ}) = -\frac{1}{2}(qn - p)\log 2\pi - \frac{1}{2}\log|\mathbf{K}'\mathbf{V}_y\mathbf{K}| - \frac{1}{2}\boldsymbol{\eta}^{\circ t}\mathbf{P}\boldsymbol{\eta}^{\circ}, \tag{1.19}$$

where

$$\mathbf{P} = \mathbf{V}_y^{-1} - \mathbf{V}_y^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_y^{-1} \quad \text{and} \quad \mathbf{K} = \mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}.$$

## 1.8 References

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley, New York.

Andersen, E.B. (1997). *Introduction to the Statistical Analysis of Categorical Data*. Springer-Verlag, New York.

Arnold, S. (1981). *The Theory of Linear Models and Multivariate Analysis*. John Wiley, New York.

Booth, J.G. and Hobert, J.P (1999). Maximizing generalized linear mixed model likelihoods with and automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, **61**, 265285.

Breslow, N.E., Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 925.

Demidenko E. (2004). *Mixed models, Theory and Applications*. John Wiley, New York.

Dobson A.J. (1997). *An introduction to Generalized linear models*. Chapman & Hall, London. Reprinted from 1990.

Christensen R. (1990). *Log-Linear Models and Logistic Regression*. Springer-Verlag, New York.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York.

Goldstein, H. (1991). Nonlinear multilevel models, with applications to discrete response data. *Biometrika*, **78**, 4551

Goldstein, H. (2003). *Multilevel Statistical Models*. Arnold, London.

González-Manteiga, W., Lombarda, M.J., Molina, I., Morales, D. and Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under logistic mixed model. *Computational Statistics and Data Analysis*, **51**, 2720–2733.

Graybill, F.A. (1976). *Theory and Application of Linear Model*. Duxbury, Mass.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, **9**, 226–252.

Hocking, R.R. (1985). *The Analysis of Linear Models*. Brooks/Cole, Pacific Grove, California.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley, New York.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and their Applications*. Springer-Verlag, New York.

Lindsey, J.K. (2000). *Applying Generalized linear models*. Springer-Verlag, New York.

Lloyd, C.J. (1999). *Statistical Analysis of Categorical Data*. John Wiley, New York.

Longford, N.T. (1994). Logistic regression with random coefficients. *Computational Statistics and Data Analysis*, **17**, 115.

Longford, N.T. (1995). *Random coefficient models*. Clareton Press, London.

McCullogh, P. and Nelder, J.A. (1989). *Generalized Linear Models, 2nd Ed.* Chapman&Hall, London.

McCulloch, C.E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, **89**, 330335.

McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**, 162170.

McCullogh, P. and Searle, S.R. (2001). *Generalized, Linear and Mixed Models*. John Wiley, New York.

McGilchrist, C.A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society, Series B*, **56**, 6169.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370384.

Rencher, A.C. (2000). *Linear models in Statistics*. John Wiley, New York.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719727.

Searle, S.R., Casella, G. and McCullogh, P. (1992). *Variance Components*. John Wiley, New York.

Searle, S.R. (1997). *Linear Models*- Classic Edition, John Wiley, New York. Reprinted from 1971.

Seber, G.F.A. (1977). *Linear Regression Analysis*, John Wiley, New York.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**, 233243.

# Chapter 2

# Small Area EBLUP

## 2.1 Introduction

In small area estimation samples are drawn from a finite population, but estimations are required for subsets (called small areas or domains) where the effective sample sizes are too small to produce reliable direct estimates. An estimator of a small area parameter is called direct if it is calculated just with the sample data coming from the corresponding small area. Thus, the lack of sample data from the target small area affects seriously the accuracy of the direct estimators, and this fact has given rise to the development of new tools for obtaining more precise estimates. See a description of this theory in the monograph of Rao (2003), or in the reviews of Ghosh and Rao (1994), Rao (1999), Pfeffermann (2002) and more recently Jiang and Lahiri (2006). Mixed models increase the effective information used in the estimation process by linking all observations of the sample, and at the same time they can allow for between-area variation. Models of this kind have been used for instance in the United States to estimate per capita income for small areas (Fay and Herriot, 1979), for estimating census undercount (Ericksen and Kadane 1985, and by Dick 1995), and for estimating poor school-age children (National Research Council, 2000). It is worth to mention that "using these estimates, the U.S. Department of Education allocates annually over $7 billion of general funds to counties, and then the states distribute these funds among school districts" (Rao, 2003). The usage of these techniques is not restricted to socioeconomic data; an example in the field of agriculture is the work of Battese, Harter and Fuller (1988), who used a mixed linear model to estimate county crop areas.

Consider a linear parameter $\eta$, i.e. a parameter that is a linear combination of the values that the target variable takes in the population units. The *Best Linear Unbiased Predictor* (BLUP) of $\eta$ depends on unknown quantities; typically, variance components and/or correlations. When those quantities are replaced by suitable estimators, then the resulting predictor is called *Empirical* BLUP (EBLUP). Due to the estimation of the variance components, these predictors are not linear on the values of the target variable. Then, the exact *Mean Squared Error* (MSE) of an EBLUP cannot be analytically derived and this fact has given rise to the development of approximations based on Taylor expansions. First simplification of the MSE was given by Kackar and Harville (1981), assuming normality of the model errors and the random effects. In a second paper, Kackar and Harville (1984) provided an approximation of the mentioned

MSE and proposed an estimator based on it. Prasad and Rao (1990) gave a new approximation for models with block-diagonal covariance matrices. Under certain regularity assumptions for the model and the estimators of variance components, they showed that when the number of blocks $D$ tends to infinity, their approximation is of order $o(D^{-1})$. They also studied a new estimator of the MSE and gave the specific expressions of this estimator for Fay-Herriot, nested-error and random-coefficient models. The conditions imposed on the estimators of the variance components are satisfied by estimators obtained by the Henderson 3 method, but they cannot be verified for maximum likelihood estimators. Datta and Lahiri (2000) provided the analogue MSE estimator for general models with block-diagonal covariance matrices, when variance components are estimated by *Maximum Likelihood* (ML) or by *Residual Maximum Likelihood* (REML) methods. More recently, Das, Jiang and Rao (2004) studied the approximation of the MSE for a wider class of models, including ANOVA and longitudinal random effects models, when variance components are estimated by ML or REML. In Section 2.2 we introduce the closed-formula estimator of the MSE given by Prasad and Rao (1990) with the extension to REML estimators given by Das, Jiang and Rao (2004).

A more complex situation for MSE estimation appears when a generalized linear mixed model is assumed for the target variable. There is some interesting research done in this area using hierarchical Bayes methodology (see, e.g., Malec et al., 1997; Ghosh et al., 1998; Farrell et al., 1997). In the frequentist framework, Jiang and Lahiri (2001) proposed an empirical best predictor (in terms of the MSE) and obtained an approximation of the MSE correct up to order $o(D^{-1})$, where D is the number of small areas. Jiang (2003) extended the Jiang-Lahiri results to generalized linear mixed models. González-Manteiga et al. (2007) gave a simple (although not best) predictor, called generalized EBLUP (GEBLUP). They also gave an easy-to-apply closed formula estimator of the MSE of the GEBLUP when the GLMM is fitted by using the approach given in Section 2.2. This MSE estimator is described in Section 2.3.

Resampling represents a solution when some characteristic of the distribution (or the distribution itself) of a statistic is required, but its exact analytical expression is not available. This happens often when the statistic is not linear on the values of the target variable/s; for instance, in the case of the median; or when the statistic is defined under a semi-parametric or nonparametric setting. Even in some cases where large sample approximations are available, bootstrap might provide more accurate alternatives because of its second-order accuracy, usually not achieved by asymptotic methods. This property is mentioned in Efron and Tibshirani (1993), and proved by Hall (1992). For these reasons, we consider of interest using bootstrap methods, which are applicable for estimating the MSE under more general model assumptions and extensible to other types of small-area parameters and corresponding estimators, linear or not.

Some resampling methods for estimation of the MSE of empirical predictors can already be found in the literature. The jackknife methodology proposed by Jiang et al. (2002) provides estimators with bias of order $O(D^{-3/2})$. Pfeffermann and Tiller (2005) proposed a parametric and a nonparametric bootstrap methods for estimating the same quantity under state-space models. Recently, Hall and Maiti (2006a, 2006b) introduced a parametric and a matched-moment double-bootstrap algorithms, and González-Manteiga et al. (2007, 2008, 2009) applied bootstrap procedures to logistic and normal mixed models. In Section 2.4 a parametric bootstrap method is given, based on the works of González-Manteiga et al.

## 2.2 The EBLUP and its mean squared error

Let $\Omega = \{1, \dots, N\}$ be a finite population, $s \subset \Omega$ a sample of size $n \leq N$ drawn from $\Omega$ and $r = \Omega - s$ be the set of non sampled units. Let $\mathbf{y} = (y_1, \dots, y_N)'$ be the vector containing the values of the target variable for the population units and consider a decomposition of $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$ in the sample elements $\mathbf{y}_s$ and the non-sample elements $\mathbf{y}_r$. Let $\mathbf{a} = (\mathbf{a}_s', \mathbf{a}_r')'$ be a vector of known constants. We are interested in predicting the linear quantity

$$\eta = \mathbf{a}'\mathbf{y} = \mathbf{a}_s'\mathbf{y}_s + \mathbf{a}_r'\mathbf{y}_r,$$

where $\mathbf{y}$ follows model (1.1). Replacing $\mathbf{y}_r = \mathbf{X}_r\boldsymbol{\beta} + \mathbf{Z}_r\mathbf{u} + \mathbf{e}_r$, we obtain $\eta = \mathbf{a}_s'\mathbf{y}_s + \tau + \mathbf{a}_r'\mathbf{e}_r$, where $\tau = \mathbf{a}_r'(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{Z}_r\mathbf{u})$. The EBLUP of $\eta$ is $\hat{\eta}_E = \mathbf{a}_s'\mathbf{y}_s + \hat{\tau}_E$, where $\hat{\tau}_E$ is equal to (1.4) with the variance components $\boldsymbol{\varphi} = (\sigma^2, \varphi_1, \dots, \varphi_m)$ replaced by their estimates. Under the regularity assumptions stated in Das, Jiang and Rao (2004), a second order approximation to the mean squared error of $\hat{\eta}$ is

$$\begin{aligned}
MSE(\hat{\eta}_E) &\approx g_1(\boldsymbol{\varphi}) + g_2(\boldsymbol{\varphi}) + g_3(\boldsymbol{\varphi}) + g_4(\boldsymbol{\varphi}), \\
g_1(\boldsymbol{\varphi}) &= \mathbf{a}_r'\mathbf{Z}_r\mathbf{T}_s\mathbf{Z}_r'\mathbf{a}_r. \\
g_2(\boldsymbol{\varphi}) &= [\mathbf{a}_r'\mathbf{X}_r - \mathbf{a}_r'\mathbf{Z}_r\mathbf{T}_s\mathbf{Z}_s'\mathbf{V}_{es}^{-1}\mathbf{X}_s]\mathbf{Q}_s[\mathbf{X}_r'\mathbf{a}_r - \mathbf{X}_s'\mathbf{V}_{es}^{-1}\mathbf{Z}_s\mathbf{T}_s\mathbf{Z}_r'\mathbf{a}_r], \\
g_3(\boldsymbol{\varphi}) &= \text{tr}\left\{(\nabla\mathbf{b}')\mathbf{V}_s(\nabla\mathbf{b}')'E\left[(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})'\right]\right\}, \\
g_4(\boldsymbol{\varphi}) &= \mathbf{a}_r'\mathbf{V}_{er}\mathbf{a}_r,
\end{aligned}$$

where $\mathbf{Q}_s = (\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}$, $\mathbf{b}' = (b_1, \dots, b_n) = \mathbf{a}_r'\mathbf{Z}_r\mathbf{V}_u\mathbf{Z}_s'\mathbf{V}_s^{-1}$, $\mathbf{T}_s = \mathbf{V}_u - \mathbf{V}_u\mathbf{Z}_s'\mathbf{V}_u^{-1}\mathbf{Z}_s\mathbf{V}_u$,

$$\varphi_0 = \sigma^2, \quad \frac{\partial\mathbf{b}'}{\partial\varphi_j} = \left(\frac{\partial b_1}{\partial\varphi_j}, \dots, \frac{\partial b_n}{\partial\varphi_j}\right), \quad \nabla\mathbf{b}' = \begin{pmatrix} \frac{\partial\mathbf{b}'}{\partial\varphi_0} \\ \frac{\partial\mathbf{b}'}{\partial\varphi_1} \\ \vdots \\ \frac{\partial\mathbf{b}'}{\partial\varphi_m} \end{pmatrix} = \begin{pmatrix} \frac{\partial b_1}{\partial\varphi_0} & \cdots & \frac{\partial b_n}{\partial\varphi_0} \\ \frac{\partial b_1}{\partial\varphi_1} & \cdots & \frac{\partial b_n}{\partial\varphi_1} \\ \vdots & \cdots & \vdots \\ \frac{\partial b_1}{\partial\varphi_m} & \cdots & \frac{\partial b_n}{\partial\varphi_m} \end{pmatrix}_{(m+1)\times n}.$$

A closed-formula estimator of $MSE(\hat{\eta}_E)$ can be obtained from the results of Prasad and Rao (1990) or Das, Jiang and Rao (2001). When $\hat{\boldsymbol{\varphi}}$ is unbiased or approximately unbiased (Henderson 3 and REML methods), this estimator is given by

$$mse(\hat{\eta}_E) = g_1(\hat{\boldsymbol{\varphi}}) + g_2(\hat{\boldsymbol{\varphi}}) + 2g_3(\hat{\boldsymbol{\varphi}}) + g_4(\hat{\boldsymbol{\varphi}}), \tag{2.1}$$

and when $\hat{\boldsymbol{\varphi}}$ is estimated by ML, the MSE estimator is

$$mse(\hat{\eta}_E) = g_1(\hat{\boldsymbol{\varphi}}) + g_2(\hat{\boldsymbol{\varphi}}) + 2g_3(\hat{\boldsymbol{\varphi}}) + g_4(\hat{\boldsymbol{\varphi}}) - \mathbf{b}_{\hat{\boldsymbol{\varphi}}}(\boldsymbol{\varphi})\nabla g_1(\boldsymbol{\varphi}). \tag{2.2}$$

The term $\mathbf{b}_{\hat{\boldsymbol{\varphi}}}(\boldsymbol{\varphi})$ is the approximate bias of $\hat{\boldsymbol{\varphi}}$. When $\mathbf{V}$ satisfies

$$\mathbf{V} = diag(\mathbf{V}_1, \dots, \mathbf{V}_m), \quad \text{with } \mathbf{V}_i = \mathbf{Z}_i\mathbf{V}_{ui}\mathbf{Z}_i' + \mathbf{V}_{ei}, \quad i = 1, \dots, m,$$

then model (1.1) can be decomposed in $m$ submodels

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i, \quad i = 1, \dots, m, \tag{2.3}$$

where $\mathbf{y} = (\mathbf{y}_1', \dots, \mathbf{y}_m')'$, $\mathbf{X} = (\mathbf{X}_1', \dots, \mathbf{X}_m')'$, $\mathbf{Z} = diag(\mathbf{Z}_1, \dots, \mathbf{Z}_m)'$, $\mathbf{u} = (\mathbf{u}_1', \dots, \mathbf{u}_m')'$, $\mathbf{e} = (\mathbf{e}_1', \dots, \mathbf{e}_m')'$, $\mathbf{X}_i$ is $n_i \times p$, $\mathbf{Z}_i$ is $n_i \times q_i$, $\mathbf{y}_i$ is $n_i \times 1$, $n = \sum_{i=1}^m n_i$ and $q = \sum_{i=1}^m q_i$. Then, an approximation of the bias (see Rao (2003)) of the ML estimator $\hat{\boldsymbol{\varphi}}$ is

$$\mathbf{b}_{\hat{\boldsymbol{\varphi}}}(\boldsymbol{\varphi}) = \frac{1}{2m}\left\{\mathcal{I}^{-1}(\boldsymbol{\varphi}) \underset{1\le j\le m}{\text{col}} \left[\text{trace}\left[\sum_{i=1}^m (\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\left(\sum_{i=1}^m \mathbf{X}_i'\mathbf{V}_i^{(j)}\mathbf{X}_i\right)\right]\right]\right\},$$

where $\underset{1\le j\le m}{\text{col}}[a_j]$ is a column vector with components $a_j$, $j = 1, \dots, m$, and

$$\mathbf{V}_i^{(j)} = \frac{\partial \mathbf{V}_i^{-1}}{\partial \boldsymbol{\varphi}_j} = -\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\varphi}_j}\mathbf{V}_i^{-1},\ j = 1, \dots, m,$$

and $\mathcal{I}(\boldsymbol{\varphi})$ is the Fisher information matrix, whose elements are

$$\mathcal{I}_{jk}(\boldsymbol{\varphi}) = \frac{1}{2}\sum_{i=1}^m \text{trace}\left[\left(\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\varphi}_j}\right)\left(\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \boldsymbol{\varphi}_k}\right)\right],\ j, k = 1, \dots, m.$$

## 2.3 The GEBLUP and its mean squared error

In this section we assume that $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$ follows the model (1.13)-(1.14). We are interested in predicting the linear quantity

$$\boldsymbol{\delta} = \mathbf{a}'\mathbf{y} = \mathbf{a}_s'\mathbf{y}_s + \mathbf{a}_r'\mathbf{y}_r.$$

The model equation for sample data is $\boldsymbol{\mu}(\boldsymbol{\theta}_s) = \mathbf{h}_s(\boldsymbol{\eta}_s)$, where $\boldsymbol{\eta}_s = \mathbf{X}_s\boldsymbol{\beta} + \mathbf{Z}_s\mathbf{u}$ and $\boldsymbol{\theta}_s$ is the vector of sampled natural parameters. Fitting this model we obtain the predictor $\hat{\boldsymbol{\eta}}_r = \mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\hat{\mathbf{u}}$ and then we can take $\mathbf{h}_r(\hat{\boldsymbol{\eta}}_r)$ as a predictor of $\mathbf{y}_r$. Thus a predictor of $\boldsymbol{\delta}$, called generalized EBLUP (or GEBLUP), is

$$\hat{\boldsymbol{\delta}} = \mathbf{a}_s'\mathbf{y}_s + \mathbf{a}_r'\mathbf{h}_r(\hat{\boldsymbol{\eta}}_r),$$

and the prediction error is $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta} = \mathbf{a}_r'[\mathbf{h}_r(\hat{\boldsymbol{\eta}}_r) - \mathbf{y}_r]$. The MSE of $\hat{\boldsymbol{\delta}}$ is given by

$$MSE(\hat{\boldsymbol{\delta}}) = E[(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})'] = E\left[\mathbf{a}_r\left\{\mathbf{h}_r(\hat{\boldsymbol{\eta}}_r) - \mathbf{y}_r\right\}\left\{\mathbf{h}(\hat{\boldsymbol{\eta}}_r) - \mathbf{y}_r\right\}'\mathbf{a}_r\right].$$

Let us define $\mathbf{V}_{Gr} = E[(\mathbf{y}_r - \mathbf{h}_r(\boldsymbol{\eta}_r))(\mathbf{y}_r - \mathbf{h}_r(\boldsymbol{\eta}_r))']$. Then

$$\begin{aligned} MSE(\hat{\boldsymbol{\delta}}) &= E\left[\mathbf{a}_r'\left\{\mathbf{h}_r(\hat{\boldsymbol{\eta}}_r) - \mathbf{h}_r(\boldsymbol{\eta}_r)\right\}\left\{\mathbf{h}_r(\hat{\boldsymbol{\eta}}_r) - \mathbf{h}_r(\boldsymbol{\eta}_r)\right\}'\mathbf{a}_r\right] + \mathbf{a}_r'\mathbf{V}_{Gr}\mathbf{a}_r \\ &+ \mathbf{a}_r'E\left[\left\{\mathbf{h}_r(\hat{\boldsymbol{\eta}}_r) - \mathbf{h}_r(\boldsymbol{\eta}_r)\right\}\left\{\mathbf{h}_r(\boldsymbol{\eta}_r) - \mathbf{y}_r\right\}'\right]\mathbf{a}_r \\ &+ \mathbf{a}_r'E\left[\left\{\mathbf{h}_r(\boldsymbol{\eta}_r) - \mathbf{y}_r\right\}\left\{\mathbf{h}_r(\hat{\boldsymbol{\eta}}_r) - \mathbf{h}_r(\boldsymbol{\eta}_r)\right\}'\right]\mathbf{a}_r = S_1 + S_2 + S_3 + S_4. \end{aligned}$$

A Taylor linearization of $\mathbf{h}_r(\hat{\boldsymbol{\eta}}_r)$ yields to the approximation of the first term; i.e.

$$S_1 = E\left[\mathbf{a}_r'\left\{\mathbf{h}_r(\hat{\boldsymbol{\eta}}_r) - \mathbf{h}_r(\boldsymbol{\eta}_r)\right\}\left\{\mathbf{h}(\hat{\boldsymbol{\eta}}_r) - \mathbf{h}_r(\boldsymbol{\eta}_r)\right\}'\mathbf{a}_r\right] \approx E\left[\mathbf{a}_r'\mathbf{H}_r(\hat{\boldsymbol{\eta}}_r - \boldsymbol{\eta}_r)(\hat{\boldsymbol{\eta}}_r - \boldsymbol{\eta}_r)'\mathbf{H}_r'\mathbf{a}_r\right].$$

Let us define the parameter $\tau_G = \boldsymbol{\alpha}_r'\boldsymbol{\eta}_r = \boldsymbol{\alpha}_r'(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{Z}_r\mathbf{u})$, the vector $\boldsymbol{\alpha}_r' = \mathbf{a}_r'\mathbf{H}_r$ and the predictor $\hat{\tau}_{GE} = \boldsymbol{\alpha}_r'\hat{\boldsymbol{\eta}}_r = \boldsymbol{\alpha}_r'(\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\hat{\mathbf{u}})$, where

$$\mathbf{H}_r = \text{diag}(\dot{h}_1, \ldots, \dot{h}_{N_r}) \quad \text{and} \quad \dot{h}_j = \frac{\partial h_j(\eta_j)}{\partial \eta_j},\ j = 1, \ldots, N_r.$$

Then

$$S_1 \approx E[(\hat{\tau}_{GE} - \tau_G)(\hat{\tau}_{GE} - \tau_G)'] = MSE(\hat{\tau}_{GE}).$$

In the case of using the PQL-ML method to estimate $\boldsymbol{\beta}$ and $\varphi_i$, $i = 1, \ldots, m$, the GLMM is approximated by the linear (normal) mixed model with log-likelihood function (1.18). Alternatively we can maximize the log-likelihood (1.19), to apply the PQL-REML method. These approximations to the normal model allow to apply the results in Kackar and Harville (1984) to approximate the last two terms of $MSE(\hat{\boldsymbol{\delta}})$, $S_3$ and $S_4$, to zero. Further, under the PQL-REML approximation to a normal model we can apply the results of Section 2.2. In this way, we obtain

$$MSE(\hat{\boldsymbol{\delta}}) \approx \mathcal{G}_1(\boldsymbol{\varphi}) + \mathcal{G}_2(\boldsymbol{\varphi}) + \mathcal{G}_3(\boldsymbol{\varphi}) + \mathcal{G}_4(\boldsymbol{\varphi}),$$

where

$$\begin{aligned}
\mathcal{G}_1(\boldsymbol{\varphi}) &= \sigma^2\mathbf{a}_r'\mathbf{H}_r\mathbf{Z}_r\mathbf{T}_s\mathbf{Z}_r'\mathbf{H}_r'\mathbf{a}_r, \\
\mathcal{G}_2(\boldsymbol{\varphi}) &= \sigma^2\mathbf{a}_r'\mathbf{H}_r\left[\mathbf{X}_r - \mathbf{Z}_r\mathbf{T}_s\mathbf{Z}_s'\mathbf{W}_s\mathbf{X}_s\right]\mathbf{Q}_s\left[\mathbf{X}_r' - \mathbf{X}_s'\mathbf{W}_s\mathbf{Z}_s\mathbf{T}_s\mathbf{Z}_r'\right]\mathbf{H}_r'\mathbf{a}_r, \\
\mathcal{G}_3(\boldsymbol{\varphi}) &= \text{tr}\left\{(\nabla\mathbf{b}')\mathbf{V}_s(\nabla\mathbf{b}')'E\left[(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})'\right]\right\}, \\
\mathcal{G}_4(\boldsymbol{\varphi}) &= \mathbf{a}_r'\mathbf{V}_{Gr}\mathbf{a}_r' \approx \mathbf{a}_r'\text{cov}(\mathbf{y}_r|\mathbf{u})\mathbf{a}_r,
\end{aligned}$$

$\mathbf{T}_s = \left(\mathbf{V}_u^{-1} + \mathbf{Z}_s'\mathbf{W}_s\mathbf{Z}_s\right)^{-1}$, $\mathbf{Q}_s = (\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}$, $\mathbf{V}_s = \mathbf{Z}_s\mathbf{V}_u\mathbf{Z}_s' + \mathbf{W}_s^{-1}$, $\mathbf{b}' = (b_1, \ldots, b_n) = \mathbf{a}_r'\mathbf{H}_r\mathbf{Z}_r\mathbf{V}_u\mathbf{Z}_s'\mathbf{V}_s^{-1}$, $\nabla\mathbf{b}'$ is defined in Section 2.2 and $\mathbf{W}_s$ is defined in (1.16). Finally, a closed-formula estimator of $MSE(\hat{\boldsymbol{\delta}})$ is

$$mse(\hat{\boldsymbol{\delta}}) = \mathcal{G}_1(\hat{\boldsymbol{\varphi}}) + \mathcal{G}_2(\hat{\boldsymbol{\varphi}}) + 2\mathcal{G}_3(\hat{\boldsymbol{\varphi}}) + \mathcal{G}_4(\hat{\boldsymbol{\varphi}}).$$

## 2.4 Bootstrap estimation of the MSE

This section introduces a parametric bootstrap method to estimate the mean squared error of estimators of finite population quantities. Parametric bootstrap procedures rely on the assumption that the vector of observations $\mathbf{y}$ follows a given parametric distribution. Here we describe the bootstrap method considering that $\mathbf{y}$ follows a nested error regression model, but extension to other types of models is straightforward. Thus, assume the superpopulation model

$$\xi : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{2.4}$$

where $\mathbf{y}_{N\times 1}$ is the vector of response variables, $\boldsymbol{\beta}_{p\times 1}$ is the vector of the coefficients of the explanatory variables or fixed effects, $\mathbf{u}_{D\times 1} \sim N(\mathbf{0}, \mathbf{V}_u)$ is the vector of random effects with covariance matrix

$\mathbf{V}_u = \sigma_u^2 \mathbf{I}_D$, where $\mathbf{I}_D = \text{diag}(1, \ldots, 1)_{D \times D}$, $\mathbf{X}_{N \times p}$ is an incidence matrix with known elements, $\mathbf{Z} = \text{diag}(\mathbf{1}_{N_d}; d = 1, ..., D)$ is a block-diagonal matrix with $\mathbf{1}_a = (1, ..., 1)'_{a \times 1}$, and $\mathbf{e}_{N \times 1} \sim N(\mathbf{0}, \mathbf{V}_e)$, with $\mathbf{V}_e = \sigma_e^2 \mathbf{W}^{-1}$, is the vector of random perturbations, independent of $\mathbf{u}$.

Let $\Omega$ be a finite population generated by the superpopulation model $\xi$. Let $\eta = \eta(\xi)$ be a linear function of the fixed and random effects of $\xi$. Let $s$ be a sample extracted from $\Omega$ using a certain sampling design. The following steps describe a bootstrap procedure designed for estimating the mean squared error of the EBLUP $\hat{\eta}_E$, $MSE(\hat{\eta}_E)$:

*Step 1.* From the sample $s$, calculate estimators $\hat{\sigma}_u^2$, $\hat{\sigma}_e^2$ and $\hat{\boldsymbol{\beta}}_E$ of $\sigma_u^2$, $\sigma_e^2$ and $\boldsymbol{\beta}$ respectively.

*Step 2.* Generate $D$ independent copies of a variable $T_1$ with $E(T_1) = 0$ and $Var(T_1) = 1$. Construct the vector $\mathbf{u}^* = \hat{\sigma}_u T_1$ of size $D$, with mean $\mathbf{0}_D$ and covariance matrix $\hat{\mathbf{V}}_u = \hat{\sigma}_u^2 \mathbf{I}_D$.

*Step 3.* Generate $N = \sum_{d=1}^{D} N_d$ independent copies of a random variable $T_2$ with $E(T_2) = 0$ and $Var(T_2) = 1$, independent of $T_1$. Construct the vector $\mathbf{e}^* = \hat{\sigma}_e \mathbf{W}^{-1/2} T_2$ of size $N$, with mean $\mathbf{0}_N$ and covariance matrix $\hat{\mathbf{V}}_e = \hat{\sigma}_e^2 \mathbf{W}^{-1}$.

*Step 4.* With the elements of the incidence matrices $\mathbf{X}$ and $\mathbf{Z}$ known for each unit in the population, construct the bootstrap superpopulation model

$$\xi^* : \mathbf{y}^* = \mathbf{X}\hat{\boldsymbol{\beta}}_E + \mathbf{Z}\mathbf{u}^* + \mathbf{e}^*. \tag{2.5}$$

For the model $\xi^*$, the parameter $\eta^* = \eta(\xi^*)$ is defined by analogy to the parameter $\eta = \eta(\xi)$, but as function of the components of the bootstrap model $\xi^*$. Further, $s^*$ denotes a sample generated from $\xi^*$, $\hat{\eta}_B^*$ the BLUP and $\hat{\eta}_E^*$ the EBLUP of $\eta^*$ constructed from $s^*$ in the same way as $\hat{\eta}_B$ and $\hat{\eta}_E$ were obtained from $s$. Under model $\xi^*$, given the initial sample $s$, the mean squared error of $\hat{\eta}_E^*$ is denoted by $MSE_*(\hat{\eta}_E^*)$. Thus, for estimating the MSE of $\hat{\eta}_E$, our first proposal is the bootstrap mean squared error $\text{MSE}_{*1}(\hat{\eta}_E) = MSE_*(\hat{\eta}_E^*)$. In practice, this estimator is approximated via Monte Carlo in the following way:

*Step 5.* Given the bootstrap superpopulation model $\xi^*$, generate independent and identically distributed samples $s^{*(b)}$, $b = 1, ..., B$, containing the same units as $s$, and calculate bootstrap parameters $\eta^{*(b)} = \eta(\xi^{*(b)})$, $b = 1, ..., B$.

*Step 6.* From each sample $s^{*(b)}$, calculate the bootstrap EBLUP $\hat{\eta}_E^{*(b)}$ of $\eta^{*(b)}$. The Monte Carlo approximation of the bootstrap estimator $\text{MSE}_{*1}(\hat{\eta}_E^*)$ is given by

$$\text{mse}_{*1}(\hat{\eta}_E^*) = B^{-1} \sum_{b=1}^{B} (\hat{\eta}_E^{*(b)} - \eta^{*(b)})^2. \tag{2.6}$$

**Remark 2.4.1.** The bootstrap procedure described here is applicable to other parameters $\eta(\xi)$ linear or not, and their corresponding predictors $\hat{\eta}$.

**Remark 2.4.2.** In the case of LMMs it is possible to obtain a less biased estimator of the MSE by correcting for the bias of $g_1(\hat{\boldsymbol{\theta}}) + g_2(\hat{\boldsymbol{\theta}}) + g_4(\hat{\boldsymbol{\theta}})$ (cf. Section 2.2). Following Pfefferman and Tiller (2005), a bias-corrected bootstrap estimator is

$$\text{MSE}_{*2}(\hat{\eta}_E^*) = 2[g_1(\hat{\boldsymbol{\theta}}) + g_2(\hat{\boldsymbol{\theta}}) + g_4(\hat{\boldsymbol{\theta}})] - E_*[g_1(\hat{\boldsymbol{\theta}}^*) + g_2(\hat{\boldsymbol{\theta}}^*) + g_4(\hat{\boldsymbol{\theta}}^*)] + E_*[(\hat{\tau}_E^* - \hat{\tau}_B^*)^2]. \quad (2.7)$$

A Monte Carlo approximation of (2.7), denoted $\text{mse}_{*2}(\hat{\eta}_E^*)$, is obtained similarly as (2.6).

## 2.5 References

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28-36.

Datta, G.S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, **10**, 613-627.

Das, K., Jiang, J. and Rao, J. N. K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, **32**, 818-840.

Dick, P. (1995). Modelling net undercoverage in the 1991 Canadian census. *Survey Methodology*, **21**, 4554.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

Ericksen, E. P. and Kadane, J. B. (1985). Estimating the population in census year: 1980 and beyond (with discussion). *Journal of the American Statistical Association*, **80**, 98-131.

Farrell, P.J., MacGibbon, B., Tomberlin, T.J. (1997). Bootstrap adjustments for empirical Bayes interval estimates of small-area proportions. *Canadian Journal of Statistics*, **25**, 7589.

Fay, R. E. and Herriot, R. A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.

Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, **9**, 55-93.

Ghosh, M., Natarajan, K., Stroud, T.W.F., Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, **93**, 273282.

González-Manteiga, W., Lombarda, M.J., Molina, I., Morales, D. and Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under logistic mixed model. *Computational Statistics and Data Analysis*, **51**, 2720-2733.

González-Manteiga, W., Lombarda, M.J., Molina, I., Morales, D. and Santamaría, L. (2008). Bootstrap Mean Squared Error of a Small-Area EBLUP. *Journal of Statistical Computation and Simulation*, **78**, 5, 443-462.

González-Manteiga, W., Lombarda, M.J., Molina, I., Morales, D. and Santamaría, L. (2008). Analytic and Bootstrap Approximations of Prediction Errors under a Multivariate Fay-Herriot model. *Computational Statistics & Data Analysis*, **52**, 5242-5252.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *Annals of Statistics*, **34**, 17331750.

Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small-area prediction. *Journal of the Royal Statistical Society, Series B*, **68**, 221238.

Jiang, J., Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*. **53**, 217243.

Jiang, J., Lahiri, P. and Wan, S. (2002). A united jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics*, **30**, 17821810.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference*, **111**, 117127.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, **15**, 1-96.

Kackar, R. and Harville, D. A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics-Theory and Methods*, **10**, 1249-1261.

Kackar, R. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, **79**, 853-862.

Malec, D., Sedransk, J., Moriarity, C.L., LeClere, F.B. (1997). Small area inference for binary variables in the national health interview survey. *Journal of the American Statistical Association*, **92**, 815826.

Pfeffermann, D. (2002). Small Area Estimation - New Developments and Directions. *International Statistical Review*, **70**, 1, 125-143.

Pfefferman, D. and Tiller, R. (2005). Bootstrap approximation to prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, **26**, 893916.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163-171.

Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, **25**, 175-186.

Rao, J.N.K. (2003). *Small Area Estimation*. John Wiley.

# Chapter 3

# Time Models

## 3.1 introduction

Mixed models that borrow strength across time are well suited for the analysis of longitudinal data, where each time series constitutes an individual curve. Diggle et al. (2005) describe statistical models and methods for the analysis of longitudinal data and they show that mixed models play a fundamental role. However main applications of these models have been to biological and health sciences.

As pointed out in Pfeffermann (2002) a typical time series model fitted to survey data consists of two parts: a model fitted to the population quantity of interest and a model fitted to the sampling errors. He considered a general class of state-space models for a single area $d$ with the index $t$ designing time that was previously introduced by Pfeffermann and Burck (1990). This model is

$$\begin{aligned} y_{dt} &= \theta_{dt} + e_{dt} = \mathbf{x}_{dt}\boldsymbol{\beta}_{dt} + e_{dt}, \\ \boldsymbol{\beta}_{dt} &= \mathbf{T}_t\boldsymbol{\beta}_{d,t-1} + \mathbf{u}_{dt}, \quad e_{dt} \sim ARMA(a,b), \end{aligned} \tag{3.1}$$

where $\boldsymbol{\beta}_{dt}$ $(p \times 1)$ is a random state vector, $\mathbf{T}_t$ $(p \times p)$ is a fixed transition matrix and $e_{dt}$ and $\mathbf{u}_{dt} = (u_{dt1}, \ldots, u_{dtp})'$ are independent random errors with $Var(e_{dt}) = \sigma^2$ and $Var(\mathbf{u}_{dt}) = \mathbf{Q}$ respectively. It is also assumed that $E(\mathbf{u}_{dt}\,\mathbf{u}'_{d,t-j}) = 0$ for $j > 0$. In model (3.1) $y_{dt}$ is the direct sample estimate for area $d$ at time $t$ and $\theta_{dt} = \mathbf{x}_{dt}\boldsymbol{\beta}_{dt}$ is the target quantity, modeled as a linear combination of known concomitant variables with random coefficients, and $e_{dt} = y_{dt} - \theta_{dt}$ is the sampling error. The notation $ARMA(a,b)$ defines an Auto-Regressive Moving Average model of order $(a,b)$, as in Box and Jenkins (1976). Model (3.1) accounts for the time series relationships between the true area quantities via the model postulated for the state vectors and for the autocorrelations between the sampling errors. Model (3.1) contains as unknown hyper-parameters the variances and covariances appearing in $\mathbf{Q}$, the parameters of the ARMA model of the sampling error and possibly also some of the elements of the transition matrices $\mathbf{T}_t$. Because of possible identification problems and in order to simplify the maximization of the likelihood, it is customary to estimate the ARMA parameters based on external estimates of the variance and autocorrelations of the sampling error. Pfeffermann et al. (1998) developed a simple method to estimate the sampling error autocorrelations for rotating panel sampling designs. The remaining parameters are estimated with the values maximizing the likelihood when fixing the ARMA parameters at their estimated values. The likelihood function is conveniently obtained by using the Kalman filter, which for known hyper-parameters yields the BLUP of the state vector and the corresponding prediction error variance-covariance matrix of every time $t$. See Harvey (1989) for more details.

In the field of small area estimation, data are often available for many small areas simultaneously, although possibly for only a few time points. In such cases, it is desired to borrow information both cross-sectionally and over time. A way of accounting for cross-sectional relationships under model (3.1) is by allowing non null correlations among the components of the error terms $\mathbf{u}_{dt}$. Pfeffermann and Burck (1990) derive the explicit expression of the small area predictor for the case in which the state vector is a random walk and the sampling errors are uncorrelated.

Rao and Yu (1994) gave a simple way of borrowing information cross-sectionally and over time by introducing a model containing both contemporary random effects and time varying effects. They proposed the extension of the basic Fay Herriot model

$$y_{dt} = \mathbf{x}_{dt}\boldsymbol{\beta} + v_d + u_{dt} + e_{dt}, \quad d = 1, \ldots, D, \quad t = 1, \ldots, T, \tag{3.2}$$

where $y_{dt}$ is a direct estimator of the indicator of interest and $\mathbf{x}_{dt}$ is a vector containing the aggregated (population) values of $p$ auxiliary variables. The index $d$ is used for domains and the index $t$ for time instants. They assume that $v_1, \ldots, v_D$ are i.i.d. normal, $(u_{d1}, \ldots, u_{dT})$'s follow i.i.d. AR(1) processes (i.e. they follow autoregressive processes of order 1), $e_{11}, \ldots, e_{DT}$ are i.i.d. normal, and the $v_d$'s, the $(u_{d1}, \ldots, u_{dT})$'s and the $e_{dt}$'s are independent.

Ghosh, Nangia and Kim (1996) proposed a slightly more complicated time correlated area level model to estimate the median income of four-person families for the fifty American states and the district of Columbia. You and Rao (2000) and Datta, Lahiri and Maiti (2002) used the Rao-Yu model, but replacing the AR(1) process by a random walk model. Datta, Lahiri, Maiti and Lu (1999) considered a similar model but added extra terms to the linking models to reflect seasonal variation in their application. They applied their model to estimate monthly unemployment rates for the nine American states and the district of Columbia. You, Rao and Gambino (2001) applied the the Rao-Yu model to estimate monthly unemployment rates for census metropolitan areas in Canada. All these models are formulated at the area level. Sections 3.4 and 3.5 introduces the area-level time linear mixed models to be developed in the SAMPLE project. They are related to the model of Rao and Yu (1994) in the sense that only $u_{dt}$ is considered to take into account the area-by-time variability through specific random effects.

Regarding applications, most time models used for small area estimation have been formulated at the area level. Unit-level models needs more requirements (e.g. the same data at the the unit level and at the aggregated level and with the same variable definitions) and more sophisticated software. The natural extension of (3.2) to a unit-level model is

$$y_{dtj} = \mathbf{x}_{dtj}\boldsymbol{\beta} + u_{1,d} + u_{2,t} + w_{dtj}^{-1/2} e_{dtj}, \quad d = 1, \ldots, D, t = 1, \ldots, m_d, j = 1, \ldots, n_{dt}, \tag{3.3}$$

where $y_{dtj}$ is the characteristic of interest for unit $j$, time instant $t$ and area $d$, $\mathbf{x}_{dtj}$ is a vector containing the values of $p$ auxiliary variables and $w_{dtj}$ is a heteroscedasticity weight. This model was used in the EURAREA project assuming that area effects $u_{1,1}, \ldots, u_{1,D}$ are i.i.d. normal, time effects $u_{2,1}, \ldots, u_{2,m_D}$

are either i.i.d. normal or follows an AR(1) process, errors $e_{11},\dots,e_{Dm_D}$ are i.i.d. normal, and the three sets of random terms in the model (the $u_{1,d}$'s, the $u_{2,t}$'s and the $e_{dtj}$'s) are independent. Furthermore the EURAREA team developed a SAS-IML code to fit these models. See the project reports in http://www.statistics.gov.uk/eurarea/download.asp. Later Fabrizi, Ferrante and Pacei (2007) considered a linear mixed model with the same equation as (3.3), but assuming that area and time effects are i.i.d. normal, errors follow an AR(1) process and they are all mutually independent.

Stukel and Rao (1997, 1999) studied the two-fold nested error regression model

$$y_{dtj} = \mathbf{x}_{dtj}\boldsymbol{\beta} + u_{1,d} + u_{2,dt} + w_{dtj}^{-1/2}e_{dtj}, \quad d=1,\dots,D, t=1,\dots,m_d, j=1,\dots,n_{dt}, \tag{3.4}$$

where the area effects $u_{1,d}$'s are i.i.d. $N(0,\sigma_1^2)$, the time by area effects $u_{2,dt}$'s are $N(0,\sigma_2^2)$, the errors $e_{dtj}$'s are i.i.d. $N(0,\sigma_0^2)$, and the $u_{1,d}$'s, the $u_{2,t}$'s and the $e_{dtj}$'s are independent. Datta and Ghosh (1991) and Pfeffermann and Barnard (1991) used the two-fold model for the special case of cluster-specific covariates, i.e. when $\mathbf{x}_{dtj} = \mathbf{x}_{dt} = (x_{dt1},\dots,x_{dtp})'$ for all $j$. Ghosh and Lahiri (1988) studied the case $\mathbf{x}'_{dtj}\boldsymbol{\beta} = \beta$ of no auxiliary information.

Sections 3.2 and 3.3 introduce the unit-level time linear mixed models to be developed in the SAMPLE project. They are more general than the model developed in the EURAREA project in the sense that $u_{2,t}$ is substituted by $u_{2,dt}$, as in Stukel and Rao (1997, 1999). In this way we allow the time effects to vary across areas. Finally, Sections 3.6 and 3.7 are devoted to unit-level GLMM models.

## 3.2 Individual-level linear model with correlated time effects

Consider a particular case of the linear mixed model (1.5) with two nested random factors, the first one with $D$ levels and, for each level $d$ ($d=1,\dots,D$) the second one with $m_d$ sublevels. More concretely, assume that

$$y_{dtj} = \mathbf{x}_{dtj}\boldsymbol{\beta} + u_{1,d} + u_{2,dt} + w_{dtj}^{-1/2}e_{dtj}, \quad d=1,\dots,D, t=1,\dots,m_d, j=1,\dots,n_{dt}, \tag{3.5}$$

where $y_{dtj}$ is the characteristic of interest for unit $j$, time instant $t$ and area $d$, $\mathbf{x}_{dtj}$ is a vector containing the values of $p$ auxiliary variables and $w_{dtj}$ is a heteroscedasticity weight. To complete the definition of model (3.5) we assume that the random effects $u_{1,d}$'s are i.i.d. $N(0,\sigma_1^2)$, the vectors $(u_{2,d1},\dots,u_{2,dm_d})$, $d=1,\dots,D$, follow i.i.d. AR(1) processes with variance and auto-correlation parameters $\sigma_2^2$ and $\rho$ respectively, the errors $e_{dtj}$'s are i.i.d. $N(0,\sigma_0^2)$, and the $u_{1,d}$'s, the $u_{2,dt}$'s and the $e_{dtj}$'s are independent.

Model (3.5) has a regression component that incorporates the auxiliary information. For the variability not explained by the $x$-variables, two random factors are considered. The first random factor models the between-areas variability and the second one models the time variability within each given area. The components $u_{1,1},\dots,u_{1,D}$ are assumed independent. The generalization to the dependent case yields to time-space linear models.

Model (3.5) can be expressed in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{W}^{-1/2}\mathbf{e}, \tag{3.6}$$

where $\mathbf{u}_1 = \mathbf{u}_{1,D\times 1} \sim \mathcal{N}(\mathbf{0}, \sigma_1^2 \mathbf{I}_D)$, $\mathbf{u}_2 = \mathbf{u}_{2,M\times 1} \sim \mathcal{N}(\mathbf{0}, \sigma_2^2 \Omega(\rho))$ and $\mathbf{e} = \mathbf{e}_{n\times 1} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_n)$ are independent and contain the $u_{1,d}$'s, the $u_{2,dt}$'s and the $e_{dtj}$'s respectively, $\mathbf{y} = \mathbf{y}_{n\times 1}$ and $\mathbf{X} = \mathbf{X}_{n\times p}$ contain the $y_{dtj}$'s and the $\mathbf{x}_{dtj}$'s respectively, $\boldsymbol{\beta} = \boldsymbol{\beta}_{p\times 1}$, $\mathbf{Z}_1 = \underset{1\le d\le D}{\text{diag}}\,(\mathbf{1}_{n_d})$, $\mathbf{Z}_2 = \underset{1\le d\le D}{\text{diag}}\,(\underset{1\le t\le m_d}{\text{diag}}\,(\mathbf{1}_{n_{dt}}))$, $M = \sum_{d=1}^{D} m_d$, $n = \sum_{d=1}^{D} n_d$, $n_d = \sum_{t=1}^{m_d} n_{dt}$, $\mathbf{I}_a$ is the identity matrix of order $a$, $\mathbf{1}_a$ is a column vector of dimension $a$ with all its elements equal to 1, $\mathbf{W} = \underset{1\le d\le D}{\text{diag}}\,(\mathbf{W}_d)$, $\mathbf{W}_d = \underset{1\le t\le m_d}{\text{diag}}\,(\mathbf{W}_{dt})$, $\mathbf{W}_{dt} = \underset{1\le j\le n_{dt}}{\text{diag}}\,(w_{dtj})_{n\times n}$ with $w_{dtj} > 0$ known, $d = 1, \dots, D$, $t = 1, \dots, m_d$, $j = 1, \dots, n_{dt}$ and the covariance matrix of $\mathbf{u}_2$ is $\Omega(\rho) = \underset{1\le d\le D}{\text{diag}}\,(\Omega_d)$, where

$$\Omega_d = \Omega_d(\rho) = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \dots & \rho^{m_d-2} & \rho^{m_d-1} \\ \rho & 1 & \ddots & & \rho^{m_d-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{m_d-2} & & \ddots & 1 & \rho \\ \rho^{m_d-1} & \rho^{m_d-2} & \dots & \rho & 1 \end{pmatrix}_{m_d\times m_d}. \tag{3.7}$$

The model will be fitted by using the REML method with the parametrization

$$\sigma^2 = \sigma_0^2, \quad \varphi_1 = \frac{\sigma_1^2}{\sigma_0^2}, \quad \varphi_2 = \frac{\sigma_2^2}{\sigma_0^2}, \quad \rho = \rho.$$

## 3.3 Individual-level linear model with independent time effects

In this section we present a simplification of model (3.5) that is useful for those cases where survey data is only available for a reduced number of time instants. The new model is defined in the same way as model of Section 3.2, but assuming that $\rho = 0$. Parameter estimates of model (3.8) can also be used as seeds for an iterative fitting method in model (3.5). We assume that

$$y_{dtj} = \mathbf{x}_{dtj}\boldsymbol{\beta} + u_{1,d} + u_{2,dt} + w_{dtj}^{-1/2} e_{dtj}, \quad d = 1, \dots, D, t = 1, \dots, m_d, j = 1, \dots, n_{dt}, \tag{3.8}$$

where the area effects $u_{1,d}$'s are i.i.d. $N(0, \sigma_1^2)$, the time by area effects $u_{2,dt}$'s are $N(0, \sigma_2^2)$, the errors $e_{dtj}$'s are i.i.d. $N(0, \sigma_0^2)$, and the $u_{1,d}$'s, the $u_{2,t}$'s and the $e_{dtj}$'s are mutually independent. Model (3.8) is equal to the model (3.4) studied by Stukel and Rao (1999).

Model (3.8) can be expressed alternatively as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{W}^{-1/2}\mathbf{e}, \tag{3.9}$$

where $\mathbf{u}_1 = \mathbf{u}_{1,D\times 1} \sim N(\mathbf{0}, \sigma_1^2 \mathbf{I}_D)$, $\mathbf{u}_2 = \mathbf{u}_{2,M\times 1} \sim N(\mathbf{0}, \sigma_2^2 \mathbf{I}_M)$ and $\mathbf{e} = \mathbf{e}_{n\times 1} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_n)$ are independent and contain the $u_{1,d}$'s, the $u_{2,t}$'s and the $e_{dtj}$'s respectively. The remaining vectors and matrices appearing in the model equation are defined in the same way as for model (3.6). This model will be fitted by using the REML method with the parametrization

$$\sigma^2 = \sigma_0^2, \quad \varphi_1 = \frac{\sigma_1^2}{\sigma_0^2}, \quad \varphi_2 = \frac{\sigma_2^2}{\sigma_0^2}.$$

## 3.4 Area-level linear model with correlated time effects

Let us consider the model

$$y_{dt} = \mathbf{x}_{dt}\boldsymbol{\beta} + u_{dt} + e_{dt}, \quad d = 1, \ldots, D, \quad t = 1, \ldots, m_d, \tag{3.10}$$

where $y_{dt}$ is a direct estimator of the indicator of interest for area $d$ and time instant $t$, and $\mathbf{x}_{dt}$ is a vector containing the aggregated (population) values of $p$ auxiliary variables. The index $d$ is used for domains and the index $t$ for time instants. We further assume that the random vectors $(u_{d1}, \ldots, u_{dm_d})$, $d = 1, \ldots, D$, follow i.i.d. AR(1) processes with variance and auto-correlation parameters $\sigma_u^2$ and $\rho$ respectively, the errors $e_{dtj}$'s are independent $N(0, \sigma_{dt}^2)$ with known $\sigma_{dt}^2$'s, and the $u_{dt}$'s are independent of the $e_{dt}$'s.

In matrix notation the model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{3.11}$$

where $\mathbf{y} = \underset{1\leq d\leq D}{\text{col}}(\mathbf{y}_d)$, $\mathbf{y}_d = \underset{1\leq t\leq m_d}{\text{col}}(y_{dt})$, $\mathbf{u} = \underset{1\leq d\leq D}{\text{col}}(\mathbf{u}_d)$, $\mathbf{u}_d = \underset{1\leq t\leq m_d}{\text{col}}(u_{dt})$, $\mathbf{e} = \underset{1\leq d\leq D}{\text{col}}(\mathbf{e}_d)$, $\mathbf{e}_d = \underset{1\leq t\leq m_d}{\text{col}}(e_{dt})$, $\mathbf{X} = \underset{1\leq d\leq D}{\text{col}}(\mathbf{X}_d)$, $\mathbf{X}_d = \underset{1\leq t\leq m_d}{\text{col}}(\mathbf{x}_{dt})$, $\mathbf{x}_{dt} = \underset{1\leq j\leq p}{\text{col}'}(x_{dtj})$, $\boldsymbol{\beta} = \underset{1\leq j\leq p}{\text{col}}(\beta_j)$, $\mathbf{Z} = \mathbf{I}_{M\times M}$ and $M = \sum_{d=1}^{D} m_d$. In this notation, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{V}_u)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{V}_e)$ are independent with covariance matrices

$$\mathbf{V}_u = \sigma_u^2\Omega(\rho), \quad \Omega(\rho) = \underset{1\leq d\leq D}{\text{diag}}\,(\Omega_d(\rho)), \quad \mathbf{V}_e = \underset{1\leq d\leq D}{\text{diag}}\,(\mathbf{V}_{ed}), \quad \mathbf{V}_{ed} = \underset{1\leq t\leq m_d}{\text{diag}}\,(\sigma_{dt}^2),$$

where the $\sigma_{dt}^2$ are known and $\Omega_d(\rho)$ is defined in (3.7). The REML method will be used to fit this model.

## 3.5 Area-level linear model with independent time effects

This section presents a simplification of model (3.10) that is useful for those cases where survey data is only available for a reduced number of time instants. The new model is defined in the same way as model of Section 3.4, but assuming that $\rho = 0$. Parameter estimates of model (3.12) can also be used as seeds for an iterative fitting method in model (3.10). We assume that

$$y_{dt} = \mathbf{x}_{dt}\boldsymbol{\beta} + u_{dt} + e_{dt}, \quad d = 1, \ldots, D, \quad t = 1, \ldots, m_d, \tag{3.12}$$

where the vectors $u_{dt}$'s are $N(0, \sigma_u^2)$, the errors $e_{dt}$'s are independent $N(0, \sigma_{dt}^2)$, and the $u_{dt}$'s are independent of the $e_{dt}$'s. This model is more general than model (3.2) considered by Rao and Yu (1994) in the sense that $u_d$ is substituted by $u_{dt}$ to take into account the area-by-time variability through specific random effects.

Model (3.12) can be alternatively written in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{3.13}$$

where $\mathbf{u} \sim N(\mathbf{0}, \mathbf{V}_u)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{V}_e)$ are independent with covariance matrices

$$\mathbf{V}_u = \sigma_u^2 \mathbf{I}_M, \quad \mathbf{I}_M = \underset{1\leq d\leq D}{\text{diag}} (\mathbf{I}_{m_d}), \quad \mathbf{V}_e = \underset{1\leq d\leq D}{\text{diag}} (\mathbf{V}_{ed}), \quad \mathbf{V}_{ed} = \underset{1\leq t\leq m_d}{\text{col}} (\sigma_{dt}^2),$$

and error variances $\sigma_{dt}^2$ assumed to be known. The vectors $\mathbf{y}$ and $\boldsymbol{\beta}$ and the matrices $\mathbf{X}$ and $\mathbf{Z}$ are defined in the same way as for model (3.6). This model will be fitted by the REML method.

## 3.6 Logistic model with correlated time effects

In this section we introduce a logistic mixed model at the unit level with time correlated random effects. The model is useful when response variables are dichotomic or binomial and we are willing to model area and time variability. To introduce the model we first give the assumptions on the random factors.

Let $u_{1,d}$ and $u_{2,dt}$ denote random effects of area $d$ and of time instant $t$ within area $d$ respectively. Let $\mathbf{u}_1 = \underset{1\leq d\leq D}{\text{col}} (u_{1,d})$ and $\mathbf{u}_2 = \underset{1\leq d\leq D}{\text{col}} (\mathbf{u}_{2,d})$, with $\mathbf{u}_{2,d} = \underset{1\leq t\leq m_d}{\text{col}} (u_{2,dt})$, be the random vectors containing the random effects and define $\mathbf{u} = (\mathbf{u}_1', \mathbf{u}_2')'$. Assume that

$$\mathbf{u}_1 \sim N(\mathbf{0}, \varphi_1 \mathbf{I}_D) \quad \text{and} \quad \mathbf{u}_2 \sim N(\mathbf{0}, \varphi_2 \Omega(\rho))$$

are independent with $\Omega(\rho) = \underset{1\leq d\leq D}{\text{diag}} (\Omega_d)$ and with $\Omega_d$ defined in (3.7). Then $\mathbf{V}_u = \text{var}(\mathbf{u}) = \text{diag}(\varphi_1 \mathbf{I}_D, \varphi_2 \Omega(\rho))$.

Concerning the target variable, we assume that the observations $y_{dtj}$, given $\mathbf{u}$, are independent with binomial distributions

$$y_{dtj}|_{u_{1,d},u_{2,dt}} \sim \text{Bin}(\nu_{dtj}, p_{dtj}),\ d = 1, \ldots, D,\ t = 1, \ldots, m_d,\ j = 1, \ldots, n_{dt}, \tag{3.14}$$

where $\sum_{t=1}^{m_d} n_{dt} = n_d$, $\sum_{d=1}^{D} n_d = n$ and $\sum_{d=1}^{D} m_d = M$. For the natural parameter $\eta_{dtj} = \log(p_{dtj}/(1-p_{dtj}))$, we assume the model

$$\eta_{dtj} = \mathbf{x}_{dtj}\boldsymbol{\beta} + u_{1,d} + u_{2,dt},\ d = 1, \ldots, D,\ t = 1, \ldots, m_d,\ j = 1, \ldots, n_{dt}, \tag{3.15}$$

where $\mathbf{x}_{dtj}$ is the row $(d, i, j)$ of matrix $\mathbf{X} = \underset{1\leq d\leq D}{\text{col}} (\mathbf{X}_d)$, $\mathbf{X}_d = \underset{1\leq t\leq m_d}{\text{col}} (\mathbf{X}_{dt})$, $\mathbf{X}_{dt} = \underset{1\leq j\leq n_{dt}}{\text{col}} (\mathbf{x}_{dtj})$.

The mean and the variance of $y_{dtj}$, given $u_{1,d}$ and $u_{2,dt}$, are $\mu_{dtj} = \nu_{dtj} p_{dtj}$ and $w_{dtj} = \nu_{dtj} p_{dtj}(1 - p_{dtj})$ respectively. The probabilities $p_{dtj}$ are obtained from the natural parameters by using the formulas

$$p_{dtj} = \frac{\exp\{\eta_{dtj}\}}{1 + \exp\{\eta_{dtj}\}},\ d = 1, \ldots, D,\ t = 1, \ldots, m_d,\ j = 1, \ldots, n_{dt}.$$

In matrix notation model (3.15) can be expressed in the form

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u},$$

where

$$\mathbf{Z}_1 = \underset{1\leq d\leq D}{\text{diag}} (\mathbf{1}_{n_d}), \quad \mathbf{Z}_2 = \underset{1\leq d\leq D}{\text{diag}} (\underset{1\leq t\leq m_d}{\text{diag}} (\mathbf{1}_{n_{dt}})) \quad \text{and} \quad \mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2).$$

This model can be fitted by the PQL-REML method.

## 3.7 Logistic model with independent time effects

This section presents a simplification of the model of Section 3.6 that is useful for those cases where survey data is only available for a reduced number of time instants. The new model is defined in the same way as before, but assuming that $\rho = 0$. Parameter estimates of model (3.16) can also be used as seeds for an iterative fitting method in model (3.15).

Let $u_{1,d}$ and $u_{2,dt}$ be the random effects for area $d$ and time instant $t$ (within area $d$). Let $\mathbf{u}_1 = \underset{1\leq d\leq D}{\text{col}}(u_{1,d})$, $\mathbf{u}_2 = \underset{1\leq d\leq D}{\text{col}}(\mathbf{u}_{2,d})$, with $\mathbf{u}_{2,d} = \underset{1\leq t\leq m_d}{\text{col}}(u_{2,dt})$, be the random vectors containing the random effects and define $\mathbf{u} = (\mathbf{u}_1', \mathbf{u}_2')'$. We assume that

$$\mathbf{u}_1 \sim N(\mathbf{0}, \varphi_1 \mathbf{I}_D) \quad \text{and} \quad \mathbf{u}_2 \sim N(\mathbf{0}, \varphi_2 \mathbf{I}_M)$$

are independent with $\sum_{t=1}^{m_d} n_{dt} = n_d$, $\sum_{d=1}^{D} n_d = n$ and $\sum_{d=1}^{D} m_d = M$.

We also assume that the observations $y_{dtj}$, conditioned to $\mathbf{u}$, are independent with binomial distributions introduced in (3.14). For the natural parameter $\eta_{dtj} = \log \frac{p_{dtj}}{1-p_{dtj}}$ we assume the model

$$\eta_{dtj} = \mathbf{x}_{dtj}\boldsymbol{\beta} + u_{1,d} + u_{2,dt}, \; d = 1, \ldots, D, \; t = 1, \ldots, m_d, \; j = 1, \ldots, n_{dt}, \qquad (3.16)$$

where $\mathbf{x}_{dtj}$ is the row $(d, i, j)$ of matrix $\mathbf{X} = \underset{1\leq d\leq D}{\text{col}}(\mathbf{X}_d)$, $\mathbf{X}_d = \underset{1\leq t\leq m_d}{\text{col}}(\mathbf{X}_{dt})$, $\mathbf{X}_{dt} = \underset{1\leq j\leq n_{dt}}{\text{col}}(\mathbf{x}_{dtj})$. In matrix notation (3.16) can be expressed in the form

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u},$$

where $\mathbf{Z}_1$, $\mathbf{Z}_2$ are defined in Section 3.6 and $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$. This model can be fitted by the PQL-REML method.

## 3.8 References

Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*, 2nd ed. San Francisco: Holden-Day.

Datta, G.S. and Ghosh, M. (1991). Bayesian prediction in linear models: applications to small area estimation. *Annals of Statistics*, **19**, 1748–1770.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the U.S. states. *Journal of the American Statistical Association*, **94**, 1074–1082.

Datta, G.S., Lahiri, P. and Maiti, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, **102**, 83-97.

Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2005). *Analysis of Longitudinal Data*. Oxford University Press. Oxford.

Frabrizi, E., Ferrante, M. R., Pacei, S. (2007). Small area estimation of average household income based on unit level models for panel data. *Survey Methodology*, **33**, 187–198.

Ghosh, M. and Lahiri, P. (1988). Bayes and empirical Bayes analysis in multistage sampling. In: *Gupta, SS., Berger, J.O. (Eds.), Statistical Decision Theory and Related Topics IV*, vol. 1. Springer, New York.

Ghosh, M., Nangia, N. and Kim, D. (1996). Estimation of median income of four-person families: a Bayesian time series approach. *Journal of the American Statistical Association*, **91**, 1423–1431.

Harvey, A. C. (1989). *Forecasting Structural Time Series with the Kalman Filter*. Cambridge: Cambridge University Press.

Pfeffermann, D. and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, **16**, 217-237.

Pfeffermann, D., Feder, M. and Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business and Economic Statistics*, **16**, 339–348.

Pfeffermann, D. and Barnard, C. (1991). Some new estimators for small area means with applications to the assessment of farmland values. *Journal of Business and Economic Statistics*, **9**, 73–84.

Pfeffermann, D. (2002). Small Area Estimation - New Developments and Directions. *International Statistical Review*, **70**, 1, 125-143.

Rao, J.N.K. and Yu, M. (1994). Small area estimation by combining time series and cross sectional data. *Canadian Journal of Statistics*, **22**, 511-528.

Stukel, D.M and Rao, J.N.K (1997). Estimation of regression models with nested error structure and unequal error variances under two and three stage cluster sampling. *Statistics and Probability Letters*, **35**, 401–407.

Stukel, D.M and Rao, J.N.K (1999). On small-area estimation under two-fold nested error regression models. *Journal of Statistical Planning and Inference*, **78**, 131–147.

You, Y. and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, **26**, 173–181.

You, Y., Rao, J.N.K. and Gambino, J. (2001). Model-based unemployment raye estimation for the Canadian Labour Force Survey: a hierarchical approach. Technical report, Household Survey Method Division. Statistics Canada.

# Chapter 4

# Spatial models

## 4.1 Introduction

In small area estimation, models with random effects for the areas introduce a correlation structure for the elements within the same area, but elements in different small areas are considered to be uncorrelated. However, it is known that socioeconomic characteristics of individuals in neighboring regions are usually more alike than those of individuals in distant regions. In statistical terms, this means that there is some kind of dependency relationship between individuals that are in neighboring regions. When this dependency is not completely captured by the auxiliary variables in the model, it should be somehow incorporated in the correlation structure of the model. Not doing it may affect seriously the performance of inferential procedures (Cressie, 1993). Nevertheless, the introduction of a dependence structure among small areas entails a serious conceptual difference with respect to the traditional framework of independent small areas, in which the overall covariance matrix is block-diagonal, see Prasad and Rao (1990). Thus, these models require new specific theoretical developments.

Cressie (1991) used a model with spatially correlated random effects to predict census undercount in small areas. More recently, Singh et al. (2005) considered an extension of the Fay-Herriot model (Fay and Herriot, 1979) in which the area random effects follow a Simultaneously Autoregressive (SAR) process. They obtained an approximation of the mean squared error of the EBLUP under this model, and also studied a spatio-temporal model in which, for each time point, the area random effects follow a SAR process, but the spatial autocorrelation parameter is constant along time. Petrucci and Salvati (2006) used the same spatial model to estimate erosion at the Rathbun Lake Watershed in Iowa. Pratesi and Salvati (2007) analysed the performance of the spatial EBLUP by simulations, and found that the introduction of the spatial correlation reduces both the variance and the bias of the EBLUP. They discussed the estimation of the mean squared error and applied the results to estimation of annual mean income in small areas of Tuscany.

Particular models with spatially correlated area random effects were also considered in the EURAREA project, see Section C4 of the first report (http://www.statistics.gov.uk/eurarea/download.asp). Instead of a neighborhood structure among areas, they followed a different approach in which the covariance between the effects associated to two different areas is a decreasing function of the distance between these areas. They also introduced a parameter that determines the strength of the correlation between areas.

This chapter introduces several models with spatially correlated area effects, namely the Fay–Herriot model, a unit level nested-error regression model, and a unit level Multinomial logit model, which includes the area level model and the univariate logistic model as particular cases.

## 4.2 Area level model with spatially correlated area effects

Consider a finite population that is partitioned into $D$ small areas. The basic FH model relates linearly the quantity of inferential interest for small area $d$, $\mu_d$ (usually the area mean or total) to $p$ area level auxiliary covariates $\mathbf{x}_d = (x_{d1}, x_{d2}, \ldots, x_{dp})$, and includes a random effect $v_d$ associated to the area; that is,

$$\mu_d = \mathbf{x}_d\boldsymbol{\beta} + v_d, \quad d = 1, \ldots, D. \tag{4.1}$$

Here $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression parameters, the random effects $v_d$ are independent and identically distributed with mean 0 and variance $\sigma_v^2$. Moreover, it assumes that a design-unbiased direct estimator $y_d$ of $\mu_d$ is available for each of the $D$ small areas, and that these direct estimators can be expressed as

$$y_d = \mu_d + e_d, \quad e_d \sim \text{iid } N(0, \psi_d), \quad d = 1, \ldots, D, \tag{4.2}$$

where the $e_d$'s are sampling errors, which are independent of the random effects $v_d$ (Ghosh and Rao, 1994) and the variances $\psi_d$ are known for all $d$. Combining (4.1) and (4.2), the full model is

$$y_d = \mathbf{x}_d\boldsymbol{\beta} + v_d + e_d, \quad d = 1, \ldots, D. \tag{4.3}$$

Let us define vectors $\mathbf{y} = (y_1, \ldots, y_D)'$, $\mathbf{v} = (v_1, \ldots, v_D)'$ and $\mathbf{e} = (e_1, \ldots, e_D)'$, and matrices $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_D')'$ and $\Psi = \text{diag}(\psi_1, \ldots, \psi_D)$. Then, the model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v} + \mathbf{e}, \quad \mathbf{e} \sim N(\mathbf{0}_D, \Psi), \tag{4.4}$$

where the notation $\mathbf{0}_k$ is used for a column vector of zeros of size $k$. When the elements of $\mathbf{v}$ are independent, this is a special case of the general linear mixed model with diagonal covariance structure. However, this model can be extended to allow for spatially correlated area effects by considering that $\mathbf{v}$ is the result of a SAR process with unknown autorregression parameter $\varrho$ and proximity matrix $\mathbf{W}$ (Anselin, 1992; Cressie, 1993), specified as

$$\mathbf{v} = \varrho\mathbf{W}\mathbf{v} + \mathbf{u}, \quad \text{where} \quad \mathbf{u} \sim N(\mathbf{0}, \sigma_u^2\mathbf{I}_D). \tag{4.5}$$

Here, $\mathbf{I}_D$ denotes the $D \times D$ identity matrix and $\sigma_u^2$ is an unknown parameter. The diagonal elements of the proximity matrix $\mathbf{W}$ are zero and we consider that the rows of this matrix are standardized in the sense that they sum up to one. Under this setup, $\varrho \in (-1, 1)$ is called spatial autocorrelation parameter (Banerjee et al., 2004). We also assume that the matrix $(\mathbf{I}_D - \varrho\mathbf{W})$ is non-singular. Then $\mathbf{v}$ can be expressed as

$$\mathbf{v} = (\mathbf{I}_D - \varrho\mathbf{W})^{-1}\mathbf{u}. \tag{4.6}$$

Hereafter, the vector of variance components will be denoted $\boldsymbol{\theta} = (\theta_1, \theta_2)' = (\sigma_u^2, \varrho)'$. Equation (4.6) together with (4.5) imply that $\mathbf{v}$ has mean vector $\mathbf{0}$ and covariance matrix equal to

$$\mathbf{G} = \sigma_u^2[(\mathbf{I}_D - \varrho\mathbf{W})'(\mathbf{I}_D - \varrho\mathbf{W})]^{-1}. \tag{4.7}$$

Combining (4.4) and (4.6), the model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_D - \varrho\mathbf{W})^{-1}\mathbf{u} + \mathbf{e}, \tag{4.8}$$

where $\mathbf{e}$ is independent of $\mathbf{v}$ and the covariance matrix of $\mathbf{y}$ is equal to

$$\mathbf{V} = \mathbf{G} + \Psi.$$

## 4.3 Unit level model with spatially correlated area effects

Here we consider a unit level model in which the area effects are spatially correlated following a SAR process. That is, we consider that the response variable for unit $j$ in area $d$ follows the model

$$y_{dj} = \mathbf{x}_{dj}\boldsymbol{\beta} + v_d + e_{dj}, \quad e_{dj} \sim \text{iid } N(0, \sigma^2), \quad j = 1, \dots, n_d, \quad d = 1, \dots, D,$$

where the vector $\mathbf{v} = (v_1, \dots, v_D)'$ of random effects follow the SAR process (4.6). Let us define vectors and matrices containing sample elements of area $d$,

$$\mathbf{y}_d = \underset{1\le j\le n_d}{\text{col}}(y_{dj}), \quad \mathbf{e}_d = \underset{1\le j\le n_d}{\text{col}}(e_{dj}), \quad \mathbf{X}_d = \underset{1\le j\le n_d}{\text{col}}(\mathbf{x}_{dj}),$$

and vectors and matrices containing all sample elements

$$\mathbf{y} = \underset{1\le d\le D}{\text{col}}(\mathbf{y}_d), \quad \mathbf{e} = \underset{1\le d\le D}{\text{col}}(\mathbf{e}_d), \quad \mathbf{X} = \underset{1\le d\le D}{\text{col}}(\mathbf{x}_d), \quad \mathbf{Z} = \underset{1\le d\le D}{\text{diag}}(\mathbf{1}_{n_d}),$$

where $\mathbf{1}_k$ stands for a vector of ones of size $k$. Then, the unit level model can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \quad \mathbf{v} \sim N(\mathbf{0}_D, \mathbf{G}), \quad \mathbf{e} \sim N(\mathbf{0}_n, \sigma^2\mathbf{I}_n),$$

where $\mathbf{v}$ and $\mathbf{e}$ are independent, $\mathbf{G}$ is defined in (4.7) and $n = \sum_{d=1}^D n_d$ is the overall sample size.

## 4.4 Unit level logit model with spatially correlated area effects

In this section, the study variable is a factor with two categories. We assume that for each area $d$, there are $n_d$ variables $y_{dj}$, $j = 1, \dots, n_d$, containing the counts of individuals out of a total of $m_{dj}$ that belong to the first category of the factor. The count for the second category of the factor is equal to $m_{dj} - y_{dj}$. We assume the existence of a random vector $\mathbf{v}$ of size $D$ whose properties will be described later. Given a realization of $\mathbf{v}$, the counts $y_{dj}$ are independent, and each $y_{dj}$ follows a Binomial distribution with size

$m_{dj}$ and probability $p_{dj}$. The probability of belonging to the second category is $1 - p_{dj}$. Thus, the mass function of $y_{dj}$ given $\mathbf{v}$ is

$$f(y_{dj}|\mathbf{v}) = \left( \begin{array}{c} m_{dj} \\ y_{dj} \end{array} \right) p_{dj}^{y_{dj}} (1 - p_{dj})^{m_{dj} - y_{dj}}, \quad j = 1, \ldots, n_d, \quad d = 1, \ldots, D \tag{4.9}$$

and can be expressed as

$$f(y_{dj}|\mathbf{v}) = c(y_{dj}) \exp \left\{ y_{dj} \log[p_{dj}/(1 - p_{dj})] + m_{dj} \log(1 - p_{dj}) \right\},$$

where $c(y_{dj})$ is a function only of $y_{dj}$ and $m_{dj}$. Thus, this mass function belongs to the natural exponential family with natural parameter $\theta_{dj} = \log[p_{dj}/(1 - p_{dj})]$. The probability can be written as a function of $\theta_{dj}$ as

$$p_{dj} = \exp(\theta_{dj}) / \left[1 + \exp(\theta_{dj})\right].$$

The mean and the variance of $y_{dj}$ are equal to

$$\mu_{dj} = m_{dj} p_{dj}, \quad \sigma_{dj}^2 = m_{dj} p_{dj} (1 - p_{dj}),$$

We further assume that the probability $p_{dj}$ is related to a vector $\mathbf{x}_{dj}$ containing the values of $p$ explanatory variables and to the random effect of area $d$, $v_d$, through the logit link (which is the natural link), in the form

$$\log[p_{dj}/(1 - p_{dj})] = \mathbf{x}_{dj} \boldsymbol{\beta} + v_d, \quad j = 1, \ldots, n_d, \ d = 1, \ldots, D. \tag{4.10}$$

Here, again the vector of random effects $\mathbf{v} = (v_1, \ldots, v_D)'$ is assumed to follow the SAR process (4.5) with proximity matrix $\mathbf{W}$ and autocorrelation parameter $\varrho$, where the autocorrelation parameter $\varrho \in (-1, 1)$ and the variance $\sigma_u^2 \in (0, \infty)$ are unknown. Again, solving for $\mathbf{v}$ in equation (4.5) we obtain that

$$\mathbf{v} \sim N(\mathbf{0}_D, \mathbf{G}), \tag{4.11}$$

where $\mathbf{G}$ is given in (4.7). Model (4.10) can be written in terms of the natural parameter as

$$\theta_{dj} = \mathbf{x}_{dj} \boldsymbol{\beta} + v_d, \quad d = 1, \ldots, D.$$

Now let us define the vector and matrices with the area elements,

$$\boldsymbol{\theta}_d = \underset{1 \le j \le n_d}{\text{col}} (\theta_{dj}), \quad \mathbf{X}_d = \underset{1 \le j \le n_d}{\text{col}} (\mathbf{x}_{dj}).$$

Then, the model formulated for the areas is

$$\boldsymbol{\theta}_d = \mathbf{X}_d \boldsymbol{\beta} + \mathbf{1}_{n_d} v_d, \quad d = 1, \ldots, D.$$

Finally, defining

$$\boldsymbol{\theta} = \underset{1 \le d \le D}{\text{col}} (\boldsymbol{\theta}_d), \quad \mathbf{X} = \underset{1 \le d \le D}{\text{col}} (\mathbf{X}_d), \quad \mathbf{Z} = \underset{1 \le d \le D}{\text{diag}} (\mathbf{1}_{n_d}),$$

the model becomes simply

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}.$$

Interesting particular cases of this model are the area level model, obtained simply by setting $n_d = 1$, $d = 1, \ldots, D$, and the model with Binomial sizes $m_{dj} = 1$ for all $j = 1, \ldots, n_d$ and all $d = 1, \ldots, D$, in which the response variables $y_{dj}$ indicate whether individual $j$ belongs or not to the first category of the factor.

## 4.5 References

Anselin, L. (1988). *Spatial Econometrics. Methods and Models*. Boston: Kluwer Academic Publishers.

Banerjee, S., Carlin, B. and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. New York: Chapman and Hall.

Cressie, N. (1991). Small-area prediction of undercount using the general linear model. *Proceedings of Statistic Symposium 90: Measurement and Improvement of Data Quality, Ottawa: Statistics Canada*, 93–105.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons.

Fay, R. and Herriot, R. (1979). Estimates of income for small places: an application of James–Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269–277.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science* **9**, 55–76 (Discussion: 76–93).

Petrucci, A. and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological and Environmental Statistics* **11**, 169–182.

Prasad, N. and Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association* **85**, 163–171.

Pratesi, M. and Salvati, N. (2007). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods & Applications* **17**, 113–141.

Singh, B., Shukla, G. and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology* **31**, 183–195.

# Chapter 5

# Small Area Estimation of Poverty indicators

## 5.1 Introduction

The model-based approach to small area estimation, in which an explicit model is used to “borrow strength” from related areas, is a relatively new research field; the first known work is that of Fay and Herriot (1979), who used an area level model to estimate mean per capita income in U.S. small places. A good source of information about small area estimation with emphasis on the model-based approach is the monograph of Rao (2003), which gathers most of the works done until the publication date. However, despite the great importance of the disposal of accurate and sufficiently disaggregated poverty statistics for policy makers, it seems that the estimation of poverty in small areas has been studied only very recently. It probably started with the SAIPE program (Small Area Income & Poverty Estimates) of the U.S. Census Bureau, see the web page http://www.census.gov/hhes/www/saipe with all the literature therein. The main objective of this program is to provide updated estimates of income and poverty statistics for the administration of federal programs, and the allocation of federal funds to local jurisdictions. The county level methodology, summarized by Bell (1997), basically uses a Fay-Herriot model to produce county estimates of school-age children under poverty. Maiti and Slud (2002) compare the previous model with a logistic model including county random effects to produce poverty rates.

The World Bank is releasing poverty and inequality estimates for small areas of some countries using the methodology of Elbers et al. (2003). This methodology is currently widely extended, see, e.g., the works of Neri et al. (2005), Ballini et al. (2006), Tarozzi and Deaton (2007) and Haslett and Jones (2006). They assume a unit level model that combines both census and survey data. Using that model, they produce disaggregated maps that describe the spatial distribution of poverty and inequality. However, in many European countries the census is decennial, and in the years in the middle of the period between two consecutive censuses data are outdated due to the rapid change in the distribution of socioeconomic variables, particularly in recent years. Another inconvenience of the method of Elbers et al. (2003) is that their proposed model does not allow for between-area variation beyond that explained by the auxiliary variables, and it considers that individuals belonging to the same area are independent.

Here we propose methods that yield poverty estimators in small areas for intercensal years along with census years and that take into account the variation between areas. This will be achieved by considering models with area random effects that also admit correlation of within-area units. The methods will be adapted to the data structures supplied by national statistical offices and other administrative registers.

Measures of inequality include the Gini coefficient, the Sen index, the general entropy and the Theil index. Here we focus mainly on poverty measures and the methods will be illustrated for these measures, although some of the methods that will be introduced in this chapter allow the estimation of these inequality measures as well.

A common definition of poverty classifies a person as “under poverty” when the selected welfare variable for this person in below the 60% of the median. Indeed, the relative nature of this and other definitions of poverty, and the low frequency of the outcome for small domains or geographical areas makes it necessary to appeal to small area techniques that improve the estimation procedures by the assumption of models. These models make use of relatively realistic relationships between the variables of interest to link all sample data, in such a way that the estimation errors can be drastically reduced as long as model assumptions are (at least approximately) true.

## 5.2 FGT poverty measures for small areas

Consider a population of size $N$ that is partitioned into $D$ small areas of sizes $N_1, \dots, N_D$. Let $E_{dj}$ be a suitable quantitative measure of welfare for individual $j$ in small area $d$, such as income or expenditure and let $z$ be the poverty line; that is, the threshold for $E_{dj}$ under which a person is considered as “under poverty”. The family of poverty measures of Foster, Greer and Thorbecke (1984), called FGT poverty measures, for a small area $d$ is

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left( \frac{z - E_{dj}}{z} \right)^{\alpha} I(E_{dj} < z), \quad \alpha = 0, 1, 2, \quad d = 1, \dots, D,$$

where $I(E_{dj} < z) = 1$ if $E_{dj} < z$ (person under poverty) and $I(E_{dj} < z) = 0$ if $E_{dj} \geq z$ (person not under poverty). For $\alpha = 0$ we get the proportion of individuals under poverty in small area $d$, also called poverty incidence or head count ratio. The measure for $\alpha = 1$ is called poverty gap, and measures the area mean of the relative distance to non-poverty (the poverty gap) of each individual. For $\alpha = 2$ the measure is called poverty severity, since the poverty gap for an individual, which is a number in $[0, 1]$, is squared, and therefore decreased.

## 5.3 Direct estimators of poverty measures

In the inference process, a random sample of size $n < N$ is drawn from the population according to a specified sampling design. Let $\Omega$ denote the set of indexes of the population units. Let $s$ be the set units selected in the sample and $r$ the set of indexes of the units that are not selected (with size $N - n$). The restrictions of $\Omega$, $s$, $N$ and $n$ to area $d$ are denoted by $\Omega_d$, $s_d$, $N_d$ and $n_d$ respectively, where

$n = n_1 + \cdots + n_D$. The sample FGT poverty measures are given by

$$f_{\alpha d} = \frac{1}{n_d} \sum_{j \in s_d} \left( \frac{z - E_{dj}}{z} \right)^{\alpha} I(E_{dj} < z), \quad \alpha = 0, 1, 2, \quad d = 1, \dots, D. \tag{5.1}$$

Direct estimators, as sample estimators, use only the sample data from the corresponding small area. However, direct estimators are usually design-based, which means that they have desirable properties with respect to the sampling design. Thus, they are unbiased in the sense that the mean over all possible samples $s_d$ is the true quantity of interest. Let $w_{dj}$ be the sampling weight (inverse of the probability of inclusion) of individual $j$ from area $d$. Direct estimators of the FGT measures are

$$f^{w}_{\alpha d} = \frac{1}{\hat{N}_d} \sum_{j \in s_d} w_{dj} \left( \frac{z - E_{dj}}{z} \right)^{\alpha} I(E_{dj} < z), \quad \alpha = 0, 1, 2, \quad d = 1, \dots, D, \tag{5.2}$$

where $\hat{N}_d = \sum_{j \in s_d} w_{dj}$ is the direct estimator of the population size of small area $d$, $N_d$.

The limited sample sizes $n_d$ within some of the areas prevent the use of estimators such as (5.1) or (5.2). For this reason, it is necessary to appeal to “indirect” estimators that make use of related data from other areas. We will use the model-based approach, in which all sample data are linked by a model that establishes the relationships between the small areas.

Parametric models assume knowledge of the probability distribution generating the response values. Sometimes the distribution is not known, and other times it is too complicated to derive suitable model-based estimators. In the next sections we describe possible models that could allow small area estimation of the FGT poverty measures.

In the following sections we describe how to model each FGT measure. Section 5.4 is referred to the poverty incidence or head count ratio, Section 5.5 to the poverty gap, and Section 5.6 to the poverty severity. Section 5.7 introduces a different approach that provides small area estimators for the whole family of FGT measures without considering different models for the different members of the family.

## 5.4 Model-based estimation of the poverty incidence

### 5.4.1 Unit level model

For $\alpha = 0$, the FGT poverty measure is simply the proportion of individuals under poverty,

$$P_d = F_{0d} = \frac{1}{N_d} \sum_{j=1}^{N_d} I(E_{dj} < z), \quad d = 1, \dots, D.$$

Modelling these quantities is not difficult. The model can be established at the unit-level, at the area level, or even in an intermediate level. At the unit level, let $Y_{dj} = I(E_{dj} < z)$; that is, $Y_{dj} = 1$ if the person is under poverty and $Y_{dj} = 0$ if the person is not under poverty. Let $p_{dj}$ be the probability of individual $j$ in small area $d$ being under poverty. Then a suitable model for this is the logistic model with random area effects. Let $u_d$ be the random effect of area $d$, $d = 1, \ldots, D$. We assume that $Y_{dj}|u_d \sim \text{Bern}(p_{dj})$, and that the probabilities $p_{dj}$ vary with the values of $p$ auxiliary variables in the form

$$\log\left[p_{dj}/(1-p_{dj})\right] = \mathbf{x}_{dj}\boldsymbol{\beta} + u_d, \quad u_d \sim \text{iid } N(0, \sigma_u^2), \quad j = 1, \ldots, N_d, \quad d = 1, \ldots, D. \tag{5.3}$$

The proportion of people under poverty can be expressed as

$$P_d = \frac{1}{N_d}\sum_{j=1}^{N_d} Y_{dj} = \frac{1}{N_d}\left[\sum_{j \in s_d} Y_{dj} + \sum_{j \in r_d} Y_{dj}\right].$$

The first sum in this expression is over individuals in the sample, which is known. The second sum is over out-of-sample individuals. The out-of-sample values $Y_{dj}$ for $j \in r_d$ can be predicted by fitting model (5.3); that is, taking

$$\hat{Y}_{dj} = \hat{E}[Y_{dj}|u_d] = \hat{p}_{dj} = \exp(\mathbf{x}_{dj}\hat{\boldsymbol{\beta}} + \hat{u}_d)/[1 + \exp(\mathbf{x}_{dj}\hat{\boldsymbol{\beta}} + \hat{u}_d)],$$

for a suitable estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, and a predictor $\hat{u}_d$ of $u_d$. Then, a model-based predictor of $P_d$ is

$$\hat{P}_d = \frac{1}{N_d}\left[\sum_{j \in s_d} Y_{dj} + \sum_{j \in r_d} \hat{Y}_{dj}\right].$$

### 5.4.2 Area level model

When the values of the auxiliary variables are available only at the area level, the model is usually stated at that level. In this case, let $\mathbf{x}_d$ be the vector of $p$ area level covariates, $Y_d = \sum_{j \in s_d} Y_{dj}$ the number of persons under poverty in the sample from area $d$, and $p_d$ the probability of being under poverty in area $d$. We assume that $Y_d|u_d \sim \text{Bin}(n_d, p_d)$, where

$$\log\left[p_d/(1-p_d)\right] = \mathbf{x}_d\boldsymbol{\beta} + u_d, \quad u_d \sim \text{iid } N(0, \sigma_u^2), \quad d = 1, \ldots, D.$$

This model yields the predictor

$$\hat{P}_d = \hat{p}_d = \exp(\mathbf{x}_d\hat{\boldsymbol{\beta}} + \hat{u}_d)/[1 + \exp(\mathbf{x}_d\hat{\boldsymbol{\beta}} + \hat{u}_d)], \quad d = 1, \ldots, D.$$

Consider a grouping variable $A$ such as Sex, Age or Sex crossed with Age, with $K$ levels. Sometimes the totals of the auxiliary variables in each area for the different levels $a = 1, 2, \ldots, K$, are available. Then the model can be stated at the level of area crossed with the grouping variable $A$. More concretely, let $\mathbf{x}_{da}$ be the vector of totals of the covariates for area $d$ and level $a$, $m_{da}$ the number of people sampled in area $d$ and group $a$, $Y_{da}$ be the total number of people under poverty in the sample from area $d$ and group $a$, and $p_{da}$ the probability of being under poverty in the same group and area. Then, we assume that $Y_{da}|u_d \sim \text{Bin}(m_{da}, p_{da})$, where

$$\log\left[p_{da}/(1-p_{da})\right] = \mathbf{x}_{da}\boldsymbol{\beta} + u_d, \quad u_d \sim \text{iid } N(0, \sigma_u^2), \quad d = 1, \ldots, D.$$

Let $N_{da}$ be the total number of individuals in area $d$ and group $a$. Then $N_{da} - m_{da}$ is the number of individuals in that same group and area that are not in the sample. If we assume that the probability of being under poverty is the same for the individuals in and out of the sample, a predictor of the total of out-of-sample people that are under poverty $Y_{da}^r$ is obtained after fitting the model, as

$$\hat{Y}_{da}^r = (N_{da} - m_{da})\hat{p}_{da} = (N_{da} - m_{da})\exp(\mathbf{x}_{da}\hat{\boldsymbol{\beta}} + \hat{u}_d)/[1 + \exp(\mathbf{x}_{da}\hat{\boldsymbol{\beta}} + \hat{u}_d)].$$

Finally, a predictor of $P_d$ is given by

$$\hat{P}_d = \sum_{a=1}^{K}(Y_{da} + \hat{Y}_{da}^r).$$

For more details on this approach, but applied to the estimation of labor force quantities, see Molina, Saei and Lombardía (2007).

## 5.5 Model-based estimation of the poverty gap

### 5.5.1 Two parts unit level model – Approach 1

This section deals with model-based estimation of the poverty gap, (FGT measure for $\alpha = 1$) given by

$$G_d = F_{1d} = \frac{1}{N_d}\sum_{j=1}^{N_d}\left(\frac{z - E_{dj}}{z}\right) I(E_{dj} < z), \quad d = 1, \dots, D.$$

Let us define the random variables

$$G_{dj} = \left(\frac{z - E_{dj}}{z}\right) I(E_{dj} < z), \quad j = 1, \dots, N_d, \quad d = 1, \dots, D.$$

Then, the quantities of interest are simply the small area means of the $G_{dj}$'s,

$$G_d = \frac{1}{N_d}\sum_{j=1}^{N_d} G_{dj} = \frac{1}{N_d}\left[\sum_{j \in s_d} G_{dj} + \sum_{j \in r_d} G_{dj}\right], \quad d = 1, \dots, D.$$

Thus, a model for the individual gaps $G_{dj}$ would provide estimators of the average gap $G_d$. The problem is that the distribution of the gaps $G_{dj}$ has positive mass at zero. Indeed, the mass of $G_{dj}$ at zero is equal to the probability that $E_{dj}$ is over the poverty line $z$. Using the ideas of Pfeffermann, Terryn and Moura (Presentation at SAE2007 Conference), $G_{dj}$ can be modelled in two parts, one for the positive values (poverty gaps $(z - E_{dj})/z$ ) and another for the probability of positive poverty gap $P(E_{dj} < z)$. For this, let us use the Total Probability Theorem and the properties of the expectation,

$$\begin{aligned}
E[G_{dj}] &= E\left[G_{dj}|E_{dj} \geq z\right] P\left(E_{dj} \geq z\right) + E\left[G_{dj}|E_{dj} < z\right] P\left(E_{dj} < z\right) \\
&= E\left[G_{dj}|E_{dj} < z\right] P\left(E_{dj} < z\right) \\
&= \frac{z - E\left[E_{dj}|E_{dj} < z\right]}{z} P\left(E_{dj} < z\right).
\end{aligned}$$

Consider random variables $U_d$ and $V_d$ that vary over areas, normally distributed with zero means and variances $\sigma_u^2$ and $\sigma_v^2$ respectively. Then it holds that

$$E[G_{dj}|U_d = u_d, V_d = v_d] = \frac{z - E\left[E_{dj}|U_d = u_d, V_d = v_d, E_{dj} < z\right]}{z} P\left(E_{dj} < z|U_d = u_d, V_d = v_d\right).$$

We assume that $E_{dj}$ and $Y_{dj} = I(E_{dj} < z)$ given $U_d$ and $V_d$ have probability distributions

$$(E_{dj}|U_d = u_d, V_d = v_d, E_{dj} < z) \sim \text{ind } N(\mu_{dj}, \sigma^2), \quad (Y_{dj}|U_d = u_d, V_d = v_d) \sim \text{ind Bern}(p_{dj}), \tag{5.4}$$

where the mean income of people under poverty (i.e. with positive poverty gaps) $\mu_{dj}$ and the probability of being under poverty (i.e. of having positive gap) $p_{dj}$ are related to the values of $p$ explanatory variables as

$$\mu_{dj} = \mathbf{x}_{dj}\boldsymbol{\alpha} + v_d, \quad \log[p_{dj}/(1 - p_{dj})] = \mathbf{x}_{dj}\boldsymbol{\beta} + u_d, \quad j = 1, \ldots, N_d,\ d = 1, \ldots, D. \tag{5.5}$$

It seems reasonable to assume that there is correlation between the area effects associated to the mean income, $V_d$, and the area effects of the proportions, $U_d$. Thus, a possible model for these two quantities is

$$\begin{pmatrix} V_d \\ U_d \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho_{uv} \\ \rho_{uv} & \sigma_v^2 \end{pmatrix}\right).$$

However, in the simulations carried out by Pfeffermann, Terryn and Moura, the correlation between the random effects associated to each part of the model did not play an important role, and the model with independent random effects for the two parts performed almost the same as the model with correlated random effects for the two parts of the model. Thus, another approach is to consider that $V_d$ and $U_d$ are independent. This would allow us to fit the linear model for $\mu_{dj}$ and the logistic model for $p_{dj}$ separately. For these two models there are fitting procedures and software available.

A predictor of $G_d$ can then be obtained after fitting the two part model (5.4)–(5.5), which yields predictors of $\mu_{dj}$ and $p_{dj}$ given by

$$\hat{\mu}_{dj} = \mathbf{x}_{dj}\hat{\boldsymbol{\alpha}} + \hat{v}_d, \quad \hat{p}_{dj} = \exp(\mathbf{x}_{dj}\hat{\boldsymbol{\beta}} + \hat{u}_d)/[1 + \exp(\mathbf{x}_{dj}\hat{\boldsymbol{\beta}} + \hat{u}_d)], \quad j = 1, \ldots, N_d,\ d = 1, \ldots, D.$$

Then, predictors of $G_{dj}$ are $\hat{G}_{dj} = z^{-1}(z - \hat{\mu}_{dj})\hat{p}_{dj}$. Finally, the small area predictor of $G_d$ would be

$$\hat{G}_d = \frac{1}{N_d}\left[\sum_{j \in s_d} G_{dj} + \sum_{j \in r_d} \hat{G}_{dj}\right], \qquad d = 1, \ldots, D.$$

### 5.5.2 Area level model

At the area level, consider the direct estimator of the poverty gap $g_d^w = f_{1d}^w$ given in (5.2) and its sampling error (with respect to the design) $\psi_d$. Since $g_d^w$ is obtained as a mean, its distribution may not be too far from Normal distribution if the frequency of the outcome ”being under poverty” is not too low. In this case, we could use a Fay-Herriot model for this measure (Fay and Herriot, 1979). In a first stage, this model assumes that the direct estimators $g_d^w$ given the random effects $v_d$ follow a Normal distribution with mean equal to the true average poverty gap $E[g_d|u_d] = G_d$, and variance given by a known constant $\psi_d$, that is,

$$g_d^w|v_d \sim N(G_d, \psi_d), \quad d = 1, \ldots, D. \tag{5.6}$$

In a second stage, it is assumed that $G_d$ is linearly related to the values of the explanatory variables for area $d$ and the random effect $v_d$ in the form

$$G_d = \mathbf{x}_d\boldsymbol{\beta} + v_d, \quad v_d \sim \text{iid } N(0, \sigma_v^2), \quad d = 1, \ldots, D. \tag{5.7}$$

Equations (5.6) and (5.7) are equivalent to the linear mixed model

$$g_d^w = \mathbf{x}_d\boldsymbol{\beta} + v_d + e_d, \quad v_d \sim \text{iid } N(0, \sigma_v^2), \quad e_d \sim \text{ind } N(0, \psi_d), \quad d = 1, \ldots, D.$$

The variances $\psi_d$ are assumed to be known, but in practice they are replaced by the design-based sampling variances of the direct estimators $g_d^w$. Then a predictor of $G_d$ is obtained directly from the fitted model, as

$$\hat{G}_d = \mathbf{x}_d\hat{\boldsymbol{\beta}} + \hat{u}_d, \quad d = 1, \ldots, D.$$

## 5.6 Model-based estimation of the poverty severity

The poverty severity is given by

$$S_d = F_{2d} = \frac{1}{N_d}\sum_{j=1}^{N_d}\left(\frac{z - E_{dj}}{z}\right)^2 I(E_{dj} < z), \quad \alpha = 0, 1, 2, \quad d = 1, \ldots, D.$$

This can be modelled in similar way as the poverty gap. Writing

$$S_d = \frac{1}{N_d}\sum_{j=1}^{N_d} S_{dj},$$

and decomposing the expectation of $S_{dj}$ as

$$E[S_{dj}] = E\left[\left(\frac{z - E_{dj}}{z}\right)^2 |E_{dj} < z\right] P\left(E_{dj} < z\right).$$

If $(E_{dj}|E_{dj} < z)$ follows approximately a Normal distribution, then, the random variable

$$L_{dj} = \left(\frac{z - E_{dj}}{z}\right)^2 |E_{dj} < z$$

has a non-centered chi-squared distribution. A transformation of this variable such as the logarithm, could make the transformed variable closer to a normally distributed variable. Then, the linear model would be stated for $M_{dj} = \log(L_{dj})$. From this model we would obtain a predictor $\hat{M}_{dj}$, which leads to a predictor of $L_{dj}$ as $\hat{L}_{dj} = \exp(\hat{M}_{dj})$. Bias correction of this predictor is given in the literature (Molina, 2008). Area level models can be also stated for these measures.

## 5.7 Empirical Best prediction of FGT poverty measures at unit level – Approach 2

### 5.7.1 Best prediction under a finite population

In this section we introduce the best predictor (BP) of a function of a random vector in a finite population. Then, in the next section we describe the application of the BP methodology for estimating FGT poverty measures in small areas.

Consider a random vector $\mathbf{y}$ containing the values of a random variable in the units of a finite population. Let $\mathbf{y}_s$ be the sub-vector of $\mathbf{y}$ corresponding to sample elements and $\mathbf{y}_r$ the sub-vector of out-of-sample elements, that is, $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$. The target is to predict the value of a real function $\delta = h(\mathbf{y})$ of the random vector $\mathbf{y}$ using sample data $\mathbf{y}_s$. For a particular predictor $\hat{\delta}$, the mean squared error is defined as

$$MSE(\hat{\delta}) = E_{\mathbf{y}}[(\hat{\delta} - \delta)^2], \tag{5.8}$$

where $E_{\mathbf{y}}$ denotes expectation with respect to the joint distribution of the population vector $\mathbf{y}$. The BP of $\delta$ is the function of $\mathbf{y}_s$ which minimises (5.8). Consider the conditional expectation $\delta^0 = E_{\mathbf{y}_r}(\delta|\mathbf{y}_s)$, where the expectation is taken with respect to the joint distribution of $\mathbf{y}_r$ and the result is a function of sample data $\mathbf{y}_s$. Subtracting and adding $\delta^0$ in the mean squared error, we obtain

$$\begin{aligned} MSE(\hat{\delta}) &= E_{\mathbf{y}}[(\hat{\delta} - \delta^0 + \delta^0 - \delta)^2] \\ &= E_{\mathbf{y}}[(\hat{\delta} - \delta^0)^2] + 2\, E_{\mathbf{y}}[(\hat{\delta} - \delta^0)(\delta^0 - \delta)] + E_{\mathbf{y}}[\delta^0 - \delta)^2] \end{aligned}$$

In this expression, the last term does not depend on $\hat{\delta}$. For the second term, observe that

$$\begin{aligned} E_{\mathbf{y}}[(\hat{\delta} - \delta^0)(\delta^0 - \delta)] &= E_{\mathbf{y}_s}\left\{E_{\mathbf{y}_r}\left[(\hat{\delta} - \delta^0)(\delta^0 - \delta)|\mathbf{y}_s\right]\right\} \\ &= E_{\mathbf{y}_s}\left\{(\hat{\delta} - \delta^0)\left[\delta^0 - E_{\mathbf{y}_r}(\delta|\mathbf{y}_s)\right]\right\} \\ &= 0. \end{aligned}$$

Thus, the BP is the value $\hat{\delta}$ that minimises $E_{\mathbf{y}}[(\hat{\delta} - \delta^0)^2]$. Since this quantity is non-negative with its minimum value at zero, the BP of $\delta$ is

$$\hat{\delta} = \delta^0 = E_{\mathbf{y}_r}(\delta|\mathbf{y}_s). \tag{5.9}$$

Consider that the vector for the finite population $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$ follows a Normal distribution with mean vector $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ for a known matrix $\mathbf{X}$ with sample and out-of-sample decomposition $\mathbf{X} = (\mathbf{X}_s', \mathbf{X}_r')'$, and positive definite covariance matrix $\mathbf{V}$ decomposed accordingly as

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{pmatrix}.$$

Assume that the quantity of interest $\delta$ is a linear function of $\mathbf{y}$, that is, $\delta = \mathbf{a}'\mathbf{y}$, where $\mathbf{a} = (\mathbf{a}_s', \mathbf{a}_r')'$. The BLUP of $\delta = \mathbf{a}_s'\mathbf{y}_s + \mathbf{a}_r'\mathbf{y}_r$ under a finite population (Royall, 1976), is given by

$$\tilde{\delta} = \mathbf{a}_s'\mathbf{y}_s + \mathbf{a}_r'\left[\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})\right], \tag{5.10}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{y}_s$ is the BLUE of $\boldsymbol{\beta}$ as defined in (1.2). If we replace $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ in the BP (5.9), then the result is equal to the EBLUP of Royall (1976) given in (5.10).

### 5.7.2 Best prediction of FGT poverty measures

In this section we describe how to compute BPs of the FGT poverty measures for small areas. Let us define the random variables

$$F_{dj} = \left(\frac{z - E_{dj}}{z}\right)^{\alpha} I(E_{dj} < z), \quad \alpha = 0, 1, 2, \quad d = 1, \ldots, D.$$

Then the family of FGT poverty measures for small area $d$ is the mean over area $d$ of these random variables

$$F_{\alpha d} = \frac{1}{N_d}\sum_{j=1}^{N_d} F_{dj}, \quad \alpha = 0, 1, 2, \quad d = 1, \ldots, D. \tag{5.11}$$

Suppose that we know the distribution of a one-to-one transformation $Y_{dj} = T(E_{dj})$ of the welfare variables $E_{dj}$, $j = 1, \ldots, N_d$, $d = 1, \ldots, D$. Let $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')'$ be the vector containing the values of the transformed variables $Y_{dj}$ for the sample and out-of-sample units. We assume that

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V}),$$

where the mean vector $\boldsymbol{\mu}$ and the variance matrix $\mathbf{V}$ can be partitioned in submatrices corresponding to sample and out-of-sample elements

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_s \\ \boldsymbol{\mu}_r \end{pmatrix} \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_s & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_r \end{pmatrix}.$$

Then the variables $F_{dj}$ are function of the transformed variables $Y_{dj}$, and the expression of $F_{dj}$ in terms of $Y_{dj}$ is

$$F_{dj} = \left(\frac{z - T^{-1}(Y_{dj})}{z}\right)^{\alpha} I(T^{-1}(Y_{dj}) < z) = h(Y_{dj}), \quad \alpha = 0, 1, 2, \quad d = 1, \ldots, D.$$

Thus, the FGT poverty measure is a non-linear function of the vector $\mathbf{y}$. Then the BP of $F_{\alpha d}$ is given by

$$\hat{F}_{\alpha d} = E[F_{\alpha d}|\mathbf{y}_s]. \tag{5.12}$$

Using the decomposition of the mean (5.11) in terms of sample and out-of-sample elements,

$$F_{\alpha d} = \frac{1}{N_d}\left[\sum_{j\in s_d} F_{dj} + \sum_{j\in r_d} F_{dj}\right],$$

and taking conditional expectation, the BP becomes

$$\hat{F}_{\alpha d} = \frac{1}{N_d}\left[\sum_{j\in s_d} F_{dj} + \sum_{j\in r_d} \hat{F}_{dj}\right],$$

where $\hat{F}_{dj} = E(F_{dj}|\mathbf{y}_s)$ is also the BP of the out-of-sample variable $F_{dj} = h(Y_{dj})$, which is defined as

$$\hat{F}_{dj} = E[h(Y_{dj})|\mathbf{y}_s] = \int_{I\!R} h(Y_{dj}) f(Y_{dj}|\mathbf{y}_s)\, dY_{di}, \quad j \in r_d, \quad d = 1, \ldots, D.$$

where $f(Y_{dj}|\mathbf{y}_s)$ is the density of $Y_{dj}$ given the data vector $\mathbf{y}_s$. The distribution of the vector of out-of-sample data $\mathbf{y}_r$ given the sample data $\mathbf{y}_s$ is

$$\mathbf{y}_r|\mathbf{y}_s \sim N(\boldsymbol{\mu}_{r|s}, \mathbf{V}_{r|s}), \tag{5.13}$$

where

$$\boldsymbol{\mu}_{r|s} = \boldsymbol{\mu}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \boldsymbol{\mu}_s), \quad \mathbf{V}_{r|s} = \mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{V}_{sr}.$$

However, there is no explicit expression for the expectation in (5.7.2) because $F_{dj} = h(Y_{dj})$ is a complex non-linear function of $Y_{dj}$. This expectation can be approximated empirically, by generating a large number $L$ of vectors $\mathbf{y}_r$ from (5.13). Let $Y_{dj}^{(\ell)}$ be the value of the out-of-sample observation $Y_{dj}$, $j \in r_d$, obtained in the $\ell$-th generation. An approximation to the best predictor of $Y_{dj}$ is

$$\hat{F}_{dj} = E[h(Y_{dj})|\mathbf{y}_{ds}] \approx \frac{1}{L}\sum_{\ell=1}^{L} h(Y_{dj}^{(\ell)}).$$

In practice, the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{V}$ usually depend on an unknown vector of parameters $\boldsymbol{\theta}$. This means that the conditional density $f(Y_{dj}|\mathbf{y}_s)$ depends on $\boldsymbol{\theta}$, that is, $f(Y_{dj}|\mathbf{y}_s, \boldsymbol{\theta})$. We can take an estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ such as the maximum likelihood (ML) estimator. Then the expectation can be approximated by generating values from the estimated density $f(Y_{dj}|\mathbf{y}_s, \hat{\boldsymbol{\theta}})$. The resulting predictor, denoted $\hat{F}_{dj}^E$, is called empirical best predictor (EBP). Finally, the EBP of $F_{\alpha d}$ is

$$\hat{F}_{\alpha d}^E = \frac{1}{N_d}\left[\sum_{j \in s_d} F_{dj} + \sum_{j \in r_d} \hat{F}_{dj}^E\right].$$

A possible model for the elements of the vector $\mathbf{y}$ is the nested error regression model

$$Y_{dj} = \mathbf{x}_{dj}\boldsymbol{\beta} + v_d + e_{dj}, \quad v_d \sim \text{iid } N(0, \sigma_v^2) \quad e_{dj} \sim \text{iid } N(0, \sigma_e^2), \quad j = 1, \ldots, N_d, \quad d = 1, \ldots, D,$$

where the random effects $v_d$ and the random errors $e_{dj}$ are independent. Let us define vectors and matrices containing the elements for area $d$,

$$\mathbf{y}_d = \underset{1 \le j \le N_d}{\text{col}}(Y_{dj}), \quad \mathbf{e}_d = \underset{1 \le j \le N_d}{\text{col}}(e_{dj}), \quad \mathbf{X}_d = \underset{1 \le j \le N_d}{\text{col}}(\mathbf{x}_{dj}),$$

and vectors and matrices containing all population elements

$$\mathbf{y} = \underset{1 \le d \le D}{\text{col}}(\mathbf{y}_d), \quad \mathbf{e} = \underset{1 \le d \le D}{\text{col}}(\mathbf{e}_d), \quad \mathbf{X} = \underset{1 \le d \le D}{\text{col}}(\mathbf{x}_d), \quad \mathbf{Z} = \underset{1 \le d \le D}{\text{diag}}(\mathbf{1}_{N_d}).$$

Then, the unit level model can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \quad \mathbf{v} \sim N(\mathbf{0}, \sigma_v^2 \mathbf{I}_D), \quad \mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_N),$$

where $\mathbf{v}$ is independent of $\mathbf{e}$. Under this model, the mean vector and the covariance matrix of $\mathbf{y}$ are

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \quad \mathbf{V} = \sigma_v^2 \mathbf{Z}\mathbf{Z}' + \sigma_e^2 \mathbf{I}_N,$$

and the distribution of the out-of-sample vector $\mathbf{y}_r$ given the data vector $\mathbf{y}_s$ is given in (5.13).

The advantages of this procedure are that it can be used to predict any function $h(\mathbf{y})$ and it requires fewer model assumptions than Approach 1 of Section 5.5.1.

## 5.8 Hierarchical Bayes estimation of FGT poverty measures – Approach 3

Consider again that the distribution of the transformed variables $Y_{dj} = T(E_{dj})$ is known and let us decompose the FGT poverty measure of order $\alpha$ for area $d$ in the sample and out-of-sample part,

$$F_{\alpha d} = \frac{1}{N_d} \left[ \sum_{j \in s_d} \left( \frac{z - T^{-1}(Y_{dj})}{z} \right)^{\alpha} I(T^{-1}(Y_{dj}) < z) + \sum_{j \in r_d} \left( \frac{z - T^{-1}(Y_{dj})}{z} \right)^{\alpha} I(T^{-1}(Y_{dj}) < z) \right].$$

Here, the first term is observed but the second term is unobserved. Thus, we need to predict the value of $Y_{dj}$ for out-of-sample units $j \in r_d$. For this, we assume that the values of $p$ auxiliary variables are known for all population units and that $Y_{dj}$ satisfies the nested-error regression model defined as

$$Y_{dj} | u_d, \rho, \sigma^2 \sim \text{ind } N(\mathbf{x}_{dj}\boldsymbol{\beta} + u_d, \sigma^2) \tag{5.14}$$

$$u_d | \rho, \sigma^2 \sim \text{ind } N\left(0, \frac{\rho}{1-\rho}\sigma^2\right), \quad j = 1, \ldots, N_d, \ d = 1, \ldots, D. \tag{5.15}$$

In (5.15), the random effects variance, $\sigma_u^2$, has been reparametrized in terms of the intraclass correlation coefficient $\rho \in (0,1)$, as $\sigma_u^2 = \rho\,\sigma^2/(1-\rho)$. For parameters $(\boldsymbol{\beta}', \sigma^2, \rho)$, let us consider the standard joint noninformative prior based on the Jeffreys rule. Assuming independence among parameters, this prior is

$$\pi(\boldsymbol{\beta}, \sigma^2, \rho) \propto \frac{1}{\sigma^2}. \tag{5.16}$$

It can be shown that the posterior density

$$\pi(\mathbf{u}, \boldsymbol{\beta}, \sigma^2, \rho | \mathbf{y}_s) = \pi_1(\mathbf{u} | \boldsymbol{\beta}, \sigma^2, \rho, \mathbf{y}_s)\, \pi_2(\boldsymbol{\beta} | \sigma^2, \rho, \mathbf{y}_s)\, \pi_3(\sigma^2 | \rho, \mathbf{y}_s)\, \pi_4(\rho | \mathbf{y}_s) \tag{5.17}$$

is proper as long as the matrix $\mathbf{X} = \text{col}_{1 \le d \le D} \text{col}_{j \in s_d}(\mathbf{x}_{dj})$ has full column rank.

The posterior distributions $\pi_1(\mathbf{u}|\boldsymbol{\beta}, \sigma^2, \rho, \mathbf{y}_s)$, $\pi_2(\boldsymbol{\beta}|\sigma^2, \rho, \mathbf{y}_s)$ and $\pi_3(\sigma^2|\rho, \mathbf{y}_s)$ are normal or inverse gamma distributions; however, $\pi_4(\rho|\mathbf{y}_s)$ is not a standard distribution. These posterior distributions can be simulated by using a Gibbs sampling scheme with a Metropolis-Hasting step in the case of parameter

$\rho$. It is necessary to study the convergence of chains in order to decide the number of burn-in iterations. Usually, in general linear models, not more than 25,000 iterations are necessary to reach convergence. After that, samples from the posterior distributions are derived (e.g. 25,000 iterations).

Consider the vector $\boldsymbol{\theta} = (\mathbf{u}', \boldsymbol{\beta}', \sigma^2, \rho)$. Now, since model (5.14) holds for all population units, then the out-of-sample units satisfy

$$Y_{dj}|\mathbf{y}_s, \boldsymbol{\theta} \sim \text{iid } N(\mathbf{x}_{dj}\boldsymbol{\beta} + u_d, \sigma^2), \quad j \in r_d, \quad d = 1, \ldots, D.$$

Moreover, the predictive posterior density of a non-sample observation $Y_{dj}$ $j \in r_d$, given the sample data, is

$$f(Y_{dj}|\mathbf{y}_s) = \int_{\boldsymbol{\theta}} f(Y_{dj}|\mathbf{y}_s, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}|\mathbf{y}_s)\, d\boldsymbol{\theta}, \quad j \in r_d,\ d = 1, \ldots, D.$$

Thus, for each sample from the posterior distribution of $\boldsymbol{\theta}$, we can draw values from the predictive posterior distribution of $Y_{dj}$ and we can compute the poverty measure

$$F_{\alpha d} = \frac{1}{N_d}\left[\sum_{j\in s_d}\left(\frac{z - T^{-1}(Y_{dj})}{z}\right)^{\alpha} I(T^{-1}(Y_{dj}) < z) + \sum_{j\in r_d}\left(\frac{z - T^{-1}(Y_{dj}^{(h)})}{z}\right)^{\alpha} I((Y_{dj}^{(h)}) < z)\right].$$

In this way, we can also compute posterior summaries (means, variances, HPD intervals) from the generated values.

## 5.9 References

Ballini, F., Betti, G. and Neri, L. (2006). Poverty and inequality mapping in the Commonwealth of Dominica. Preprint.

Bell, W. (1997). Models for county and state poverty estimates. Preprint, Census Statistical Research Division.

Elbers, C., Lanjouw, J. O. and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, **71**, 355–364.

Fay, R. E. and Herriot, R. A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures, *Econometrica*, **52**, 761–766.

Maiti, T. and Slud, E.V. (2002). Comparison of small area models in SAIPE. Preprint.

Molina, I. (2008). Uncertainty under a multivariate nested-error regression model with logarithmic transformation, *Journal of Multivariate Analysis*, **100**, 963–980.

Molina, I., Saei, A. and Lombarda, M. J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model *Journal of the Royal Statistical Society A*, **170**, 975–1000.

Molina, I., Salvati, N. and Pratesi, M. (2008). Bootstrap for estimating the MSE of the Spatial EBLUP. Under second revision in *Computational Statistics*.

Neri, L., Ballini, F. and Betti, G. (2005). Poverty and inequality in transition countries. *Statistics in Transition*, **7**, 135–157.

Pfeffermann, D., Terryn, B. and Moura, F. (2007). *Bayesian small area estimation of literacy under a two part random effects model*, Presentation at SAE2007 Conference held in Pisa.

Rao, J. N. K. (2003). *Small Area Estimation*. London: Wiley.

Tarozzi, A. and Deaton, A. (2007). Using census and survey data to estimate poverty and inequality for small areas. Preprint.

Royall, R. M. (1976). The Linear Least Squares Prediction Approach to Two-Stage Sampling, *Journal of the American Statistical Association*, **71**, 657–664.

# Chapter 6

# Quantile / M-quantile Models for Small Area Estimation

## 6.1 Introduction

In the previous chapters of this review the "classical" methods for small area estimation, based on linear and generalized linear mixed models, were introduced. In this chapter we review a new approach to small area estimation that is based on M-quantile models (Chambers and Tzavidis, 2006). The structure of this chapter is as follows. In section 6.2 we review quantile regression, in section 6.3 we introduce M-quantile regression and in section 6.4 nonparametric M-quantile regression. In section 6.5 we describe how quantile or M-quantile models can be employed for measuring area effects and estimators of small area averages. In section 6.6 we discuss mean squared error estimation for M-quantile small area predictors. In sections 6.7 and 6.8 we discuss how the M-quantile approach can be adapted for borrowing strength over space in small area estimation when spatial information is available. Finally, in section 6.9 we describe the nonparametric approach M-quantile small area estimation.

## 6.2 Quantile regression

The classical theory of linear statistical models is fundamentally a theory of conditional expectations. That is, a regression model summarises the behaviour of the mean of $y$ at each point in a set of $x$'s (Mosteller and Tukey, 1977). This summary provides a rather incomplete picture, in much the same way as the mean gives an incomplete picture of a distribution. It is usually much better to fit a family of regression models, each one summarising the behaviour of a different percentage point (quantile) of $y$ at each point in this set of $x$'s. Such a modelling exercise is referred to as quantile regression.

Let $\Omega = \{1, \dots, N\}$ be a finite population. Let $\mathbf{y} = (y_1, \dots, y_N)'$ denote the variable values for the $N$ population elements. We consider a sample $s \subset \Omega$, of $n \leq N$ units, and we denote with $r = \Omega - s$ the set of non sampled units. Let $\mathbf{X}_{n \times p}$ be a matrix of $p$ auxiliary variables, and for each population unit $j$ let $\mathbf{x}_j = (x_{1j}, \dots, x_{pj})$ be the vector corresponding to the $j$-th row of matrix $\mathbf{X}$.

In a seminal paper Koenker and Bassett (1978) developed the idea of quantile regression. In the linear case, quantile regression leads to a family (or "ensemble") 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 $x$. For example, when $q = 0.5$ the quantile regression line shows how the median of this conditional distribution changes with $x$. A linear model for the $q^{th}$ conditional quantile of $y$ given the covariates $\mathbf{X}$ is $Q_q(x) = \mathbf{X}\boldsymbol{\beta}_q$, where $\boldsymbol{\beta}_q = (\beta_1, \ldots, \beta_p)'_q$. Inclusion of the intercept is straightforward specifying $\mathbf{x}_1 = \mathbf{1}_n$. The vector $\boldsymbol{\beta}_q$ is estimated by minimising $\sum_{j=1}^{n} |y_j - \mathbf{x}_j\mathbf{b}| \{(1-q)I(y_j - \mathbf{x}_j\mathbf{b} \leq 0) + qI(y_j - \mathbf{x}_j\mathbf{b} > 0)\}$ with respect to $\mathbf{b}$. Solutions to this minimisation problem are usually obtained using linear programming methods (Koenker and D'Orey, 1987) and functions for performing quantile regression analysis now exist in standard statistical software, e.g. the R statistical package (R Development Core Team, 2005).

## 6.3 M-quantile regression

Quantile regression can be viewed as a generalisation of median regression. In the same way, expectile regression (Newey and Powell, 1987) is a "quantile-like" generalisation of mean (i.e. standard) regression. M-quantile regression (Breckling and Chambers, 1988) integrates these concepts within a common framework defined by a "quantile-like" generalisation of regression based on influence functions (M-regression).

The M-quantile of order $q$ for the conditional density of $y$ given $\mathbf{X}$ is defined as the solution $Q_q(x;\psi)$ of the estimating equation $\int \psi_q(y-Q)f(y|\mathbf{X})dy = 0$, where $\psi$ denotes the influence function associated with the M-quantile. A linear M-quantile regression model is one where we assume that $Q_q(x;\psi) = \mathbf{X}\boldsymbol{\beta}_\psi(q)$. That is, we allow a different set of regression parameters for each value of $q$. For specified $q$ and $\psi$, estimates of these regression parameters can be obtained by solving the estimating equations

$$\sum_{j=1}^{n} \psi_q(r_{jq\psi})\mathbf{x}_j = 0 \tag{6.1}$$

where $r_{jq\psi} = y_j - \mathbf{x}_j\boldsymbol{\beta}_\psi(q)$, $\psi_q(r_{iq\psi}) = 2\psi\{s^{-1}r_{iq\psi}\}\ \{qI(r_{jq\psi} > 0) + (1-q)I(r_{jq\psi} \leq 0)\}$ and $s$ is a suitable robust estimate of scale, e.g. the MAD estimate $s = median\,|r_{jq\psi}|\,/0.6745$. The division by $0.6745$ is made so that for normally distributed data, $s$ is an estimate of the standard deviation. A popular choice for the influence function is the Huber Proposal 2, $\psi(u) = uI(-c \leq u \leq c) + c\,\mathrm{sgn}(u)$. However, other influence functions are also possible. Provided $c$ is bounded away from zero, straightforward modification of widely available iteratively reweighted least squares software for fitting robust regression models (e.g. *rlm* in R, see Venables and Ripley, 2002) then leads to a solution of (6.1).

## 6.4 Nonparametric M-quantile regression

M-quantile models do not depend on strong distributional assumptions nor on a predefined hierarchical structure, and outlier robust inference is automatically performed when these models are fitted. However, M-quantile regression assumes that the quantiles of the distribution are some known parametric function of the covariates. When the functional form of the relationship between the $q^{th}$ quantile and the covariates deviates from the assumed one, the traditional M-quantile regression can lead to biased estimators of the small area parameters. Pratesi et al. (2006) extended M-quantile regression to nonparametric modeling via penalized splines. Penalized splines (or $p$-splines) regression is a flexible smoothing technique popularized by Eilers and Marx (1996). Ruppert et al. (2003) provide a thorough treatment of $p$-splines and their applications. Bollaerts et al. (2006) introduce quantile regression based on $p$-splines to estimate quantile growth curves and quantile antibody levels as a function of age. Lee and Oh (2007), independently of Pratesi et al. (2006), use M-regression to make $p$-splines robust against outliers. Using $p$-splines for M-quantile regression, beyond having the properties of M-quantile models, allows for dealing with an undefined functional relationship that can be estimated from the data. When the relationship between the $q^{th}$ quantile and the covariates is not linear, a $p$-splines M-quantile regression model may have significant advantages compared to the linear M-quantile model.

The nonparametric specification of the conditional M-quantile of $y$ given $\mathbf{X}$ can be summarized as follows. Given an influence function $\psi$, a nonparametric model with one covariate $x_1$ for the $q^{th}$ quantile can be written as $Q_q(x_1, \psi) = \tilde{m}_{\psi,q}(x_1)$, where the function $\tilde{m}_{\psi,q}(\cdot)$ is unknown, but assumed to be approximated sufficiently well by the following function

$$m_{\psi,q}[x_{1j}; \boldsymbol{\beta}_\psi(q), \boldsymbol{\gamma}_\psi(q)] = \beta_{1\psi}(q)x_{1j} + \ldots + \beta_{p\psi}(q)x_{1j}^p + \sum_{k=1}^{K} \gamma_{k\psi}(q)(x_{1j} - \kappa_k)_+^p, \tag{6.2}$$

where $p$ is the degree of the spline, $(t)_+^p = t^p$ if $t > 0$ and 0 otherwise, $\kappa_k$ for $k = 1, \ldots, K$ is a set of fixed knots, $\boldsymbol{\beta}_\psi(q) = (\beta_{1\psi}(q), \ldots, \beta_{p\psi}(q))'$ is the coefficient vector of the parametric portion of the model and $\boldsymbol{\gamma}_\psi(q) = (\gamma_{1\psi}(q), \ldots, \gamma_{K\psi}(q))'$ is the coefficient vector for the spline one. The latter portion of the model allows for handling nonlinearities in the structure of the relationship. If the number of knots $K$ is sufficiently large, the class of functions in (6.2) is very large and can approximate most smooth functions. In particular, in the $p$-splines context, a knot is placed every 4 or 5 observations at uniformly spread quantiles of the unique values of $x_1$. The spline model (6.2) uses a truncated polynomial spline basis to approximate the function $\tilde{m}_{\psi,q}(\cdot)$. Other bases can be used; more details on bases and knots choice can be found in Ruppert et al. (2003).

The influence of the knots is limited by putting a constraint on the size of the spline coefficients: typically $\sum_{k=1}^{K} \gamma_{k\psi}^2(q)$ is bounded by some constant, while the parametric coefficients $\boldsymbol{\beta}_\psi(q)$ are left unconstrained. Therefore, estimation can be accommodated by mimicking penalization of an objective function and solving the following set of estimating equations

$$\sum_{j=1}^{n} \psi_q(y_j - \mathbf{x}_j\boldsymbol{\beta}_\psi(q) - \mathbf{z}_j\boldsymbol{\gamma}_\psi(q))(\mathbf{x}_j, \mathbf{z}_j)' + \lambda \begin{bmatrix} \mathbf{0}_{(1+p)} \\ \boldsymbol{\gamma}_\psi(q) \end{bmatrix} = \mathbf{0}_{(1+p+K)}, \tag{6.3}$$

where $\mathbf{x}_j$ is the $j$-th row of the $\mathbf{X}$ matrix while $\mathbf{z}_j$ is the $j$-th row of the $n \times K$ matrix

$$\mathbf{Z} = \begin{bmatrix} (x_{11} - \kappa_1)_+^p & \cdots & (x_{11} - \kappa_K)_+^p \\ \vdots & \ddots & \vdots \\ (x_{1n} - \kappa_1)_+^p & \cdots & (x_{1n} - \kappa_K)_+^p \end{bmatrix},$$

and $\lambda$ is a Lagrange multiplier that controls the level of smoothness of the resulting fit.

An algorithm based on iteratively reweighted penalized least squares is proposed in Pratesi et al. (2006) to effectively compute the parameter estimates. Once those estimates are obtained, $\hat{m}_{\psi,q}[x_1] = m_{\psi,q}[x_1; \hat{\boldsymbol{\beta}}_{\psi}(q), \hat{\gamma}_{\psi}(q)]$ can be computed as an estimate for $Q_q(x_1, \psi)$.

Extension to bivariate smoothing can be handled by assuming $Q_q(x_1, x_2, \psi) = \tilde{m}_{\psi,q}(x_1, x_2)$. This is of central interest in a number of application areas as environment and public health. It has particular relevance when referenced responses need to be converted to maps.

In particular, the following model is assumed at quantile $q$ for unit $j$:

$$m_{\psi,q}[x_{1j}, x_{2j}; \boldsymbol{\beta}_{\psi}(q), \boldsymbol{\gamma}_{\psi}(q)] = \beta_{1\psi}(q)x_{1j} + \beta_{2\psi}(q)x_{2j} + \mathbf{z}_j\boldsymbol{\gamma}_{\psi}(q). \tag{6.4}$$

Here $\mathbf{z}_j$ is the $j$-th row of the following $n \times K$ matrix

$$\mathbf{Z} = [C(\tilde{\mathbf{x}}_j - \boldsymbol{\kappa}_k)]_{\substack{1 \leqslant j \leqslant n \\ 1 \leqslant k \leqslant K}} [C(\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'})]^{-1/2}_{1 \leqslant k \leqslant K}, \tag{6.5}$$

where $C(\mathbf{t}) = ||\mathbf{t}||^2 \log ||\mathbf{t}||$, $\tilde{\mathbf{x}}_j = (x_{1j}, x_{2j})$ and $\boldsymbol{\kappa}_k$, $k = 1, \ldots, K$ are knots. See Pratesi et al. (2006) for details on this. Here, it is enough to note that the estimation procedure can again be pursued with (6.3) where $\mathbf{x}_j = (1, \tilde{\mathbf{x}}_j)$.

The choice of knots in two dimensions is more challenging than in one. Two solutions suggested in literature that provide a subset of observations nicely scattered to cover the domain are *space filling designs* (Nychka and Saltzman, 1998) and the *clara* algorithm (Kaufman and Rousseeuw, 1990, Chapter 3). The first one is based on the maximal separation principle of $K$ points among the unique $\tilde{\mathbf{x}}_i$ and is implemented in the `fields` package of the R language (R Development Core Team, 2005). The second one is based on clustering and selects $K$ representative objects out of $n$; it is implemented in the package `cluster` of R.

It should be noted, then, that the estimating equations in (6.3) can be used to handle univariate smoothing and bivariate smoothing by suitably changing the parametric and the spline part of the model, i.e. once the $\mathbf{X}$ and the $\mathbf{Z}$ matrices are set up. Finally, other continuous or categorical variables can be easily inserted parametrically in the model by adding columns to the $\mathbf{X}$ matrix. This allows for semiparametric modeling, as intended in Ruppert et al. (2003), to be inherited and applied to M-quantile regression.

## 6.5 Small area estimation with M-quantile models

Mixed effects models assume that variability associated with the conditional distribution of $y$ given $x$ can be at least partially explained by a pre-specified hierarchical structure, e.g. the small areas of interest. As we saw in the previous sections, an alternative approach to modelling the variability in this conditional distribution is via M-quantile regression, which does not depend on a hierarchical structure. Let us index population units only by $j$ in what follows and, following Kokic et. al. (1997) and Aragon et. al. (2005), characterize conditional variability across the population of interest by the M-quantile coefficients of the population units. For unit $j$ with values $y_j$ and $\mathbf{x}_j$, this coefficient is the value $q_j$

such that $Q_{q_j}(\mathbf{x}_j;\psi) = y_j$. Note that these M-quantile coefficients are determined at population level. Consequently, if a hierarchical structure does explain part of the variability in the population data, then we expect units within clusters defined by this hierarchy to have similar M-quantile coefficients. By definition,

$$\overline{Y}_d = N_d^{-1}\left(\sum_{j\in s_d} y_j + \sum_{j\in r_d} \mathbf{x}_j\boldsymbol{\beta}_\psi(\theta_d)\right) + N_d^{-1}\sum_{j\in r_d}\mathbf{x}_j\left[\boldsymbol{\beta}_\psi(q_j) - \boldsymbol{\beta}_\psi(\theta_d)\right] \tag{6.6}$$

when the conditional M-quantiles follow a linear model. Here $\theta_d = N_d^{-1}\sum_{j\in d} q_j$ is the average value of the M-quantile coefficients of the units in area $d$, with $d = (1, \ldots, D)$, and $s_d$, $r_d$ respectively denote the sampled and non-sampled units in area $d$. Typically the first term on the right hand side of (6.6) will dominate, suggesting a predictor of small area average of the form

$$\hat{\overline{Y}}_d = N_d^{-1}\left(\sum_{j\in s_d} y_j + \sum_{j\in r_d} \mathbf{x}_j\boldsymbol{\beta}_\psi(\hat{\theta}_d)\right). \tag{6.7}$$

We refer to $\theta_d$ as the M-quantile coefficient of area $d$ in what follows. Irrespective of how the M-quantile coefficient $\theta_d$ for area $d$ is defined, (6.7) is equivalent to using $\mathbf{x}_j\boldsymbol{\beta}_\psi(\theta_d)$ to predict the unobserved value $y_j$ for population unit $j \in r_d$. This suggests that predicted values for other small area characteristics can be also calculated using these unit level predictions. Approaches to estimating the distribution function of the characteristic of interest at small area level are described in Chapter 7 of this review.

In order to compute (6.7) we need the estimated M-quantile coefficient for area $d$, i.e. $\hat{\theta}_d$. Such an estimate will depend on the sample M-quantile coefficients, which we denote by $\{q_j; j \in s\}$, and which characterize the variation in the conditional distribution of $y$ given $\mathbf{X}$ in the sample in exactly the same way as the $q_j$ characterize this distribution in the population. In order to calculate the $q_j$, we define a fine grid on the (0,1) interval, and use the sample data to fit M-quantile regression lines at each value $q$ on this grid. The required $q_j$ values are then obtained by linear interpolation over this grid. Provided the sampling method is non-informative given $\mathbf{x}$, $\hat{\theta}_d$ can be calculated as the mean of the $q_j$ values in area $d$. This is appropriate if $\theta_d$ is defined as the mean value of the population $q_j$ values in area $d$. If a more robust definition of $\theta_i$ is employed, say the median of these population values, then $\hat{\theta}_d$ can be calculated as the median of the $q_j$ values in area $d$. Given a finite population $\Omega$, the area specific empirical distribution function of $y$ for area $d$ is

$$F_d(t) = N_d^{-1}\left\{\sum_{j\in s_d} I(y_j \le t) + \sum_{j\in r_d} I(y_j \le t)\right\}. \tag{6.8}$$

The problem of predicting $F_d(t)$ essentially reduces to predicting the values $y_j$ for the non-sampled units in small area $d$. One straightforward way of achieving this is to simply replace the unknown non-sample values of $y$ in (6.8) by their predicted values $\hat{y}_j$ under an appropriate model, leading to a predictor of (6.8) of the form

$$\hat{F}_d(t) = N_d^{-1}\left\{\sum_{j\in s_d} I(y_j \le t) + \sum_{j\in r_d} I(\hat{y}_j \le t)\right\}. \tag{6.9}$$

A predictor of the mean $\overline{Y}_d$ of $y$ in area $d$ is then defined by the value of the mean functional defined by (6.9). This leads to the usual plug-in predictor of the mean,

$$\hat{\overline{Y}}_d = \int_{-\infty}^{\infty} td\hat{F}(t) = N_d^{-1}\left(\sum_{j\in s_d} y_j + \sum_{j\in r_d} \hat{y}_j\right).$$

It immediately follows that the EBLUP is the mean functional defined by (6.9) when $\hat{y}_j = \mathbf{x}_j\hat{\boldsymbol{\beta}} + \mathbf{z}_j\hat{\mathbf{u}}_d$ (see Section 2.2 of this review), while the M-quantile predictor (6.7) is also a mean functional corresponding to (6.9) but now with $\hat{y}_j = \mathbf{x}_j\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d)$. In both cases the predicted value of a non-sample unit $j$ in area $d$ corresponds to an estimate $\hat{\mu}_j$ of its expected value given that it is located in area $d$.

Tzavidis and Chambers (2007) note that the M-quantile predictor (6.7) can be biased and propose an alternative estimator based on the distribution function estimator by Chambers and Dunstan (1986) (CD hereafter). In the context of the small area distribution function (6.8), and assuming that the residuals $\epsilon_j = y_j - \mu_j$ are homoskedastic within the small area of interest (an assumption satisfied by the linear mixed model), this is of the form

$$\hat{F}_d^{CD}(t) = N_d^{-1}\left[\sum_{j\in s_d} I(y_j \le t) + n_d^{-1}\sum_{j\in s_d}\sum_{k\in r_d} I\left\{\hat{\mu}_k + (y_j - \hat{\mu}_j) \le t\right\}\right]. \tag{6.10}$$

It can be shown that the mean functional defined by (6.10) takes the value

$$\hat{\overline{Y}}_d^{CD} = \int_{-\infty}^{\infty} td\hat{F}_d^{CD}(t) = N_d^{-1}\left\{\sum_{j\in s_d} y_j + \sum_{j\in r_d}\hat{\mu}_j + (f_d^{-1} - 1)\sum_{j\in s_d}(y_j - \hat{\mu}_j)\right\} \tag{6.11}$$

where $f_j = n_d N_d^{-1}$ is the sampling fraction in area $d$. Under a linear M-quantile approach to small area estimation, substituting $\hat{\mu}_{jd} = \mathbf{x}_j\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d)$ in (6.11) then defines a bias-adjusted predictor for $\overline{Y}_d$ that is an alternative to (6.7).

Wang and Dorfman (1996) point out that the CD predictor (6.10) is model-consistent but design-inconsistent. An alternative to this predictor that is both design-consistent and model-consistent has been proposed by Rao et al. (1990). Under simple random sampling the Rao-Kovar-Mantel predictor of the finite population distribution function is

$$\begin{aligned}\hat{F}_d^{RKM}(t) &= n_d^{-1}\sum_{j\in s_d} I(y_j \le t) + N_d^{-1}\sum_{k\in s_d} n^{-1}\sum_{j\in s_d} I(y_j - \hat{y}_j \le t - \hat{y}_k) + \\ &- (n_d^{-1} - N_d^{-1})\sum_{k\in s_d} n_d^{-1}\sum_{j\in s_d} I(y_j - \hat{y}_j \le t - \hat{y}_k).\end{aligned} \tag{6.12}$$

Chambers et al. (1992) compared the large-sample mean squared errors of (6.10) and (6.12) and concluded that neither dominates the other. When the model is correctly specified we expect (6.10) to outperform (6.12). However Rao-Kovar-Mantel demonstrated that (6.10) can be substantially biased when model assumptions fail, while (6.12) is less sensitive. Here we just note that the Rao-Kovar-Mantel predictor can be used to define a predictor of a small area characteristic that can be represented as a functional of the small area distribution function in exactly the same way as the CD-type predictor (6.10). In general, the resulting predictors will not be the same. An exception is the Rao-Kovar-Mantel based predictor of the area mean, which is the same as the CD-based predictor of this mean under simple random sampling.

## 6.6 Mean squared error estimation for M-quantile predictors for domains

Mean squared error estimation for the M-quantile predictors of small area averages is described in Chambers et al. (2007). To start, we note that since an iteratively reweighted least squares algorithm is used to calculate the M-quantile regression fit at $\hat{\theta}_d$, we have

$$\hat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_d) = (\mathbf{X}_s'\mathbf{W}_{s_d}\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{W}_{s_d}\mathbf{y}_s$$

where $\mathbf{X}_s$ and $\mathbf{y}_s$ denote the matrix of sample $x$ values and the vector of sample $y$ values respectively, and $\mathbf{W}_{s_d}$ denotes the diagonal weight matrix of order $n$ that defines the estimator of the M-quantile regression coefficient with $q = \hat{\theta}_d$. It follows that the M-quantile predictor (6.7) can be expressed in a weighted form

$$\hat{\bar{Y}}_d = \mathbf{w}_{s_d}'\mathbf{y}_s$$

with weights

$$\mathbf{w}_{s_d}^{MQ} = N_d^{-1}\left[\Delta_n^{(d)} + \mathbf{W}_s(\hat{\theta}_d)\mathbf{X}_s\{\mathbf{X}_s'\mathbf{W}_s(\hat{\theta}_d)\mathbf{X}_s\}^{-1}\mathbf{X}_r\Delta_{N-n}^{(d)}\right]. \tag{6.13}$$

Here $\mathbf{W}_s(\hat{\theta}_d)$ is the diagonal matrix of final weights used in the IRLS algorithm. It also follows that (6.11) can be written

$$\hat{\bar{Y}}_d^{MQ/CD} = \mathbf{w}_{s_d}'\mathbf{y}_s$$

with weights

$$\mathbf{w}_{s_d} = (w_{jd}) = n_d^{-1}\Delta_{s_d} + (1 - N_d^{-1}n_d)\mathbf{W}_d\mathbf{X}_s(\mathbf{X}_s'\mathbf{W}_d\mathbf{X}_s)^{-1}\left\{\overline{\mathbf{x}}_{r_d} - \overline{\mathbf{x}}_{s_d}\right\}. \tag{6.14}$$

with $\Delta_{s_d}$ denoting the $n$-vector that 'picks out' the sample units from area $d$. Given the linear representation of the M-quantile predictors, methods of robust mean squared error estimation for linear predictors of population quantities (Royall and Cumberland, 1978) can be used. In particular, the prediction variance of $\hat{\bar{Y}}_d^{MQ/CD}$ is estimated by

$$v(\hat{\bar{Y}}_d^{MQ/CD}) = \frac{1}{N_d^2}\sum_{g=1}^{G}\sum_{j\in s_g}\lambda_{jdg}\left\{y_j - \mathbf{x}_j\hat{\boldsymbol{\beta}}_{\psi}(\hat{\theta}_g)\right\}^2, \tag{6.15}$$

where $\lambda_{jdg} = \left\{(w_{jd}-1)^2 + (n_d-1)^{-1}(N_d-n_d)\right\}I(g=d) + w_{jg}^2 I(g \neq d)$.

## 6.7 Borrowing strength over space with M-quantile small area models

As we have seen in this literature review, mixed effects models are widely used in small area estimation. Typically, such models assume independence of random area effects and individual effects. This assumption of unit level independence is also implicit in the M-quantile approach to small area estimation. In economic, environmental and epidemiological applications, however, observations that are spatially close may be more related than observations that are further apart. One approach for modeling this spatial correlation is by extending random effects models to allow for spatially correlated area effects, e.g. via a Simultaneous Autoregressive Regression (SAR) random effects model. This approach to small area estimation was described in Chapter 4 of this document. SAR models allow for spatial correlation in the error structure. An alternative approach to incorporate 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., 1996) extends the traditional regression model by allowing local rather than global parameters to be estimated. That is, GWR directly models spatially non-stationarity in the mean structure of the outcome variable. The use of GWR in small area estimation with M-quantile models was proposed by Salvati et al (2007). In doing so the authors first propose an M-quantile GWR model, i.e. a local model for the M-quantiles of the conditional distribution of the outcome variable given the covariates. This model is then used to define a predictor of the small area characteristic of interest (here we focus on the small area mean) that accounts for spatial association in the data. An important spin-off from this approach are more efficient synthetic estimators for out of sample areas.

Given $n$ observations at a set of $L$ locations $\{u_l; l = 1, \dots, L; L \leq n\}$, with $n_l$ data values $\{y_{jl}, x_{jl}; j = 1, \dots, n_l\}$ observed at location $u_l$, a GWR model is defined as follows

$$y_{jl} = \mathbf{x}_{jl}\boldsymbol{\beta}(u_l) + \epsilon_{jl} \tag{6.16}$$

The value of the regression "function" $\boldsymbol{\beta}(u)$ at an arbitrary location $u$ is estimated using weighted least squares

$$\hat{\boldsymbol{\beta}}(u) = \left\{ \sum_{l=1}^{L} w(u_l, u) \sum_{j=1}^{n_l} \mathbf{x}_{jl}\mathbf{x}'_{jl} \right\}^{-1} \left\{ \sum_{l=1}^{L} w(u_l, u) \sum_{j=1}^{n_l} \mathbf{x}_{jl} y_{jl} \right\}$$

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. One popular approach to defining such a weighting function puts

$$w(u_l, u) = \begin{cases} exp\left[1 - (d_{u_l,u}/b)^2\right]^2 & if\ d_{u_l,u} \leq b, \\ 0 & otherwise \end{cases} \tag{6.17}$$

where $d_{u_l,u}$ denotes the Euclidean distance between $u_l$ and $u$ and $b$ is the bandwidth, which can be optimally defined using a least squares criterion (Fotheringham et al., 2002). It should be noted, however, that alternative weighting functions can also be used. The GWR model (6.16) is a model for the conditional expectation of $\mathbf{y}$ given $x$ at location $u$. This is easily generalised to a model for the M-quantile of order $q$ of the conditional distribution of $\mathbf{y}$ given $x$ at $u$. That is, we write

$$Q_q(\mathbf{x};\psi,u)=\mathbf{X}\boldsymbol{\beta}_{\psi}(u;q) \tag{6.18}$$

where now $\boldsymbol{\beta}_{\psi}(u;q)$ varies with $u$ as well as with $q$. That is, (6.18) allows the entire conditional distribution (not just the mean) of $y$ given $\mathbf{X}$ to vary from location to location. The parameter $\boldsymbol{\beta}_{\psi}(u;q)$ in (6.18) can be estimated by solving

$$\sum_{l=1}^{L} w(u_l,u)\sum_{j=1}^{n_l}\psi_q\left\{y_{jl}-\mathbf{x}'_{jl}\boldsymbol{\beta}_{\psi}(u;q)\right\}\mathbf{x}_{jl}=0 \tag{6.19}$$

where $\psi_q(t)=2\psi(s^{-1}t)\left\{qI(t>0)+(1-q)I(t\leq 0)\right\}$. Here s is a suitable robust estimate of the scale of the sample $y$ values, e.g. the MAD estimate of scale and we will typically assume a Huber Proposal 2 influence function, $\psi(t)=tI(-c\leq c)+c\mathtt{sgn}(t)I(|t|>c)$. Provided $c$ is bounded away from zero, an iteratively re-weighted least squares algorithm that combines the iteratively re-weighted least squares algorithm used to fit a "spatially stationary" M-quantile model and the weighted least squares algorithm used to fit a GWR model can then be used to solve (6.19), leading to estimates of the form

$$\hat{\boldsymbol{\beta}}_{\psi}(u;q)=\left\{\mathbf{X}'_s\mathbf{W}^*_s(u;q)\mathbf{X}_s\right\}\mathbf{X}'_s\mathbf{W}^*_s(u;q)\mathbf{y}_s \tag{6.20}$$

Here $\mathbf{y}_s$ is the vector of $n$ sample $y$ values and $\mathbf{X}_s$ is the corresponding matrix of order $n\times p$ of sample $x$ values. The matrix $\mathbf{W}^*_s(u;q)$ is a diagonal matrix of order $n$ with each entry corresponding to a particular sample observation equal to the product of this observation's spatial weight, which depends on its distance from location $u$, with the weight that this observation has when the sample data are used to calculate the "spatially stationary" M-quantile estimate $\hat{\beta}_{\psi}(q)$.

Having defined an M-quantile variant of the GWR model, we can then use this model to predict the area $d$ mean $\overline{Y}_d$ of $y$. Following Chambers and Tzavidis (2006), we can first estimate M-quantile GWR coefficients $\{q_j; j\in s\}$ of the sampled population units without reference to the small areas of interest. This can be done using the grid-based interpolation described in section 6.5. In particular, we adapt this approach to the GWR M-quantile model by first defining a fine grid of $q$ values over the interval (0,1) and then using the sample data to fit the model for each distinct value of $q$ on this grid and at each sample location. The M-quantile GWR coefficient for unit $j$ with values $y_j$ and $\mathbf{x}_j$ at location $\mathbf{u}_j$ is finally calculated by interpolating over this grid to find the value $q_j$ such that $Q_q(\mathbf{x}_j;\psi,u_j)=y_j$. Provided that there are sample observations in area $d$, an area $d$ specific M-quantile GWR coefficient $\hat{\theta}_d$ can be defined as the average value of the sample M-quantile GWR coefficients in area $d$. Following Tzavidis and Chambers (2007), a bias-adjusted M-quantile GWR predictor of the mean $\overline{Y}_d$ in small area $d$ is

$$\hat{\overline{Y}}_d^{MQGWR/CD}=N_d^{-1}\left[\sum_{j\in s_d}y_j+\sum_{j\in r_d}\hat{Q}_{\hat{\theta}_d}(\mathbf{x}_j;\psi,u_j)+\frac{N_i-n_i}{n_i}\sum_{j\in s_i}\left\{y_j-\hat{Q}_{\hat{\theta}_d}(\mathbf{x}_j;\psi,u_j)\right\}\right] \tag{6.21}$$

where $\hat{Q}_{\hat{\theta}_d}(\mathbf{x}_j;\psi,u_j)$ is defined via the MQGWR model.

There are situations where we are interested in estimating small area characteristics for domains (areas) with no sample observations. The conventional approach to estimating a small area characteristic, say the mean, in this case is synthetic estimation. Under the M-quantile model the synthetic mean predictor for out of sample area $d$ is $\hat{\bar{Y}}_d^{MQ/SYNTH} = N_d^{-1}\sum_{j\in\Omega_d}\mathbf{x}_j\hat{\boldsymbol{\beta}}_\psi(0.5)$. We note that with synthetic estimation all variation in the area-specific predictions comes from the area-specific auxiliary information. One way of potentially improving the conventional synthetic estimation for out of sample areas is by using a model that borrows strength over space such as an M-quantile GWR model. In this case a synthetic-type mean predictor for out of sample area $d$ is defined by

$$\hat{\bar{Y}}_d^{MQGWR/SYNTH} = N_d^{-1}\sum_{j\in\Omega_d}\hat{Q}_{0.5}(\mathbf{x}_j;\psi,u_j)$$

## 6.8 Mean squared error estimation for M-quantile GWR predictors for domains

The arguments outlined in section 6.6 can be extended to define an estimator of the mean squared error of (6.21). To start we note that (6.21) can be expressed as a weighted sum of the sample $y$-values

$$\hat{\bar{Y}}_d^{MQGWR/CD} = N_d^{-1}\mathbf{w}'_{s_d}\mathbf{y}_s \tag{6.22}$$

where

$$\mathbf{w}_{s_d} = \frac{N_d}{n_d}\mathbf{1}_{s_d} + \sum_{j\in r_d}\mathbf{H}'_{jd}\mathbf{x}_j - \frac{N_d-n_d}{n_d}\sum_{j\in s_d}\mathbf{H}'_{jd}\mathbf{x}_j. \tag{6.23}$$

Here $\mathbf{1}_{s_d}$ is the $n$-vector with $j^{th}$ component equal to one whenever the corresponding sample unit is in area $d$ and is zero otherwise and

$$\mathbf{H}_{jd} = \left\{\mathbf{X}'_s\mathbf{W}^*_s(u_j;\hat{\theta}_d)\mathbf{X}_s\right\}^{-1}\mathbf{X}'_s\mathbf{W}^*_s(u_j;\hat{\theta}_d).$$

Given the linear representation (6.22), an estimator of a first order approximation to the mean squared error of this predictor can be computed following methods of robust mean squared error estimation for linear predictors of population quantities (Royall and Cumberland, 1978). Put $w_{s_d} = (w_{jd})$. This estimator is of the form

$$v(\hat{\bar{Y}}_d^{MQGWR/CD}) = \frac{1}{N_d^2}\sum_{g:n_g>0}\sum_{j\in s_g}\lambda_{jdg}\left\{y_j - \hat{Q}_{\hat{\theta}_g}(x_j,\psi,u_j)\right\}^2 \tag{6.24}$$

where $\lambda_{jdg} = \left\{(w_{jd}-1)^2 + (n_d-1)^{-1}(N_d-n_d)\right\}I(g=d) + w_{jg}^2 I(g\neq d)$.

## 6.9 Nonparametric M-quantile small area estimation

The $p$-splines M-quantile regression methodology presented in section 6.4 can be applied to the estimation of small area quantities. To estimate the small area mean, the first step is to estimate the M-quantile coefficients $q_j$ for each unit $j$ in the probabilistic sample $s$ of size $n$ without reference to the $D$ small areas of interest. This is done defining a fine grid of values on the interval $(0, 1)$ and using the sample data to fit the $p$-splines M-quantile regression functions at each value $q$ on this grid. If a data point lies exactly on the $q$th fitted curve, then the coefficient of the corresponding sample unit is equal to $q$. Otherwise, to obtain $q_j$, a linear interpolation over the grid is used.

If a hierarchical structure does explain part of the variability in the population data, we expect units within clusters defined by this hierarchy to have similar M-quantile coefficients. Therefore, an estimate of the mean quantile for area $d$, $\theta_d$, is obtained by taking the corresponding average value of the sample M-quantile coefficient of each unit in area $d$, $\hat{\theta}_d = \sum_{j=1}^{n_d} q_j$. The small area estimator of the mean $\bar{Y}_d$ is then

$$\hat{\bar{Y}}_d = \frac{1}{N_d}\Big\{\sum_{j\in s_d} y_{jd} + \sum_{j\in r_d} \hat{y}_{jd}\Big\}, \tag{6.25}$$

where $s_d$ and $r_d$ denote the sampled and non sampled units in area $d$, respectively, with $\Omega_d = s_d \cup r_d$, and $N_d$ is the known population size of area $d$. Note that the unobserved value for population unit $d \in r_d$ is predicted using

$$\hat{y}_{jd} = \mathbf{x}_{jd}\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d) + \mathbf{z}_{jd}\hat{\boldsymbol{\gamma}}_\psi(\hat{\theta}_d),$$

where $\hat{\boldsymbol{\beta}}_\psi(\hat{\theta}_d)$ and $\hat{\boldsymbol{\gamma}}_\psi(\hat{\theta}_d)$ are the coefficient vectors of the parametric and spline portion, respectively, of the fitted $p$-splines M-quantile regression function at $\hat{\theta}_d$.

The predicted values $\hat{y}_{jd}$ can also be used to obtain a bias-adjusted estimator of the mean through the Chambers and Dunstan adjustment:

$$\hat{\bar{Y}}_d = \frac{1}{N_d}\Big\{\sum_{j\in s_d} y_{jd} + \sum_{j\in r_d} \hat{y}_{jd} + \frac{N_d - n_d}{n_d}\sum_{j\in s_d}(y_{jd} - \hat{y}_{jd})\Big\}, \tag{6.26}$$

where $\hat{y}_{jd}$ denotes the predicted values for the population units in $s_d$ and in $r_d$.

In many instances we are interested in estimating parameters for out of sample areas, that is areas where there are not sampled units even if in those areas there are population units with the characteristic of interest. In this case no area effects can be computed and the small area characteristic is estimated by using synthetic estimation. We can note that with synthetic estimation all variation in the area-specific predictions comes from the area-specific auxiliary information. One approach to improving estimation for out of sample areas is by borrowing strength over space (Saei and Chambers, 2005). In the case of $p$-splines M-quantile regression, this can be achieved using model (6.4) and setting $\hat{\theta}_d = 0.5$. A synthetic type mean predictor for out of sample area $d$ is given by

$$\hat{\bar{Y}}_d = \frac{1}{N_d}\left\{\sum_{j\in r_d} \mathbf{x}_{jd}\hat{\boldsymbol{\beta}}_\psi(0.5) + \mathbf{z}_{jd}\hat{\boldsymbol{\gamma}}_\psi(0.5)\right\}. \tag{6.27}$$

We expect that when a truly spatially process is present, (6.27) will improve the efficiency of the other traditional synthetic estimators.

Following the approach described in Chandra and Chambers (2005) and Chambers and Tzavidis (2006), for fixed $q$ and $\lambda$, the $\hat{\bar{Y}}_d$ in (6.26) can be written as the following linear combination of the observed $y_j$,

$$\hat{\bar{Y}}_d = \frac{1}{N_d}\sum_{j\in s} w_{jd}y_j, \tag{6.28}$$

where the $n$-vector of weights $\mathbf{w}_d = (w_{1d}, \ldots, w_{nd})'$ is given by

$$\mathbf{w}_d = \frac{N_d}{n_d}\mathbf{1}_{s_d} + \mathbf{W}(\hat{\theta}_d)\,[\mathbf{X}\ \mathbf{Z}]\,\Big([\mathbf{X}\ \mathbf{Z}]'\,\mathbf{W}(\hat{\theta}_d)\,[\mathbf{X}\ \mathbf{Z}] + \lambda\mathbf{G}\Big)^{-1}\Big(\mathbf{T}_{r_d} - \frac{N_d - n_d}{n_d}\mathbf{T}_{s_d}\Big) \tag{6.29}$$

with $\mathbf{1}_{s_d}$ the $n$-vector with $j^{th}$ component equal to one whenever the corresponding sample unit is in area $d$ and to zero otherwise, $\mathbf{W}(\hat{\theta}_d)$ a diagonal $n\times n$ matrix that contains the final set of weights produced by the iteratively reweighted penalized least squares algorithm used to estimate the regression coefficients, $\mathbf{G} = \mathtt{diag}\{\mathbf{0}_P, \mathbf{1}_K\}$ with $P$ the number of columns of $\mathbf{X}$ and $K$ the number of columns of $\mathbf{Z}$, and with $\mathbf{T}_{r_d}$ and $\mathbf{T}_{s_d}$ the totals of the covariates for the non-sampled and the sampled units in area $d$, respectively.

The weights derived from (6.29) are treated as fixed and a "plug in" estimator of the mean squared error of estimator (6.28) given by

$$MSE(\hat{\bar{Y}}_d) = var(\hat{\bar{Y}}_d - \bar{Y}_d) + [bias(\hat{\bar{Y}}_d)]^2 \tag{6.30}$$

can be proposed by using standard methods for robust estimation of the variance of unbiased weighted linear estimators (Royall and Cumberland, 1978) and by following the results due to Tzavidis and Chambers (2007). The prediction variance of (6.28) can be approximated by

$$var(\hat{\bar{Y}}_d - \bar{Y}_d) \approx \frac{1}{N_d^2}\Big(\sum_{j\in s_d}\Big\{d_{jd}^2 + \frac{N_d - n_d}{n_d - 1}\Big\}var(y_{jd}) + \sum_{j\in s\backslash s_d} d_{jd}^2 var(y_{jd})\Big) \tag{6.31}$$

with $d_{jd} = w_{jd} - 1$ if $j \in s_d$ and $d_{jd} = w_{jd}$ otherwise, and $s\backslash s_d$ the set of sampled units outside area $d$. The bias can be written as

$$bias(\hat{\bar{Y}}_d) \approx \frac{1}{N_d}\Big(\sum_{k=1}^{D}\sum_{j\in s_k} w_{jd}\tilde{y}_{jk} - \sum_{j\in\Omega_d}\tilde{y}_{jd}\Big) \tag{6.32}$$

where $\tilde{y}_{jk} = \mathbf{x}_{jk}\boldsymbol{\beta}_{\psi}(\hat{\theta}_k) + \mathbf{z}_{jk}\boldsymbol{\gamma}_{\psi}(\hat{\theta}_k)$ are the study variable values under the $p$-splines M-quantile regression model. Following the area level residual approach (Tzavidis and Chambers, 2006), we can interpret $var(y_{jd})$ conditionally to the specific area $d$ from which $y_j$ is drawn and hence replace $var(y_{jd})$ in (6.31) by $(y_{jd} - \hat{y}_{jd})^2$. An estimate of the bias is obtained replacing $\tilde{y}_{jk}$ by $\hat{y}_{jk}$ in (6.32). A robust estimator of the mean squared error of (6.28) is given by the sum of the estimator of the variance

$$\widehat{var}(\hat{\bar{Y}}_d) = \frac{1}{N_d^2}\left[\sum_{j\in s_d}\left\{d_{jd}^2 + \frac{N_d - n_d}{n_d - 1}\right\}(y_{jd} - \hat{y}_{jd})^2 + \sum_{j\in s\backslash s_d} d_{jd}^2(y_{jd} - \hat{y}_{jd})^2\right] \tag{6.33}$$

and the squared estimate of the bias

$$\hat{b}^2(\hat{\hat{Y}}_d) = \frac{1}{N_d^2}\left(\sum_{k=1}^{D}\sum_{j\in s_k} w_{jd}\hat{y}_{jk} - \sum_{j\in\Omega_d}\hat{y}_{jd}\right)^2. \qquad (6.34)$$

Since the bias-adjusted nonparametric M-quantile estimator is an approximately unbiased estimator of the small area mean, the squared bias term will not impact significantly the mean squared error estimator. The main limitation of the MSE estimator is that it does not account for the variability introduced in estimating the area specific $q$'s and $\lambda$. We note also that we can obtain an estimate only for areas where there are at least two sampled units. For all these reasons, we are currently investigating the use of bootstrap as an alternative approach for estimating the MSE.

## 6.10 References

Aragon, Y., Casanova, S., Chambers, R.L., Leconte, E. (2005). Conditional ordering using nonparametric expectiles. *Journal of Official Statistics*, **21**, 617-633.

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**, 28-36.

Bollaerts, K., Eilers, P.H.C., Aerts, M. (2006). Quantile regression with monotonocity restrictions using $p$-splines and the l-1-norm. *Statistical Modelling*, **6**, 189-207.

Breckling, J. and Chambers, R.L. (1988). M-quantiles. *Biometrika*, **75**, 761-71.

Breidt, F.J., Opsomer, J.D., Johnson, A.A., Ranalli, M.G. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey methodology*, **33**, 35-44.

Brunsdon, C., Fotheringham, A.S., Charlton, M. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, **28**, 281-298.

Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, **81**, 1063-1069.

Chambers, R.L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, **73**, 597-604.

Chambers, R.L., Dorfman, A.H., Hall, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika*, **79**, 577-82.

Chambers, R.L. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, **93**, 255-268.

Chambers, R.L., Chandra, H., Tzavidis, N. (2007). On robust mean squared error estimation for linear predictors for domains. [Paper submitted for publication. A copy is available upon request].

Chandra, H. and Chambers, R.L. (2005). Comparing EBLUP and C-EBLUP for small area estimation. *Statistics in Transition*, **7**, 637-648.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377-403.

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, **78**, 605-610.

Eilers, P.H. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89-121.

Fotheringham, A.S., Brundson, C., and Charlton M. (2002). *Geographically Weighted Regression - The analysis of spatially varying relationship*. West Sussex, England: John Wiley & Sons Ltd.

Green, P.J. and Silverman, B.W. (1994). *Non parametric regression and generalized linear models: a roughness penalty approach*. Chapman & Hall Ltd.

Haerdle, W. and Gasser, T. (1984). Robust non-parametric function fitting. *Journal of the Royal Statistical Society, Series B*, **46**, 42-51.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, **51**, 186-192.

Huber, P.J. (1981). *Robust Statistics*. John Wiley & Sons.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33-50.

Koenker R. and D'Orey, V. (1987). Computing regression quantiles. *Applied Statistics*, **36**, 383-93.

Koenker R., NG, P., Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, **81**, 673-680.

Kokic, P., Chambers, R.L., Breckling, J., Beare, S. (1997). A measure of production performance. *Journal of Business and Economic Statistics*, **15**, 445-451.

Lee, T.C. and Oh, H.S. (2007). Robust penalized regression spline fitting with application to additive mixed modeling. *Computational Statistics*, **22**, 159-171.

Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression. A Secondary Course in Statistics*. Reading, MA: Addison-Wesley.

Newey, W.K. and powell, J.L. (1987). Asymmetric least squared estimation and testing. *Econometrica*, **55**, 819-47.

Nychka, D. and Saltzman, N. (1998). Design of air quality monitoring networks. In Kychka, Douglas Piegorsch, Walter and Cox (eds), *Case studies in environmental statistics*.

Pratesi, M., Ranalli, M.G., Salvati, N. (2006). Non-parametric M-quantile regression via penalized splines. *ASA Proceedings on Survey Research Methods*, Alexandria, VA.

Pratesi, M., Ranalli, M.G., Salvati, N. (2007). Non parametric M-quantile regression using penalized splines in small area estimation. *Working paper*, University of Pisa.

Pratesi, M., Ranalli, M.G., Salvati, N. (2008). Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US. To appear in *Environmetrics*.

R Development Core Team (2005). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. URL: http://www.R-project.org.

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**, 365-75.

Royall, R.M. and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, **73**, 351 - 58.

Ruppert, D., Wand, M.P., Carroll, R. (2003). *Semiparametric regression*. Cambridge University Press, Cambridge, New York.

Saei, A. and Chambers, R.L. (2005). Empirical best linear unbiased prediction for out of sample areas. In *S3Ri Methodology Working Papers*. Southampton Statistical Science Research Institute, pp. 1-15.

Salvati, N., Tzavidis, N., Pratesi, M., Chambers, R.L. (2007). Small area estimation via M-quantile geographically weighted regression. [Paper submitted for publication. A copy is available upon request].

Tzavidis, N. and Chambers, R.L. (2006). Bias adjusted distribution estimation for small areas with outlying values. In *S3Ri Methodology Working Papers*. Southampton Statistical Science Research Institute, pp. 1-30.

Tzavidis, N. and Chambers, R.L. (2007). Robust prediction of small are means and distributions. [Paper submitted for publication. A copy is available upon request].

Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S*. New York: Springer.

Wang, S. and Dorfman, A.H. (1996). A new estimator of the finite population distribution function. *Biometrika*, **83**, 639-52.

Welsh, A.H. (1996). Robust estimation of smooth regression and spread functions and their derivatives. *Statistica Sinica*, **6**, 347-366.

# Chapter 7

# Estimation of cumulative distribution functions

## 7.1 Introduction

Cumulative distribution function estimation is often an important objective in survey practice. The distribution function allows to identify subgroups in the population whose values for a particular variable lie below or above a given limit. For example, Laeken poverty indicators, such as *at-risk-of-poverty rate*, are based on quantiles (that are easily derived from the cumulative distribution function) of the distribution of equivalized net income. Furthermore, when the variable of interest is strongly skewed or is plurimodal the knowledge of the mean can be misleading, while the knowledge of the distribution function gives full information.

The cumulative distribution function can be estimated without using auxiliary information (section 7.3) or using this available information in a parametric, semiparametric or nonparametric framework (section 7.4). This classification crosses the one between design-based and model-based estimators of the cumulative distribution function.

Another topic of interest closely related to the estimation of the cumulative distribution function is that of quantile estimation (section 7.5).

The problem of estimating the cumulative distribution function at the small area level (section 7.6) can be faced again with or without the use of auxiliary information. As for the estimation of small area mean values (chapter 6), the novel approach is represented by the use of M-quantile and nonparametric M-quantile regression models.

The estimation of the mean squared error of the cumulative distribution function estimator is an open research problem, in particular for what concern small area estimation. For the distribution function estimator in general the proposed mean squared error estimators depend on the choice of a specific superpopulation model. Alternative jackknife and bootstrap approaches have been proposed, but the problem is still open in the small area case.

## 7.2 Estimating cumulative distribution function from survey data

Let $\Omega = \{1, \ldots, N\}$ be a finite population. Let $\mathbf{y} = (y_1, \ldots, y_N)'$ denote the variable values for the $N$ population elements. The cumulative distribution function of $Y$ at population level is

$$F(t) = N^{-1} \sum_{j \in \Omega} I(y_j \leq t). \tag{7.1}$$

We consider a sample $s \subset \Omega$, of $n \leq N$ units, and we denote with $r = \Omega - s$ the set of non sampled units. For each population unit $j$, let $\mathbf{x}_j = (x_{1j}, \ldots, x_{pj})$ denote a vector of $p$ known auxiliary variables. The goal is to estimate (7.1).

One possibility to classify the contributions in the literature to the estimation of (7.1) is by considering whether they make use of auxiliary information or not.

## 7.3 Estimating cumulative distribution function without auxiliary information

The customary design-based estimator, the Hájek estimator of the finite population distribution function (7.1) is

$$\hat{F}_H(t) = \frac{\sum_{j \in s} (1/\pi_j) I(y_j \leq t)}{\sum_{j \in s} 1/\pi_j} \tag{7.2}$$

where $\pi_j$ is the probability of inclusion in the sample of unit $j$. The estimator $\hat{F}(t)$ is design-unbiased for $F(t)$ under any sampling scheme such that $\sum_{j \in s} \pi^{-1} = N$. The first proposal on how to estimate the variance of this estimator is probably in Woodruff (1952).

Alternative estimators to (7.2) were considered and compared in Kuk (1988), namely the Hájek estimator defined above, the Horvitz Thompson estimator

$$\hat{F}_{HT}(t) = \frac{\sum_{j \in s} (1/\pi_j) I(y_j \leq t)}{N} \tag{7.3}$$

and the complementary proportion estimator, that is the estimator for $S(t) = 1 - F(t)$

$$\hat{F}_{CP}(t) = 1 - \frac{\sum_{j \in s} (1/\pi_j) I(y_j > t)}{N}. \tag{7.4}$$

Theoretical results in Kuk (1988) suggest that $\hat{F}_{CP}(t)$ and $\hat{F}_H(t)$ should be preferred to $\hat{F}_{HT}(t)$. Moreover, empirical results show that $\hat{F}_{CP}(t)$ should be used for the estimation of medians. On the issue of estimating the cumulative distribution function without using auxiliary information see also Hyndman and Fan (1996), Shuster (1973) and Modarres (2002).

## 7.4 Estimating cumulative distribution function using auxiliary information

### The parametric case

The customary estimator (7.2) does not make use of auxiliary population information. To remedy this, Rao et al. (1990) proposed an alternative model-assisted estimator of the cumulative distribution function.

Consider the superpopulation model with one covariate $x$

$$y_j = \beta x_j + v(x_j)e_j \tag{7.5}$$

where $\beta$ is an unknown parameter, $v(x) = x^{1/2}$ and the $e_j$ are independent and identically distributed random variables with zero mean. Rao et al. (1990) consider two methods to include the auxiliary information at the estimation stage: the ratio and the difference estimator. The ratio estimator of $F(t)$ is

$$\hat{F}^{RKM_a}(t) = N^{-1}\left\{\sum_{j\in s}(\pi_j^{-1})I(y_j \le t)\right\}\left\{\sum_{j\in s}\pi_j^{-1}I(\hat{R}x_j \le t)\right\}^{-1}\left\{\sum_{j\in\Omega}I(\hat{R}x_j \le t)\right\} \tag{7.6}$$

where $\hat{R} = (\sum_{j\in s} y_j/\pi_j)(\sum_{j\in s} x_j/\pi_j)^{-1}$ is the customary design-consistent estimator of the population ratio $R = y/x$. This estimator is more efficient than (7.2) when $y_j$ is approximately proportional to $x_j$.

The difference estimator is

$$\hat{F}^{DIF}(t) = N^{-1}\left\{\sum_{j\in s}(\pi_j^{-1})I(y_j \le t) + \left(\sum_{j\in\Omega}I(\hat{R}x_j \le t) - \sum_{j\in s}\pi_j^{-1}I(\hat{R}x_j \le t)\right)\right\}. \tag{7.7}$$

This estimator has the same property of $\hat{F}^{RKM_a}(t)$ and it avoids the ratio bias, especially for small $n$.

Rao et al. (1990) propose an alternative difference estimator because the (7.6) and (7.7) are asymptotically design-unbiased but not model-unbiased under model (7.5). The estimator is

$$\hat{F}^{RKM_b}(t) = N^{-1}\left\{\sum_{j\in s}(\pi_j^{-1})I(y_j \le t) + \left(\sum_{j\in\Omega}\hat{G}_j - \sum_{j\in s}\pi_j^{-1}\hat{G}_{jc}\right)\right\} \tag{7.8}$$

where

$$\hat{G}_j = \left(\sum_{k\in s}1/\pi_k\right)^{-1}\left(\sum_{k\in s}\pi_k^{-1}I(\hat{R}x_j + \hat{e}_k x_j^{1/2} \le t)\right),$$

$$\hat{G}_{jc} = \left(\sum_{k\in s}\pi_j/\pi_{jk}\right)^{-1}\left(\sum_{k\in s}(\pi_j/\pi_{jk})I(\hat{R}x_j + \hat{e}_k x_j^{1/2} \le t)\right),$$

$\hat{e}_j = x_j^{-1/2}(y_j - \hat{R}x_j)$ and $\pi_{jk}/\pi_j$ is the conditional probability of selecting unit $j$ and $k$ given that $j \in s$.

The variance estimators for (7.6), (7.7) and (7.8) are proposed in Rao et al. (1990). Wu and Sitter (2001b) argue that these variance estimators rely on the existence of an asymptotic expansion which may not exist for a particular design and is difficult to verify in general. Their proposal is to use a jackknife approach to obtain a design-consistent estimator of the variance of the design-based estimator (7.8).

A model-based parametric approach to the estimation of the finite population cumulative distribution function has been proposed by Chambers and Dunstan (1986).

Consider the (7.1) and note that it can be decomposed as

$$F(t) = N^{-1}\left\{\sum_{j\in s} I(y_j \leq t) + \sum_{k\in r} I(y_k \leq t)\right\}.$$

The unknown quantity $\sum_{k\in r} I(y_k \leq t)$ needs to be estimated. Assuming that a superpopulation model exists, we can predict the unobserved $y$ values under this model and get a model-unbiased estimator for $F(t)$. Consider again as a superpopulation the model (7.5), where the auxiliary variable $x$ is known for all the units in the population. A naive estimator is obtained replacing the unknown quantity $\sum_{k\in r} I(y_k \leq t)$ with the estimated one $\sum_{k\in r} I(\hat{y}_k \leq t)$ under the superpopulation model. Since this leads to a biased estimator, Chamber and Dunstan (1986) (CD hereafter) propose this estimator of $F(t)$

$$\hat{F}^{CD}(t) = N^{-1}\left\{\sum_{j\in s} I(y_j \leq t) + \sum_{k\in r} n^{-1} \sum_{j\in s} I(\hat{\beta} x_k + v(x_j)\hat{e}_j \leq t)\right\} \tag{7.9}$$

where $\hat{\beta} = \left\{\sum_{j\in s} y_j x_j / v^2(x_j)\right\}\left\{\sum_{j\in s} x_j^2 / v^2(x_j)\right\}^{-1}$. This estimator is correct under the superpopulation model (7.5). Consistency and asymptotic results for the bias and variance of the Chambers and Dunstan estimator are derived in Chambers et al. (1992) under a simple linear model. Considering alternative superpopulation models, asymptotic results must be completely re-derived. For this reason, Wu and Sitter (2001b) proposed a jackknife consistent estimator of the variance of (7.9). Alternatively, Lombardia et al. (2003) presented a consistent bootstrap procedure to estimate the mean squared error.

Several estimators were proposed as variants of the RKM and CD estimators.

Wang and Dorfman (1996) proposed a weighted estimators of the RKM and CD estimators of the form $\hat{F}^{WD}(t) = \omega \hat{F}^{CD} + (1-\omega)\hat{F}^{RKM_b}$, where $\omega$ is a weight estimated from available data to achieve minimal asymptotic mean squared error.

Mak and Kuk (1993) proposed a variant to the CD estimator of this form

$$\hat{F}^{MK}(t) = N^{-1}\left\{\sum_{j\in s} I(y_j \leq t) + \sum_{k\in r} \Phi\left(\frac{t - \hat{\beta} x_k}{\hat{\sigma} v^{1/2}(x_k)}\right)\right\} \tag{7.10}$$

where $\Phi(\cdot)$ is the normal cumulative distribution function and the estimate of the standard deviation $\hat{\sigma}$ is obtained from the weighted regression fit.

A robust approach to the estimation of the cumulative distribution function is due to Welsh and Ronchetti (1998). They proposed a methodology based on the use of robust estimates and a bias-calibrated form of the CD estimator. This method is useful when the sample data contain a small number of representative outliers. Representative outliers are correct observations from typical units which are extreme relative to the bulk of the data (Chambers, 1986).

Some other proposals can be found in Lombardia et al. (2005), Dorfman (1993), Chambers et al. (1993).

### The nonparametric case

The susceptibily of parametric methods to mean function and variance misspecification bias provides motivation to consider a superpopulation model $y_j = m(x_j) + v(x_j)e_j$ using nonparametric methods. In this case, the only hypothesis is that the mean function $m(x_j)$ is some smooth function of the auxiliary data $x_j$, and $v(x_j)$ is smooth and strictly positive. Another possibility is to specify $I(y_j \leq t)$ given $x$ as a smooth function.

Dorfman and Hall (1993) discuss in detail a general class of nonparametric estimators for the finite population distribution function; for example, a model-based nonparametric estimator of the total of $Y$ which can be adapted to the case of the cumulative distribution function is:

$$\hat{F}^{DOR}(t) = (N)^{-1}\left\{\sum_{j\in s} I(y_j \leq t) + \sum_{k\in r} \hat{g}(x_k)\right\}$$

where $\hat{g}(x_k)$ is a model-based nonparametric estimator of $g(x_k)$, the kernel smooth estimator of $G\{(t - m(x_j))v^{-1}(x_j)\}$.

To estimate the bias, the variance and the mean squared error of nonparametric CD-like estimators of the cumulative distribution function, Lombardia et al. (2004) proposed a bootstrap methodology.

Under the hypothesis of a smooth function for $I(y_j \leq t)$ given $x$, Kuo (1988) proposed the first nonparametric estimator. See also Kuk (1993).

The nonparametric approach to the estimation of the cumulative distribution function can be extended to the case where even the assumption that the nonparametric model always holds is removed. More specifically, a local polynomial regression approach based on the population total estimator of Breidt and Opsomer (2000) can be derived. A semiparametric approach for situations were some auxiliary variables are treated nonparametrically and the others are treated parametrically have also been proposed in Johnson et al. (2008), as an extension to the semiparametric estimator of the finite population total by Breidt and Opsomer (2002).

### Calibration estimators

A different class of estimators of the cumulative distribution function calibrate the estimate with respect to the auxiliary variable $x$. All the proposals try to achieve design consistency for protection in case of model failure, and improvement on design-based estimators in case of model correctness.

One possibility is to calibrate using a difference between the total in the population and the corresponding weighted sum in the sample of some function $g(\cdot)$ of $x$: $\sum_{k\in\Omega} g(x_k) - \sum_{j\in s} w_j g(x_j)$. Under this framework and considering $I(y_j \leq t)$ three class of estimators can be defined.

- Generalized regression estimators (GREG). Wu and Sitter (2001a) suggest a $g(x_j)$ producing an estimators that recall the RKM one.

- Estimators specified through the calibration weights. Kovacevic (1997) proposes different distance functions and calibration constrains assuming to have or to have not full population information. Rueda et al. (2007a) consider an estimator which performs on a level with the CD estimator and better then the RKM; Rueda et al. (2007b) propose a nonparametric version of the same estimator. Harms and Duchesne (2006) propose a calibration estimator for the cumulative distribution function as an intermediate step to estimate quantiles, since it is obtained constraining on a given population quantile.

- Pseudo-empirical likelihood estimators. Chen and Wu (2002) suggest three estimators, based on different models, and they show their relative efficiency through a simulation study.

On the estimation of the cumulative distribution function using auxiliary information see also Durrant and Skinner (2006), Nascimento Silva et al. (1995), Kuk and Mak (1994), Dunstan and Chambers (1989), Kuk and Mak (1989).

## 7.5 Quantile estimation

Another topic of interest closely related to the cumulative distribution function estimation problem is that of quantile estimation.

One possibility is to estimate the quantiles through inverting an estimate of the cumulative distribution function $\hat{F}(t)$. For $\theta(\alpha) = \min \{t : F(T) \geq \alpha\}$, the $\alpha$-quantile of $y$, an estimate $\hat{\theta}(\alpha)$ is

$$\hat{\theta}(\alpha) = \min \left\{ t : \hat{F}(t) \geq \alpha \right\}$$

where $\hat{F}(t)$ is based on some finite population cumulative distribution function estimator. See Harms and Duchesne (2006), Kuk and Mak (1989).

A different possibility is the direct estimation of the quantile, $\hat{\theta}(\alpha)$, without the need to specify the cumulative distribution function. Using auxiliary information on the quantiles of $x$, some estimators of this kind has been suggested by Rao at al. (1990). Other quantiles direct estimators using auxiliary information can be found in Meeden (1995) and Mak and Kuk (1993).

As concerns confidence intervals and variance estimation for quantiles, a general method is to transform estimates of precision referring to the cumulative distribution function, as described by Woodruff (1952). Different proposals and some comparison between them can be found in Sitter and Wu (2001), Dorfman and Valliant (1993), Francisco and Fuller (1991), Rao et al. (1990).

See also Oberhofer and Haupt (2005), Rueda and Arcos (2004), Singh et al. (2001), Rueda et al. (1998), Sheather and Marron (1990), Falk (1985), Harrel and Davis (1982).

## 7.6 Small area estimation of the cumulative distribution function

Also the SAE problem of estimating $F(t)$ and the corresponding quantiles can be faced with and without the use of available auxiliary information.

### SAE cumulative distribution function without auxiliary information

The estimator of the cumulative distribution function at small area level $F_d(t)$ for any given $t$ can be viewed as the problem of estimating a small area mean. To see this, note that $F_d(t)$ can be expressed as a small area mean,

$$\hat{F}_d(t) = N_d^{-1} \sum_{j \in \Omega_d} I(y_j \leq t) = \frac{T_z}{N_d} = \bar{z}_d \tag{7.11}$$

where $\Omega_d$ represents the population of area $d$, $I(y_j \leq t)$ is an indicator variable defined for $j = 1, ..., N_d$ and for any real number $t$ as $I(y_j \leq t) = 1$ if $y_j \leq t$ and $0$ otherwise. Consequently we can estimate $F_d(t)$ for a given $t$, by the sample weighted mean at small area level

$$\hat{F}_d(t) = \frac{\hat{T}_z}{\hat{N}_d} = \frac{\sum_{s_d} \frac{I(y_j \leq t)}{\pi_j}}{\sum_{s_d} \frac{1}{\pi_j}} \tag{7.12}$$

where $s_d$ is the set of sample elements of area $d$. To divide by $\hat{N}_d$ rather than by $N_d$ (if $N_d$s were known) will often have advantages from a variance point of view. Since $\hat{F}_d(t)$ is not linear, and consequently only approximately unbiased, we can not determine its exact variance expression (Särndal et al., 1992).

### SAE cumulative distribution function using auxiliary information

$\hat{F}_d(t)$ in (7.11) is based on information about the study variable only for sampled units. However auxiliary information is often available for the entire population of the small area of interest. The relationship of the auxiliary information with the study variable across the sample allows inferences about the not sampled portion of the population in the small area of interest. In particular, the estimation of $\hat{F}_d(t)$ can be improved.

Existing auxiliary information is easily incorporated in the estimators of $F_d(t)$ under a model-based approach to inference.

The small area estimator of the cumulative distribution function under a model-based approach is

$$\hat{F}_d(t) = N_d^{-1} \left\{ \sum_{j \in s_d} I(y_j \leq t) + \sum_{j \in r_d} I(\hat{y}_j \leq t) \right\}, \tag{7.13}$$

where $s_d$ and $r_d$ represent the sample and non sampled units of small area $d$, respectively, $d = (1, \ldots, D)$ and $\hat{y}_j$ is a predictor for the $y_j$. The most popular class of models for small area to predict $\hat{y}_j$ are mixed models (chapter 1, model 1.1). Under these models the small area estimator of the $F_d(t)$ becomes

$$\hat{F}_d^{MX}(t) = N_d^{-1} \left\{ \sum_{j \in s_d} I(y_j \leq t) + \sum_{j \in r_d} I(x_j \hat{\beta} + \hat{u}_d \leq t) \right\}, \tag{7.14}$$

where $\hat{\beta}$ and $\hat{u}_d$ are obtained through the ML or REML procedures presented in chapter 1.

An alternative approach to the estimation of $\hat{F}_d(t)$ is based on M-quantile models (Chambers and Tzavidis, 2006). The M-quantile version of the $\hat{F}_d(t)$ estimator is

$$\hat{F}_d^{MQ}(t) = N_d^{-1}\left\{\sum_{j\in s_d} I(y_j \leq t) + \sum_{j\in r_d} I(x_j\hat{\beta}_\psi(\hat{\theta}_d) \leq t)\right\} \tag{7.15}$$

where the estimates $\hat{\beta}_\psi(\hat{\theta}_d)$ can be calculated through iterative weighted least square algorithms (chapter 6).

A semiparametric approach to the estimation of small area cumulative distribution function have also been proposed by Pratesi et al. (2008), following the methodology already considered in section 6.9. In this case the estimator can be expressed as:

$$\hat{F}_d^{NPMQ}(t) = N_d^{-1}\left\{\sum_{j\in s_d} I(y_j \leq t) + \sum_{j\in r_d} I(x_j\hat{\beta}_\psi(\hat{\theta}_d) + z_j\hat{\gamma}_\psi(\hat{\theta}_d) \leq t)\right\} \tag{7.16}$$

where $\hat{\beta}_\psi(\hat{\theta}_d)$ and $\hat{\gamma}_\psi(\hat{\theta}_d)$ are the coefficient vectors of the parametric and spline proportion of the fitted $p$-splines M-quantile regression function at $\hat{\theta}_d$. Pratesi et al. (2008) also consider a nonparametric bootstrap technique for estimating the mean squared error of the estimator (7.16).

Since the MQ and the NPMQ estimators are affected by bias, the CD correction (section 6.5) can be applied to the estimation of the small area cumulative distribution function as follows

$$\hat{F}_{CD,d}(t) = N_d^{-1}\left\{\sum_{j\in s_d} I(y_j \leq t) + \sum_{k\in r_d} n_d^{-1}\sum_{j\in s_d} I(\hat{y}_k + (y_j - \hat{y}_j) \leq t)\right\}.$$

The CD version of estimators (7.15) and (7.16) can then be obtained computing the corresponding $\hat{y}_k$ predictions. A first application of M-quantile models to estimate small area cumulative distribution functions can be found in Tzavidis et al. (2008), where estimates for the quantiles of the variable of interest in the small areas has been computed.

The estimation of the mean squared error of (7.13) is nowadays an open research problem.

## 7.7 References

Breidt, F.J. and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Ann. Statist.*, **28**, 1026-1053.

Breidt, F.J. and Opsomer, J.D. (2002). Design properties of semiparametric model-assisted estimators. *Working paper*, Iowa State University.

Chambers, R.L. (1986). Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association*, **81**, 1063-1069.

Chambers, R.L., Dorfman A.H., Hall P. (1992). Properties of estimators of the finite distribution function. *Biometrika*, **79**, 577-582.

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias Robust Estimation in Finite Populations using Nonparametric Calibration. *Journal of the American Statistical Association*, **88**, 268-277.

Chambers, R.L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, **73**, 597-604.

Chambers, R.L. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, **93**, 255-268.

Chen, J. and Wu, C. (2002). Estimation of Distribution Function and Quantiles using the Model-calibrated Pseudo Empirical Likelihood Function. *Statistica Sinica*, **12**, 1223-1239.

Dorfman, A.H. (1992). Non parametric regression for estimating totals in finite populations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 622-625.

Dorfman, A.H. (1993). A Comparison of Design-based and Model-based Estimators of the Finite Population Distribution Function. *Australian Journal of Statistics*, **35**, 29-41.

Dorfman, A.H. and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics*, **21**, 1452-1475.

Dorfman, A.H. and Valliant, R. (1993). Quantile Variance Estimators in Complex Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 866-871.

Durrant, G.B. and Skinner, C. (2006). Using Missing Data Methods to Correct for Measurement Error in a Distribution Function. *Survey Methodology*, **32**, 25-36.

Dunstan, R. and Chambers, R.L. (1989). Estimating Distribution Functions from Survey Data with Limited Benchmark Information. *Australian Journal of Statistics*, **31**, 1-11.

Falk, M. (1985). Asymptotic normality of the kernel quantile estimator. *The Annals of Statistics*, **13**, 428-433.

Francisco, C.A. and Fuller, W.A. (1991). Quantile Estimation with a Complex Survey Design. *The Annals of Statistics*, **19**, 454-469.

Harms, T. and Duchesne, P. (2006). On Calibration Estimation for Quantiles. *Survey Methodology*, **32**, 37-52.

Harrell, F.E. and Davis, C.E. (1982). A new distribution-free quantile estimator. *Biometrika*, **69**, 635-640.

Hyndman, R.J. and Fan Y. (1996). Sample Quantiles in Statistical Packages. *The American Statistician*, **21**, 361-365.

Johnson, A.A., Breidt, F.J., Opsomer, J.D. (2008). Estimating distribution functions from survey data using nonparametric regression. *Journal of Statistical Theory and Practice*, **2**, 419-431.

Kovacevic, M.S. (1997). Calibration Estimation of Cumulative Distribution Function and Quantile from Survey Data. *Proceedings of the Survey Methods Section, Statistical Society of Canada Meeting*, 1-7.

Kuk, A.Y.C. (1988). Estimation of Distribution Functions and Medians under Sampling with Unequal Probabilities. *Biometrika*, **75**, 97-103.

Kuk, A.Y.C. (1993). A Kernel Method for Estimating Finite Population Distribution Functions using Auxiliary Information. *Biometrika*, **80**, 385-392.

Kuk, A.Y.C. and Mak, T.K. (1989). Median Estimation in the Presence of Auxiliary Information. *Journal of the Royal Statistical Society, series B*, **51**, 261-269.

Kuk, A.Y.C. and Mak, T. K. (1994). A Functional Approach to Estimating Finite Population Distribution Functions. *Communications in Statistics, Theory and Methods*, **23**, 883-896.

Kuo, L. (1988). Classical and Prediction Approaches to Estimating Distribution Functions from Survey Data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 280-285.

Lombardia, M.J., Gonzalez-Manteiga, W., Prada-Sanchez, J.M. (2003). Bootstrapping the Chambers-Dunstan estimate of a finite population distribution function. *Journal of Statistical Planning and Inference*, **116**, 367-388.

Lombarda, M.J, Gonzlez-Manteiga, W. and Prada-Sanchez, J.M (2005). Estimation of a Finite Population Distribution Function Based on a Linear Model with Unknown Errors. *Canadian Journal of Statistics*, **33**, 181-200.

Mak, T.K. and Kuk, A. (1993). A New Method for Estimating Finite-population Quantiles using Auxiliary Information. *Canadian Journal of Statistics*, **21**, 29-38.

Meeden, G. (1995). Median Estimation using Auxiliary Information. *Survey Methodology*, **21**, 71-77.

Modarres, R. (2002). Efficient Nonparametric Estimation of a Distribution Function. *Computational Statistics and Data Analysis*, **39**, 75-95.

Nascimento Silva, P.L.D. and Skinner, C.J. (1995). Estimating Distribution Functions with Auxiliary Information using Poststratification. *Journal of Official Statistics*, **11**, 277-294.

Oberhofer, W. and Haupt, H. (2005). The asymptotic distribution of the unconditional quantile estimator under dependence. *Statistics & Probability Letters*, **73**, 243-250.

Pratesi, M., Ranalli, M.G., Salvati, N. (2008). Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US. *Environmetrics*, **19**, 687-701.

Rao, J.N.K., Kovar, J.G., Mantel, H.J. (1990). On estimating distribution function and quantiles from survey data using auxiliary information. *Biometrika*, **77**, 365-375.

Rueda, M., Martinez, S., Martinez, H., and Arcos, A. (2007a). Estimation of the Distribution Function with Calibration Methods. *Journal of Statistical Planning and Inference*, **137**, 435-448.

Rueda, M., Martinez, S., and Snchez, I. (2007b). Estimation of the Distribution Function using Non-parametric Regression. *Technical Report*, University of Granada.

Rueda, M., Arcos, A. and Artes, R. (1998). Quantile interval estimation in finite population using a multivariate ratio estimator. *Metrika*, **47**, 203-213.

Rueda, M. and Arcos, A. (2004). Improving ratio-type quantile estimates in a finite population. *Statistical Papers*, **45**, 231-248.

Särndal, C.E., Swensson, B.E. and Wretman, J.H. (1992). *Model assisted survey sampling*. Springer, New York.

Sheather, J. and Marron, J.S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, **85**, 410-416.

Shuster, E.F. (1973). Median On the Goodness-of-fit Problem for Continuous Symmetric Distributions. *Journal of the American Statistical Association*, **68**, 713-715.

Singh, S., Joarder, A.H., and Tracy, D.S. (2001). Median Estimation using Double Sampling. *Australian and New Zealand Journal of Statistics*, **43**, 33-46.

Sitter, R.R. and Wu, C. (2001). A Note on Woodruff Confidence Intervals for Quantiles. *Statistics and Probability Letters*, **52**, 353-358.

Tzavidis, N. and Chambers, R.L. (2007). Robust prediction of small areas means and distributions. *CCSR Working paper*.

Tzavidis, N., Salvati, N., Pratesi, M., Chambers, R.L. (2008). M-quantile models with application to poverty mapping. *Statistical Methods & Applications*, **17**, 393-411.

Wang, S. and Dorfman, A.H. (1996). A New Estimator of the Finite Population Distribution Function. *Biometrika*, **83**, 639-652.

Welsh, A.H. and Ronchetti, E. (1998). Bias–calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society, Series B*, **60**, 413-428.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, **47**, 635-646.

Wu, C. and Sitter, R.R. (2001a). A Model-Calibration Approach to using Complete Information from Survey Data. *Journal of the American Statistical Association*, **96**, 185-193.

Wu, C. and Sitter, R.R. (2001b). Variance estimation for the finite population distribution function with complete auxiliary information. *The Canadian Journal of Statistics*, **29**, 289-307.

# LITERATURE REVIEW

# VOLUME III

| | |
|---|---|
| Project Acronym: | SAMPLE |
| Project Full title: | Small Area Methods for Poverty and Living Conditions Estimates |
| Project/Contract No: | EU-FP7-SSH-2007-1 |
| Grant agreement No: | 217565 |
| Work Package 3: | Integration of EU-SILC with administrative data |
| Document title: | WP3 Literature review |
| Date: | 30th March 2009 |
| Type of document: | Deliverable D4 |
| Status: | Final |
| Editors: | C. Rognini, M. Casarosa, A. Giunta |
| Authors: | Claudio Rognini (c.rognini@provincia.pisa.it), PP-UROPS<br>Michela Casarosa (m.casarosa@provincia.pisa.it), PP-UROPS<br>Antonella Giunta (a.giunta@provincia.pisa.it), PP-UROPS |

# Contents

**Prologue** **II**

**1 Poverty indicators used by the Tuscany Region** **1**

1.1 IRPET indicators – Tuscany Regional Institute for Economic Planning ........ 2

1.2.Institutional Information Systems: Provincial Social Observatories and Health's Societies ........ 6

1.3.Indicators from the third sector ........ 6

**2 Poverty Indicators used in the Province of Pisa** **7**

2.1. Data and Indicators Developed by the UROPS ........ 7

2.2.The status of initiated processes ........ 9

**3 Conclusion** **18**

# Prologue

The following report presents the current status of the instruments and indicators used to measure and monitor the phenomenon of poverty and social inequality by *policy makers* in planning and implementing efficient and integrated policies to reduce and prevent poverty. The territorial levels examined are Italy in the conceptual framework and European methodology, the Tuscany Region (NUTS 2), the Province of Pisa (NUTS 3) and in particular, the local sub-provincial governments made up of aggregate municipalities (LAU 2). The Tuscany Region and the Province of Pisa are territories where the SAMPLE project shall conduct the empirical part relative to the development of the stimulus model for small areas and the permanent observatory for the analysis and monitoring of poverty phenomena.

The report is broken down into two parts:

- The first part consists of a brief overview of the spin-offs from the Lisbon strategy at the level of the Tuscany Region and explores the indicators used for its regional programming efforts, in particular, those for the Regional Development Plan (PSR, according to the Italian initials) and the Regional Social Integration Plan (PISR);
- The second portrays the current status of the results assembled from the UROPS in their review of the data and the development of a multi-method approach in acquiring knowledge of the phenomenon of poverty and vulnerability.

Before going on to the analysis of the individual parts for the territories analysed, a summary of the current conceptual approach is described as it is used by policy makers and the persons responsible for poverty in these territories.

The multi-dimensional approach to poverty is shared acquired knowledge, in the sense that the dimension of revenues and material deprivation, even though it remains crucial, is inserted into a more complex dimension of quality of life. A shared and harmonized determination of the indicators required to empirically measure the multidimensionality does not yet follow this approach. In fact, Laeken's indicators are able to create a homogenous and cognitive basis between the Member States of the European Union as we shall analyse later on. Their implementation, limited to a regional territorial level, has created a fertile field for the development of autonomous strategies to identify the information and construct indicators at the sub-regional government level, resulting in various analyses that cannot be compared with each other.

By applying five topic areas that are "vectors" to the documents analysed1, using a cognitive framework developed by CERFE2 it is possible for us to affirm that:

1. There is a consolidated convergence in the literature and national and regional regulations as to the **centrality of the knowledge of poverty**, its multidimensional nature, its intensity, and the dynamics of the phenomena of deprivation present in the territory. These serve as an essential basis for taking coherent and effective steps in the policies to fight and prevent poverty (**first vector**). Awareness of how this is crucial arose over time with the terminology moving progressively closer as to the conceptual multidimensional definition of poverty and the resultant need to ensure that the methodological approaches were also multidimensional;

2. What has still not been suitably put into practice and applied to anti-poverty policies is the **subjectivity of the poor** understood as an attitude so as to no longer identify the poor as passive subjects that are beneficiaries of state interventions, but rather as players in anti-poverty policies (**second vector**), notwithstanding the fact that this approach is meeting with ever-increasing agreement in the debate surrounding the dynamics and the effects of being poor and vulnerable. Beyond directly listening and becoming involved with poor and vulnerable persons, through actions aimed at their *empowerment*, very little has been developed or analysed in relation to the most advanced European experiences and good practices formulated by the European Anti Poverty Network or EAPN in the territories under review. The EAPN promotes the direct involvement of the poor and vulnerable in selecting the institutional and third-sector interventions that concern them, as will be analysed below;

3. There is a growing convergence in acknowledging the **social-dynamic nature of poverty** in complex societies. In the documents analysed, the image of poverty as a static condition actually overlaps either in the theoretical or political approach, highlighting how poverty is a process. In fact, by inserting the variable of time, poverty can be intermittent, persistent, recurring, chronic and intergenerational. From this understanding the longitudinal measures that are able to take advantage of these dynamics have been developed. However, in the territories under review they seem to have been developed to a very limited extent, just as the life story approaches[3] have not been very developed. These approaches highlight the the dynamics of the processes of impoverishment by identifying the cumulative micro-fractures that are often borne by the subjects in solitude and cause their descent into poverty (**third vector)**;

4. The **differentiated typology of the poor and vulnerable groups** is recognized in the territories and it is precisely this awareness that calls for development of appropriate methodological instruments to

[1] An initial application of the vectorial framework was done by analysing the following material: the PRS 2006-2010 and the PISR 2007-2010; a regional and provincial bibliography. The considerations expressed herein also bear on the results of the research and actions undertaken by UROPS, to be further described below, on Poverty, Vulnerability and Living Conditions of Families in the Province of Pisa carried out with the Health's societies and subjects from third sector associations of the provincial territory.

[2] Giancarlo Quaranta, Gabriele Quinti (under the supervision of), *Social Exclusion and Poverty. Contribution to learning and measuring social and environmental risks in the international context* CERFE, Rome, 2005.

[3] An application of this approach will be recounted later. It was promoted by the UROPS and executed by the Department of Social Sciences of the University of Pisa.

measure the exposure and risks of vulnerability and how these individuals become more fragile. This is mostly associated with those cases having a reduced capacity to protect themselves due to a lack of valuable resources, connections and support networks (**fourth vector)**;

5. Intimately connected to the previous vectors is the understanding that *policy makers* urgently need to promote **concrete policies** aimed at preventing and fighting the various phenomena of poverty and social exclusion. In other words, they need to develop articulated and appropriate policies based on the characteristics of the vulnerable individuals involved, as well as the characteristics and dynamics of the territory they live in. This will allow them to be more effective in dealing with the dynamic aspects of poverty (**fifth vector)**.

The establishment of concrete policies is an ongoing procedural effort that needs to be further reinforced in relation to this fifth vector.

# Chapter 1

# Poverty indicators used by the Tuscany Region

The Tuscany Region is involved in national efforts to draft the PAN-Incl. and is active in the European Network named RETIS whose objectives are to develop instruments to analyse poverty and policies and projects aimed at its prevention and deterrence.

Viewing poverty and social exclusion as a phenomenon, pursuant to the multidimensional acceptance stated in the introduction, the source of data and qualitative and quantitative indicators used by the Regional Council to develop its programs are mainly comprised of the following institutional resources and by Caritas, a non-institutional resource.

- Regional Statistics System;
- IRPET – Regional Institute for Economic Planning;
- ARS – Regional Health Agency;
- Regional Network of Social Observatory (OO.SS according to Italian abbreviation);
- Health's societies' System;
- Database from the CARTIAS Regional Welcome Centres.

The Regional Council therefore takes into account the analysis developed by these institutional actors and Caritas when preparing their **regional action plan.** The Regional Development Program (PRS), receives the European instructions on the Strategy for Sustainable Development and the Lisbon Strategy, and the instructions and planning document whereby the Regional Council from Tuscany defines the directions and actions to be undertaken for the integrated development of the territory based on economic, social and environmental sustainability. It does this in close cooperation with the current Government Program established by the legislature. The PRS identifies the strategic choices of the regional action and the legislature's priorities through the Integrated Regional Projects. On the one hand, it links up with the legislature's sectorial programs to determine the operations and planning. On the other hand, it acts on the project plans that were chosen and developed by the individual provincial territories through the Local

Development Conventions[1]. The Regional Council arrives at the Regional Development Program though a *governance* model with representatives from all facets of Tuscan society, respecting the specific institutional competencies. The Integrated Social Regional Plan (PISR), which more specifically intervenes in the integrated social policies, is a sectorial plan aimed at putting into effect the social rights of citizens. On the basis of the analysis of context and the indicators of social hardships it establishes the parameters for allocating regional funds for social policies.

## 1.1 IRPET Indicators - Tuscany Regional Institute for Economic Planning

The IRPET provides analysis and in-depth knowledge of the Tuscany Region. Some of the principal studies carried out on the living conditions of Tuscan families and poverty were the following: two reports on well-being; periodic estimates of poverty; an analysis of the distributive effects of the main fiscal provisions (tax reform) and social (introduction of guaranteed minimum income, the Isee (Indicator of Equivalent Economic Situation), the contribution for the purchase of a house for young couples, the check for newborns, the tax for financing a fund for persons that are not self-sufficient, etc.) that have an impact on the standard of living of Tuscan families. Since 2002, together with the "C. Dagum" Cridire Research Centre at the University of Siena (the partner in the SAMPLE project), it has initiated a survey on the living conditions in Tuscany (ICVFT). The survey was conducted in 2002, and repeated in 2004, and made it possible to gain further understanding of social mobility, the dispersion of revenues, the phenomena of economic and multidimensional poverty, housing conditions, and the perception of living conditions. The analysis provided by these studies all have a direct bearing on the programming, implementation and evaluation of the economic and social programs and the comprehensive strategy for development for the Tuscany Region.

The studies have for the most part been regional, with provincial information and in some cases, they have taken into account the aggregated municipalities of the SEL (Local Tuscan Economic Systems). The ICVFT prepares the data at a regional level. In 2006, with the publication of the **Poverty Mapping for the Tuscany Region**[2], IRPET analysed the territorial distribution of poverty and inequality, generating estimates at the territorial and municipal levels and consequently, at the higher territorial levels: provinces and SEL.

The computation methods used[3] combine two different statistical sources. One uses a sampling technique, namely the Inquiry into the Living Conditions of Tuscan Families (ICVFT), whereas the Census of Population and Households refers to the entire population of families. Through a regressive linear model

[1] The PSR also includes a programmatic reference framework of the new European programmes and the interventions related to the Framework Programme Agreements with the national government and is linked with the strategic choices of the Territorial Direction Programme (PIT); finally, it represents the guiding instrument on how to use regional, national and EU financial resources, allocated in provisional terms between the various priorities identified with the PIR – Regional Law 49/1999 "Regulations for regional programming."

[2] IRPET, Mapping Poverty in Tuscany, Florence, July 2006

[3] According to the information provided by the authors in their introduction to the text, the procedure was used by the World Bank to monitor poverty and evaluate policies in Albania. These were used to identify which measurement indicators were required to decide how funds and transfers would be distributed in the territory so that the areas most in need economically and socially would benefit.

with variable components applied to the data collected in the ICVFT, an estimate of the relative distribution of the dependent variables is arrived at (available income). This estimate can be used to generate an analogous distribution for each census unit, conditionally based on the observed characteristics. The co-variables used in the model, that is, the explanatory variables from which the dependent variables are taken, are those present in the two records.

This methodology makes it possible to associate an income value for all Tuscan families and link it to the census. From this conditional distribution, it extracts a series of poverty and inequality measures articulated to the municipal level. Then, the application of the estimate to small areas makes it possible to calculate the interest parameter values for areas not represented in the sample.

The computation methodology followed two procedural steps:

a. From the sample data it arrived at a model capable of estimating family income;

b. Subsequently, one arrives at the income level for each family in the census, taking into account relative margins of error. These figures are then calculated using the traditional indexes (headcount ratio, income gap ratio, Gini, Atkinson, etc.) for poverty and inequality in the territorial framework.

For each territorial unit the following measures of poverty and income inequality were calculated:

| | |
|---|---|
| 1 | Equivalent average family income |
| 2 | Measures of poverty and inequality |
| | a) *diffusion spread* |
| | b) *Foster, Greer and Thorbecke indexes* |
| | c) *Sen index* |
| | d) *Gini index* |
| | e) *Gini index regarding income of poor families* |
| | f) *Atkinson index* |
| | *g) average logo-rhythmic deviation* |
| | h) *inequality measurements* |

The following information was used from the census data, the most recent one having been the **14th Census of Population and Housing (2001)**:

| | |
|---|---|
| 1 | Population counted and notations made on structural characteristics |
| 2 | Determination of legal population |
| 3 | Information gathered regarding the numerical consistency and structural characteristics of housing (buildings also counted in census: housing use, and in centres lived in, including those not intended for housing) |
| 4 | Structural and family demographics of the resident foreign population |
| 5 | Family typology |
| 6 | Level of highest academic degree attained either in Italy or abroad |
| 7 | Professional status and information regarding non-residents |

The two sources of information were analysed in order to study their comparability as well as to construct common variables from the two databases with the same distribution. The information culled from the Census and Inquiry relate to persons who customarily stay in the area and their housing situation. The census also gathered information on buildings.

**Collective data common to the two records regarding housing:**

| | |
|---|---|
| 1 | Deed of use for housing |
| 2 | Housing area in square metres |
| 3 | Existence of shower and bathtub |
| 4 | Existence of sanitary fittings inside housing |
| 5 | Existence of hot water |
| 6 | Existence of heating equipment |
| 7 | Existence of private box, parking spot or garage |
| 8 | Existence of fixed and working telephone line |

**Collective data common to the two records regarding individuals:**

| | |
|---|---|
| 1 | Information on kinship ties between individuals |
| 2 | Age |
| 3 | Civil Status |
| 4 | Highest academic degree attained |
| 5 | Professional title and status, type of work contract held |
| 6 | Information regarding number of hours worked in the week preceding the interview |
| 7 | Economic business sector |

Carrying out the various phases of the process involved in producing the data made it possible for IRPET to structure the definitive working records constituting a basis for the micro-data that can be updated. This facilitates the monitoring the living conditions of families in Tuscany.

## 1.2 Institutional Information Systems: Provincial Social Observatories and Health's Societies

Other sources of regional information of note used for planning the territorial policies of the Tuscany Region are the Regional System of the Provincial Social Observatories and the Health's Societies: Public Consortiums whose principals are the local and communal health organizations.

The **Provincial Social Observatories,** including the Social Observatory for the Province of Pisa – UROPS which is a partner in the project, were established in 1997 by the Tuscany Region with the legislative task of contributing to the Information System for Social Policies providing **cognitive data on the needs of the population in the provincial territories under its jurisdiction** in order to design social policies and promote the informed participation of third-sector subjects and citizens in the overall logistics of promoting citizenship. They periodically produce surveys and carry out local investigations with a common territorial framework for their reference areas. Over time they have additionally developed, with some differences between the various provincial realities, analogous Statistical Information Databases and produced the annual Statistical Dossiers and Social Reports.

The **Health's societies**, (SdS according to Italian abbreviation) represent one of the most important innovations in the integrated health program and social assistance and they have a specific planning task for the integrated social and health policies. Their role is to encourage the involvement of local communities, social parties, third sector associations, and volunteers in **identifying health-related needs** (according to the WHO definition) in the **planning process**. For the SdS, local action is an essential element in the regional strategy to promote health and is not limited just to social and health issues, but also includes the improvement of health through an integrated inter-sectorial policy capable of influencing the **determining factors affecting the health of populations** and the quality **of the environment.**

The recent law of Regione Toscana (LR 68/09) has consolidated the Health's Societies which have the function of social and sanitary planning at supra-municipal level.

The SdS programming is formulated using the **Integrated Health Plan** instrument. Together with the **Regional Health Agency,** the SdS develops a shared cognitive system based on health markers of the population in the territories under its jurisdiction. They develop **Health Profiles and Health Portraits** every three years.

Since may 2008 UROPS participate at Regional Network of Social Observatory (OO.SS according to Italian abbreviation) coordinated by the Regional Social Observatory

One of the aims of the Regional Network is to create a regional common set of indicators in social policies and a common methodology for the acquisition of the dates

The OO.SS has selected 130 indicators to estimate the health state of local population and to evaluate local social policies realized by the Health's Societies.

For every indicator we have identified the provider and the source. Every indicator will be provided at supra-municipal level.

These activity is a first important step for the development of the "Observation System to monitor poverty, vulnerability and social exclusion" (Task 3.4).

These are the main areas which will be monitored:

- Demographic profile
- Health state (travel accidents, hospitalize, infectious desease)
- Essential level of territorial health care (sanitary services, social care)
- Elderly persons (health indicators, socio-demographic indicators, not self sufficiency)
- Families and youngers
- Immigration
- Disability
- Mental health
- Dependences

## 1.3. Indicators from the third sector

Within the framework of third sector associations, a cognitive data source for the Tuscany Region is represented by the analysis produced by the **Observatory of Resources and Poverty,** which is part of CARITAS (An Episcopal organization of the Italian Episcopal Conference), which has a national structured database from its territorial Welcome Centres. The Immigration Reports that are produced annually are also used in analysing and monitoring poverty and social exclusion. These reports represent an important and relevant cognitive source for the national, regional and provincial territory.

# Chapter 2

# Poverty Indicators used in the Province of Pisa

## 2.1. Data and Indicators Developed by the UROPS

In the provincial territory of Pisa, the municipalities, the ASL organizations, the Social-Health Areas have produced important studies over time on poverty and social exclusion specifically focus on particular target populations. The analysis produced has been limited and occasional. As part of its institutional responsibilities, the Observatory for Social Policies UROPS)[4] has produced cognitive reports on the social situation in Pisa. It has also conducted research and studies on the specific needs of various population groups (youth, adults, elderly persons living alone) that depict the social fragility of the region and the needs and vulnerabilities of its residents.

Though quite significant, the information and assessments of the phenomenon of poverty and social exclusion produced by Eurostat, Istat, as well as the sample surveys conducted by Eurispes and Censis, but also at the regional level in studies by IRPET, have up to now not taken into consideration the territorial details that are required and useful for *policy makers* in local government. In light of these considerations, the OPS has decided to strengthen its Observatory model[5] by developing a specific Information Section on Poverty, Vulnerability and Living Conditions of Families in the Province of Pisa: This is the reason that in

[4] The Observatory for Social Policies (OPS) , an organization established under Tuscany Regional rules, was founded in the Province of Pisa in 1999. Together with the provinces of Lucca, Leghorn and Massa Carrara, it prepares a model Social observatory of a Vast Area developed by the Department of Social Sciences of the University of Pisa. Over time, the OPS has developed a **statistical information database** aimed at harmonizing the different information records. Not all of these records have an analogous level of development: a) quantitative statistical data from administrative sources from the provincial territory; b) data on utilization from Social Services in order to allow for a constant monitoring of their needs and the interventions undertaken; c) data produced from quantitative investigations (surveys) and in-depth qualitative inquiries on certain topics. The territorial detail goes up to the municipal level and aggregates the data in a functional manner to the social and health programmes, according to the Health's societies. The data collected is published in research journals, statistical dossiers and reports with analysis of the statistical data that highlight the main socio-demographic trends, assets and risk and social vulnerability factors at the provincial and sub-provincial level to be used as cognitive instruments for local government.

[5] The following events were organized as part of the research and action initiatives: Study seminar on poverty "Reading poverty in the territories to fight it more effectively," Conference on "Local Development and Social Inclusion" aimed at involving social parties and trade associations of the provincial territory in the debate on poverty and social exclusion.

June 2006 it promoted provincial research and actions to progressively involve the Health's societies in the territory and then the non-institutional *stakeholders*. It first used the bottom-up **Open Coordination Method** with the cooperation of the Health's societies in the region, and then that of the non-institutional *stakeholders*. The aim of the process was to create a common concept of poverty and vulnerability in order to make it possible to share methodologies and indicators for measuring and monitoring the phenomenon in such a way that would be homogenous for the provincial territory. It has disseminated knowledge in the provincial territory of the surveying instruments, indicators and protocols proposed by the European Commission.

As part of its research and action efforts, UROPS has engaged the first two local institutions by signing **Memorandums of Understanding:**

- The first one is with **CARITAS** and it seeks to exchange data on observing poverty phenomena enabling URPS to have access to the data on absolute poverty gathered by the Caritas Pisa Welcome Centres thereby promoting a united reading and analysis of the phenomenon as it occurs in the provincial territory;

- The second was signed with **INPS** to reach a joint determination from the INPS database of the information relevant to intercepting poverty phenomena and provide UROPS with access to the database.

Finally, to arrive at its objectives, UROPS sponsored an **internal working group from the Province of Pisa** together with the contact persons from the Town Councillorships for Economic Development and Productive Activities; Education; Training and Work thereby promoting good institutional practices within the Agency itself.

The actions undertaken by UROPS, with the participation of the Health's societies of the provincial territory endeavoured to build a shared permanent territorial analysis and monitoring system of the phenomenon of poverty using a multi-method approach and capable of identifying the various aspects of the complexity of the phenomenon. The different approaches applied to the system should constitute records that can be updated and integrated between them and will be strengthened by the creation of the SAMPLE project.

1. Identify **ecological type data** relevant to measuring the phenomenon of poverty, present in the databases and institutional records already in existence in the territory as well as the national administrative data, **building a concise measuring indicator of social hardship.**

2. Conduct a **survey of families, individuals and poverty in the provincial territory,** in order to identify the extent of poverty ascertained by the European Commission and measured through the EU-SILC investigation, for the first time connecting to the wave scheduled for 2008.

3. Find additional dimensions to the processes of impoverishment through **qualitative and in-depth investigations using the life story approach.**

4. By developing the suitable methodological instruments, gather and analyse the **data/information in possession of the non-institutional *stakeholders*** produced from their direct observations and experience with deprived and vulnerable citizens.

5. Promote **self-evaluations by Social Services on their own** policies and interventions to prevent and fight poverty and social exclusion.

6. Promote **awareness of the phenomenon among policy makers and stakeholders in the territory**

7. **Link up with European Institutional networks as well as third sector networks** operating in the thematic field of poverty and social inclusion.

## 2.2. The status of initiated processes

Below is a presentation of the individual objectives, the processes that were initiated, and the instruments developed to reach the objectives.

**Objective 1**: Identify **ecological type data** related to measuring the phenomenon of poverty, that is found in the institutional databases and records that are present in the territory and national administrative data, **building up a concise indicator for measuring social hardship**.

During the monthly meetings with the established work groups, the members examined that statistical data present in the Informational Statistical Database of the Provincial Social Observatory[6], and they identified which information was relevant to poverty, which ones had already been used by the Health's societies and what new data was necessary for understanding the phenomenon of poverty and vulnerability (Attachment 1 Platform of shared data).

This sharing process was followed by two work phases:

A. Collection and analysis of the first data

B. Construction of the concise indicator of social-territorial hardship

## A. Collection and Analysis of the Primary Data

The analysis of the data was published in December 2006 in the first **Preliminary Report on "Poverty, Vulnerability and Living Conditions of Families in the Province of Pisa"** under the auspices of UROPS. The data analysed is listed below:

[6] The OPS Statistical Information Database collects and updates data on the provincial territory, with a territorial breakdown as well as a municipal one when available in the fields of *Demography, Foreign Citizens, Families, Housing, Education, job Market, Social Security, Health, Justice, Safety, Third Sector Associations*.

| | **Data and Indexes** | **Source** |
|---|---|---|
| | **Demographic Aspects** | |
| 1 | Evolution of the population per municipality 1951-2005 | ISTAT and Municipal Registers |
| 2 | Generic birth and fertility rates in the territories of the Health's societies (SdS) 2005 | ISTAT and Municipal Registers |
| 3 | Evolution of the foreign population in the municipalities 1951-2005 | ISTAT and Municipal Registers |
| 4 | Impact of the foreign population on the resident population in the SdS – 2005 | ISTAT and Municipal Registers |
| 5 | Countries of origin of the foreign citizens in the SdS – 2005 | ISTAT |
| 6 | Structured age survey of the foreign population according to gender and totals for the province – 2005. | ISTAT |
| 7 | Structured age survey of the foreign population according to gender and totals for the province - 2005 | ISTAT |
| 8 | Resident population according to civil status and gender per municipality – 2005 | ISTAT |
| 9 | Indexes on the elderly, dependency, exchange, structure of the working population per province – 2005 | according to ISTAT data |
| | **Family Components** | |
| 10 | Type of families (couples with children, without children, single parents mother/father) – 2005 | ISTAT – Census |
| 11 | Evolution in the number of families and average size 1996 – 2005 | ISTAT |
| 12 | Families according to number of members (1-2-3-4-5-6 and more) – 2005 | ISTAT |
| | **Housing Components** | |
| 14 | Housing occupied by residents according to type of ownership and use - 2001 | ISTAT – Census |
| 15 | Government housing: number of rooms vacant and occupied per income category - 2005 | Pisa Agency for Government Housing |
| 16 | Number of assignees per age category – 2005 | Pisa Agency for Government Housing |
| 17 | Number of users per age category – 2005 | Pisa Agency for Government Housing |
| 18 | Request for government subsidy to pay rental: number of families requesting aid according to nationality (Italian citizens/foreign citizens) per municipality | Municipalities: Housing Offices and Social Policies Offices |
| 19 | Evolution of eviction orders (issued – request for execution – executed) 2001-2005 per province | Ministry of the Interior on community data |
| | **Educational Components** | |
| 20 | Distribution of (lower) secondary school degrees per gender and per province | Scholastic Observatory Province of Pisa |
| 21 | Educational degrees attained by parents of students in (lower) secondary school classes. | Scholastic Observatory Province of Pisa |
| 22 | Repeating students and non-repeating students – (lower) secondary school. classes | Scholastic Observatory Province of Pisa |
| 23 | Academic counselling and training per province (lower classes) | Scholastic Observatory Province of Pisa |

| | **Data and Indexes** | **Source** |
|---|---|---|
| 24 | Choice in type of education for students depending on lower secondary school diploma | Scholastic Observatory Province of Pisa |
| | **Occupational Components** | |
| 25 | Female unemployment in the Province of Pisa per municipality – 2001 | ISTAT – Census |
| 26 | Male unemployment in the Province of Pisa per municipality – 2001 | ISTAT – Census |
| 27 | Citizens available for work (unemployed – not working – precarious with an annual income under 7,500 Euro) per gender and per Employment Centre in the Province of Pisa | Employment Centres - Province of Pisa |
| 28 | Foreign citizens available for work (unemployed – not working – precarious with an annual income under 7,500 Euro) per gender and per Employment Centre in the Province of Pisa | Employment Centres - Province of Pisa |
| 29 | Citizens available for work (unemployed – not working – precarious with an annual income under 7,500 Euro) age category, gender and Employment Centre in the Province of Pisa | Employment Centres - Province of Pisa |
| 30 | Foreign citizens available for work (unemployed – not working – precarious with an annual income under 7,500 Euro) age category, gender and Employment Centre in the Province of Pisa | Employment Centres - Province of Pisa |
| | **Social Security Components** | |
| 31 | Evolution of the ordinary unemployment subsidies 2001-2005 | INPS |
| 32 | Evolution of the unemployment subsidies in the agricultural sector 2001-2005 | INPS |
| 33 | Evolution of Mobility Indemnity 2001-2005 | INPS |
| 34 | Evolution of Redundancy Fund 2001-2005 | INPS |
| 35 | Pensions supplemented by Guaranteed minimum income (427.58 euros per month) – 2005 | INPS |
| 36 | Pensions supplemented by Guaranteed minimum income according to category (elderly – disability – reversibility) per municipality – 2005 | INPS |
| 37 | Pensions supplemented by Guaranteed minimum income per type (elderly – disability – reversibility), and gender and per municipality – 2005 | INPS |
| 38 | Total pensions according to amounts (<500; 500-999; 1000-1499; 1500-1999; 2000-2499; 2500-3000; >3000) – 2005 | INPS |
| 39 | Total pensions per gender and municipality – 2005 | INPS |
| | | |
| | **Health Components** | |
| 40 | Request for exemption from paying health services co-payment based on income – Elderly per age group (65-69; 70-74; 75-79; 80-84; 85-89; >90) and gender per municipality – 2005 | USL 5 Health Organization of the Province of Pisa |
| 41 | Request for exemption from paying health services co-payment based on income – Children below 6 years of age per municipality – 2005 | USL 5 Health Organization of the Province of Pisa |
| 42 | Request for exemption from paying health services co-payment based on income (self-certifying) – 2005 | USL 5 Health Organization of the Province of Pisa |

Additionally, the data was also represented in geographic reference maps at the municipal and provincial levels. A first tentative representation of the sub-municipal level was attempted as an experiment according to census area, using the data from the 2001 Census.

The Report was made public and itemized in 2007 in the territories of the Health's societies, stimulating the debate, promoting awareness of new approaches in evaluating the phenomenon and involving more territorial *stakeholders.*

## B. Creating a Concise Indicator of Social-Territorial Hardship

The Department of Statistics, Economics and Applied Mathematics at the University of Pisa created a Concise Indicator of Territorial Hardship. By using this indicator, an initial classification and mapping of the municipalities in the province was produced. The indicator was put together by determining the following specific elements:

| | **Indicators** | **Source** |
|---|---|---|
| | **Demographic Components** | |
| 1 | Percentage variation of the population between 2001 and 2006 | ISTAT |
| 2 | Percentage of population => 75 years-old | ISTAT |
| 3 | Dependency index of elderly population | ISTAT |
| | **Family Components** | |
| 4 | Percentage of large families (with 5 or more members) | ISTAT – Census |
| 5 | Percentage of families headed by single mother with children | ISTAT – Census |
| 6 | Percentage of single-member families 65 years of age and older | ISTAT – Census |
| | **Housing Components** | |
| 7 | Percentage of housing without heating occupied by residents | ISTAT – Census |
| 8 | Percentage of housing without hot water occupied by residents | ISTAT – Census |
| 9 | Over-crowding index | ISTAT – Census |
| | **Human Capital Components** | |
| | ***Quality of Human Capital*** | |
| 10 | Percentage of population between the ages of 15 and 52 not completing mandatory schooling | ISTAT – Census |
| 11 | Percentage of population aged 19 years and older who has obtained at least one Level II secondary school diploma | ISTAT – Census |
| | ***Employment of Human Capital*** | ISTAT – Census |
| 12 | Female unemployment rate | ISTAT – Census |
| 13 | Unemployment rate of young adults | ISTAT – Census |
| | **Economic-Social Components** | |
| 14 | Percentage of INPS pensions below 500 euros based on total INPS pensions | INPS |
| 15 | Ratio between resident population and working population | ISTAT – Census |

The results from applying the hardship indicator were published in a chapter of the methodological presentation and results of the municipalities' application of the Province of Pisa of the indicator in the "Report on the Social Situation of the Province of Pisa" under the auspices of UROPS **and** published in February 2008

**Objective 2:** By means of surveys of families, individuals and poverty in the provincial territory, **determine the extent of poverty identified by the European Commission and ascertained in the EU-SILC investigation,** linking up with the first wave scheduled for 2008.

The financial unfeasibility of conducting an investigation as complex as the one recommended by the EU-SILC of the territory lead to submitting an application so that the SAMPLE project would receive European financing. It is as part of this project framework that the sampling will be done with 800 families from 166,429 (as of 31 December 2000 who are residents of the provincial territory, according to the procedures and goals presented in the project. In 2006, the Department of Statistics, Economics and Applied Mathematics at the University of Pisa (project leader) prepared an estimate based on small areas using the micro data of the ICVF investigations made available to the IRPET. It performed this estimate as part of the preliminary Report mentioned above. As regards the methodology used to arrive at the estimate, please refer to the direct contributions in this *deliverable* prepared by the project's University partners.

**Objective 3:** Identify additional aspects of the processes of impoverishment through qualitative and in-depth investigations through the life story approach

To reconstruct the paths leading to impoverishment and coping strategies for material deficiencies and learn how social exclusion is experienced by individuals, a biographic investigation of subjects that turned to the Caritas Welcome Centres in the territory were conducted. The investigation was made possible thanks to the UROPS-Caritas Memorandum of Understanding and was carried out by the Department of Social Sciences of the University of Pisa with the direct involvement of employees of the Welcome Centres who conducted the interviews. The in-depth interviews with the persons selected who have used the Centres for several years, and how many have successfully found a way out of the poverty trap was done by means of biographical-narrative interviews. The aim was to reconstruct the subjective paths to impoverishment, and place the narrators' points of view at the forefront in order have the plurality of dimensions and representations of the conditions of poverty emerge. These biographies would also serve to analyse the daily life experiences and the links with the past, as well as the future prospects of the persons interviewed. The following components were identified:

| | **MAP of CONSIDERATIONS for qualitative indicators** |
|---|---|
| | **Main social-relationship contexts (problems confronted in daily life)** |
| 1 | Work |
| 2 | Family |
| 3 | Friendly relationship networks, groups and communities |
| 4 | Educational and training background |
| 5 | Support services (governmental, non-profit, networks). |

| | **Temporal Components (background – previous experiences – future prospects)** |
|---|---|
| 6 | Current conditions |
| 7 | Past experiences |
| 8 | Vision of the future/planning actions and strategies and identification of resources/connections |
| | **Spatial Components** |
| 9 | Place of origin |
| 10 | Residential transfers and/or migratory experiences |
| 11 | Daily life environment |
| 12 | Housing |
| 13 | Subject's housing history |
| 14 | Relationship with public and private spaces |

The outcome of the interviews was published in the already-mentioned Report on the Social Situation of the Province of Pisa.

**Objective 4:** Whilst developing the appropriate methodological instruments, collect and analyse the **data/information in the custody of non-institutional *stakeholders*** in the territory, resulting from their direct observations and experiences with impoverished and vulnerable citizens.

To bolster knowledge regarding the complexity of the phenomenon of poverty, it is necessary to collect even extreme elements that often escape the methodological approaches already mentioned. Subjects that are part of the third citizenship sector have informative reservoirs of interests, and the lack of structure in most cases, as well as the impossibility of comparing data often makes it impossible to properly take advantage of the knowledge produced in them for the benefit of the territorial analysis. In order to convert their data/informational assets into available knowledge for use by the local government, the UROPS, together with the SdS collaborators started certain actions aimed at reaching this objective. In 2007, the SdS collaborators identified the *non-institutional stakeholders* involved in preventing and fighting social exclusion[7]. This is the reason for establishing decentralized sub-working groups in the individual territories that followed the working group's reflections. They also participated actively in the seminars and conferences as well as identified statistical and quantitative data necessary for measuring the phenomenon[8].

In the territorial meetings of stakeholders they explained how they collect information and data on those citizens in need that turn to their associations. There are various methods for gathering and recording the information with differentiated configurations. Agreement was reached to use an instrument that would make

[7] Many of these actors are present, as already indicated, in the co-programming round tables (government agencies – third sector) for social inclusion policies provided for by national regulations and those of the Tuscany Region.

[8] The chapters drafted directly by the Health's societies employees in the already-mentioned "Preliminary Report on Poverty, Vulnerability and Living Conditions of Families in the Province of Pisa" provide a detailed account of the efforts made as well as the policies used to prevent and fight the phenomenon of poverty that were initiated by Social Services in their reference territories.

it possible to regularly collect[9] their data and information. At the time this report was drafted, the chart proposed was developed and shared by two territories by the Health's societies.

The chart is not yet in its definitive version, because it might be subject to additional interventions. At present it includes the following data:

| | **Structure of Form for Stakeholders (information to be collected every six months)** |
|---|---|
| | **Information on the organization** |
| 1 | Name, Type of Organization, Person in Charge |
| 2 | Year established |
| 3 | Address (street – town – administrative division; telephone number and e-mail; web site if it exists |
| 4 | Sectors (in order of prevalence) and services provided according to each sector listed |
| 5 | Services provided with the greatest frequency in the last two years |
| | **Information on Users** |
| 6 | Estimate (or exact figure) of the number of users attended to in the period identified |
| 7 | Subdivision into age – gender – citizenship – country of origin categories |
| 8 | Civil status and presence/ non presence of children |
| 9 | Part of the territory from which they come (Municipality/ administrative division) |
| 10 | Subjects requesting services: increase or decrease during the time under consideration |
| 10 | If subjects' needs are taken care of: total number of persons under care for several years and how many new subjects |
| 11 | Needs expressed (indicate order or prevalence) |
| 12 | Needs expressed according to nationality of citizen |
| | **Networks** |
| 13 | Names of institutional agencies that the organization works with (specify territorial level) |
| 14 | Names of third-sector agencies that the organization works with (specify territorial level) |
| 15 | names of institutional agencies and third sector organizations that operate in the prevention and fighting of poverty with whom you DO NOT collaborate (specify the territorial level) |
| | **Evaluation of Interventions** |
| 16 | Successful interventions and the factors that were decisive in success |
| 17 | Unfavourable conditions of the territory that do not make effective interventions possible |
| 18 | Type of individual support required by subjects in need |
| | **Observations by person filling out the form** (blank space) |

**Objective 5:** Promote **self-evaluations by Social Services on their own** policies and interventions to prevent and fight poverty and social exclusion.

Something that came to light in the work groups is the need for persons working in the Health's societies to be able to involve Social Services from the territory when it conducts self-evaluations and deliberations of the policies established. The daily work of the agencies, and the urgency of social exclusion lead to a lack of these opportunities, which instead are crucial in highlighting the critical nature and limits of their actions, develop new proposals and identify good practices adopted in individual territories by the Health's societies.

[9] An initial suggestion that was valued by the work groups was to collect and analyse the forms every six months.

Towards garnering support for this process, during the course of 2006, with the scientific support of experts in Business Economics proposed by the Department of Statistics, Economics and Applied Mathematics at the University of Pisa, a study was initiated on the logistics of planning and the dynamics of spending for interventions on social inclusion. It was from this viewpoint that the first guidelines were developed making it possible to allow the contact persons from the Healthcare Organisations, together with the territorial working groups, to draft an initial report on the status of their territorial interventions.

The first results were published in the above-mentioned Preliminary Report on Poverty, Vulnerability and Living Conditions of Families in the Province of Pisa.

| | **Framework for analysing the current status of social inclusion interventions carried out in the reference territories by the Health's societies** |
|---|---|
| | **Definition of concepts in the different territorial spheres** |
| 1 | Poverty |
| 2 | Social exclusion |
| 3 | Social inclusion |
| 4 | Types of hardship in the territories |
| | - Economic condition of family units |
| | - Precariousness of the job market |
| | - Disintegration of family units |
| | - Unravelling of social and relationship networks |
| | - Ageing of the population |
| | - Mental or physical disease |
| | - Hardship related to forms of dependency |
| | - Cultural privation |
| | - Infantile privation |
| | - Gender poverty |
| | **Local welfare policies set up to combat the phenomena** |
| 4 | Economic interventions (continuous and one time economic grants, aid agreements) |
| 5 | Interventions to promote social inclusion (systemic social policy interventions and intervention projects) |
| 6 | Social spending: balance sheet |
| | **Network procedures and institutional relationships with the third-party subjects in the territory** |
| 7 | Involvement modalities of third sector associations |
| | - Establishment of territorial tables for social inclusion |
| | - Modalities to carry out the project interventions with the third-sector subjects |
| | **Evaluation of the interventions** |
| 8 | Strong points of the territory |
| 9 | Critical points of the territory |
| 10 | Successful interventions |
| 11 | Type of individual support needed by subjects in state of need |

**Objective 6**: Promote awareness of the phenomenon among policy makers and stakeholders in the territory.

In addition to launching the efforts described herein and the involvement of territorial players, UROPS from time to time presented the results of its research in the territories, thereby stimulating knowledge of the phenomenon and increasing the awareness of new forms of poverty. During the course of the territorial presentations, it also gathered additional information on how the phenomena of social exclusion presented itself in the territories and made additional contacts to increase the non-institutional subjects that could be encouraged to participate (anti-usury centres and other associations active in the territories).

The process of territorial involvement will be pursued as part of the SAMPLE project.

**Objective 7:** Link up with European Institutional networks as well as third sector networks operating in the thematic field of poverty and social inclusion.

In its efforts to reach this objective, UROPS so far conducted meetings with certain European networks involved with the issues of poverty and social inclusion. These meetings have not yet been formalized. The objective was to develop an analogous language and approach to the phenomenon, compare the indicators used in measuring them, and learn about European *good practices* at institutions and third sector associations. Among others, one of the good practices identified was that of *empowerment* interventions encouraging the direct involvement of persons living in poverty.

The contacts that have not yet been formalized in the previous phases were made by UROPS and the above-cited EAPN, and specifically, with contacts from the Italian chapter of CILAP and with the European network RETIS.

Developing this project will make it possible to expand the European relationship networks and the planned scientific partnership, creating a fertile field of opportunities to work in collaboration with a wider number of actors in the European territory, both institutional and non-institutional that are actively fighting poverty.

# Conclusion

## Expanding the efforts of local territorial governments in their analysis and interventions in fighting and preventing poverty

The status of the efforts described in this document reveals a progressive conceptual convergence on the multi-dimensional interpretation of poverty and social exclusion by different territorial actors, both institutional and non-institutional, and varied actions and instruments used to measure the phenomena.

What does emerge from the examination of the territories is that there is a weak application of the approaches, methodologies and policies adopted and recommended by the European Commission to prevent and fight poverty.

Systematic actions were initiated that were recounted in this report, both at the regional and provincial level, to share methodologies and instruments in order to reach a convergence of analysis and a more efficient comparison of the synergetic policies to strengthen the welfare systems. The common cognitive bases constructed with the participatory approach used actually facilitates and assists the rationalization and integration efforts of the policies thereby benefiting impoverished citizens.

Specifically as regards the territory of Pisa, UROPS, acting as the promoter and organiser of the territorial sharing efforts, in a bottom-up proposal of the European Open Coordination Method, hopes to conclude and strengthen the actions undertaken thanks to SAMPLE.