ESSnet on Small Area Estimation

REPORT ON WORKPACKAGE 6

GUIDELINES

FINAL VERSION

March 2012

Contributors:

Italy, Netherlands, Norway, Poland, Spain, United Kingdom.
Guidelines for the application of the small area estimation methods in NSI sample surveys

ISTAT, CBS, GUS, INE, ONS, SSB

Contents

1 Introduction 3

2 Scope and purpose of the guidelines 4

3 Indirect estimation methods 5
  3.1 Indirect estimation from design-based perspective 7
  3.2 Role of models in small area estimation 8

4 Standardized three-stage process 9
  4.1 Clarification: Needs, data, evaluation criteria 10
  4.2 Basic smoothing: Triplet of direct, synthetic and composite estimates 16
  4.3 Enhancement: Modelling to address outstanding issues 18

5 Sources and selection of additional information 21
  5.1 Selecting auxiliary data 23
    5.1.1 Area types and auxiliary data 25
    5.1.2 Boundary systems 26
    5.1.3 Auxiliary data at boundary systems 27
    5.1.4 Auxiliary data and covariates 27

6 Main SAE estimators 28
  6.1 Direct estimation 28
    6.1.1 Some particular direct estimators 30
  6.2 Indirect estimators 31
    6.2.1 Synthetic estimators 32
    6.2.2 Composite estimates 34
  6.3 Model based estimators 35
    6.3.1 SAE models and approaches to inference 40
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3.2 Basic area level mixed model</td>
<td>43</td>
</tr>
<tr>
<td>6.3.3 Basic unit level mixed model</td>
<td>45</td>
</tr>
<tr>
<td>7 Diagnostics and accuracy assessment</td>
<td>48</td>
</tr>
<tr>
<td>7.1 Design-bias exploration</td>
<td>48</td>
</tr>
<tr>
<td>7.2 Over-shrinkage and other ensemble properties</td>
<td>49</td>
</tr>
<tr>
<td>7.3 Model diagnostics</td>
<td>51</td>
</tr>
<tr>
<td>7.4 Benchmarking</td>
<td>56</td>
</tr>
<tr>
<td>7.5 Maps</td>
<td>57</td>
</tr>
<tr>
<td>7.6 MSE estimation</td>
<td>58</td>
</tr>
<tr>
<td>7.7 Monte Carlo simulation</td>
<td>60</td>
</tr>
<tr>
<td>8 Software tools</td>
<td>63</td>
</tr>
<tr>
<td>8.1 Design based estimation</td>
<td>63</td>
</tr>
<tr>
<td>8.2 Model based estimation</td>
<td>64</td>
</tr>
<tr>
<td>8.2.1 Ad hoc software tools</td>
<td>64</td>
</tr>
<tr>
<td>8.2.2 Time consuming efficiency</td>
<td>65</td>
</tr>
<tr>
<td>8.2.3 Model selection and model evaluation</td>
<td>66</td>
</tr>
<tr>
<td>9 Overview of the case studies</td>
<td>67</td>
</tr>
<tr>
<td>10 Conclusions</td>
<td>70</td>
</tr>
</tbody>
</table>
1 Introduction

In recent years, due to the increasing demand for more detailed information, methods for small area estimation have been intensively investigated by researchers of academic institutions and NSIs. As consequence, after their evaluation through empirical studies, several small area methods have been proposed and specific software tools have been produced for their application. Instead, it still partly missing a comprehensive list of activities needed to apply the methods and a set of principles that, assuring the quality of the estimates, can boost up the application of the SAE methods for the official statistics production. In the challenge to fill this gap, it emerged the need of an ESSnet project on small area estimation. Sample surveys are used to provide estimates for target variables at population or national level and also for subpopulations by geographic area or other domains of study. The aim of small area estimation methods is to produce reliable estimates for the variable of interest for unplanned domains, under budget and time constraints. In principle, if a sample is selected in each domain or area, design consistent and unbiased estimates can be produced. However, the sample surveys do not always allow to select sample units for every small domain of interest, in these cases small area estimation procedures need to be utilised. Then, a small area estimation problem occur whenever only a small sample sizes or even no sample has been collected in the area or domain of interest.

A domain is a subgroup of a population defined by classificatory levels, such as, age, sex, ethnicity, socio-economic status. An area is a geographic area such as cities, states, local authorities, municipalities. In sample surveys an area or domain is regarded as small if the sample within the area/domain is not large enough to provide direct survey estimates of adequate precision. So that the term small area here refers to any subpopulation (or domain) regardless whether it is geographic or not.

The aim of this document is to provide a set of principles for the application of small area methods to obtain the estimates of interest in the NSIs large scale surveys, in order to make SAE techniques actually applicable. This report summarizes the outcomes of the ESSnet project on small area estimation, taking advantage of the experiences, knowledge and efforts of all its members. The objective is not to produce a manual or a complete review of methods, but a document that allows to point out some good practices to follow in order to obtain step by step small area estimates of interest with a sufficient level of quality.

These guidelines are restricted to the application of estimation methods on data from national surveys, we remand at the paper of Brackstone (2002) for a description of different situation for which detailed information for sub-domains may be needed. Moreover, the problems of addressing small area information needs at sampling design level and the issues of balancing between national and small domains information requirements is not developed in this document See Singh

This document is organized as follow: firstly there is a description of the overall process illustrating the steps that should be applied when small area estimates are needed. In this phase a clarification of user needs, the evaluation of variance of direct estimates, an initial application of some basic methods and the prerequisite for applying these methods are specified. Secondly the relevant issues of the use of explicit models are taken into account, emphasizing the advantages and the drawbacks of model based small area estimation methods.

Moreover, since a crucial point for the production of small area estimates is the quality assessment of the implemented methods a set of diagnostics and checks of the properties of the applied estimators are proposed without letting aside a description of the main sources of auxiliary information and how these extra information can be used to improve the quality of the small area estimation. Finally, a brief review of the standard and enhanced small area methods, a summary of the case studies carried out by the ESSnet members and some indication about useful tools or procedures that can be actually used to produce small area estimates will be provided. With reference to software tools, some codes developed by ESSnet members will be also available in the portal of this project.

2 Scope and purpose of the guidelines

A problem of small area estimation arises when “direct estimates” based on the domain-specific data are too unreliable to be accepted in general. All the small areas of interest form a partition of the target population and the existence of such a partition should always be kept in mind on dealing with small area estimation (SAE) problems. Thus, one can have to cope with this problem despite the direct estimator may be acceptable in a few, or even quite a few, small areas. The most important reason that a sample-based direct estimate may be unacceptable is the lack (or absence) of direct domain-specific data.

In contrast, any estimator at the national level is always direct by this definition. Small area estimation therefore presents an unfamiliar situation to many practitioners of survey sampling, and small area estimates are typically not produced on a routine basis from the numerous sample surveys conducted by the national statistical institutes (NSIs), despite considerable methodological research and software development in the past three decades or so. Clearly, there is a gap between the current practice of survey sampling and the growing demand for more detailed statistics in many subject areas.

The current guidelines aims to address this perceived gap. Above all, we shall outline a standardized approach, or process, to small area estimation based on existing sample survey data. The accompanying key concepts, methods, and
software tools are defined and classified, with references to the literature as well as reports from the other work packages of this ESSnet project where these are covered in greater details. By following the standardized process, one should be able:

1. to reach a decision of whether acceptable small area estimates of interest may be produced based on the available data, method and software tool;

2. otherwise, to identify the major obstacles that one needs to overcome in order to reach acceptable results.

It has been our intention to make the guidelines concise instead of comprehensive, prescriptive rather than discursive. Neither we shall concern ourselves with other issues of small area statistics beyond the most typical survey sampling situation. It suffices to mention two examples here as a reminder of the need to keep a broad outlook.

1. A wholesome approach, of which the statistical estimation methods constitute one of the many components, is required in order to make small area statistics, herein small area estimation, an integrated part of the statistical production system. Brackstone (2002, “Strategies and approaches for small area statistics”, Survey Methodology, vol. 28, pp. 117-123) discuss some of the issues and challenges from a more strategic point of view.

2. Sample surveys may be initiated primarily to address the need of small area statistics for policy/program funding and evaluation, without the traditional priority on the national-level estimates. There are, then, some quite different issues for the sampling design which can take into account small area needs. (Singh et al 1994, David A. Marker 2001, Longford 2006, Falorsi et al 2006) Anyway sampling design can not be planned to consider all the small area needs especially when the area population sizes are too small, due to budget constrains.

3 Indirect estimation methods

The standard problem for large scale sampling surveys is related to the estimation of a wide range of parameters of a finite population $P_R$ referred to a given area $R$. The sampling strategy, generally based on complex sampling designs, uses estimators finalized to produce reliable estimates for population $P_R$. The need to produce estimates for sub-areas of $R$ often can not be dealt with in the planning stage. This situation is showed in Figure 3.1, in which one stage stratified sampling is schematized and where $h$ is one of the $H$ strata of a partition of $R$ and
where points represent sampling units. As Figure 3.1 shows, potentially three different types of small areas can occur:

1. the first type, denoted by $d$, is an example of unplanned small area, being the union of $H_d$ strata, some of which are not complete and the corresponding sample size $n_d$ is a random variable;

2. the second kind of small area, denoted by $d'$, is a special case of small area in which no sample units belong to the target small area;

3. the third type of area, denoted by $d''$, is an example of planned small area, being the union of complete sampling strata.

A main classification of small area estimators is based on the distinction between direct and indirect estimators. For each area $d$ ($d = 1, ..., D$) and time (or survey occasion) $t$ ($t = 1, 2, ...$) of interest:

1. direct estimators, uses the values of the variable of interest observed on sampling units belonging to small area $d$ and time $t$;

2. indirect estimators may offer greater precision than the direct estimators by exploiting observed values of the target variables for a larger area, a broad-area $H$, containing the small area $d$, and/or sample values which are collected for different time occasions besides the current one.
The indirect estimators may be classified in synthetic and composite estimators and may be further classified according to the underlying inferential approach, i.e., design-based, model-assisted or model-based (predictive or Bayesian). When the indirect estimators are derived under a model-based approach, you can have a model-based synthetic estimator, if a fixed model is assumed or a model-based composite estimator, if a mixed model is taken into account.

In the first two approaches, the target parameters are considered as unknown but fixed quantities, whereas in the third one a model is formalized in order to obtain the estimator. The properties of design-based and model-assisted approaches are derived on the design randomization as with direct estimator. In the third approach, the parameters of interest are considered random variables; inference is based on statistical models which describe the relationship between the variables of interest and auxiliary information.

The models are used to link the available sample information of the small areas (within a broad-area) and/or the data coming from different survey occasions. In both cases, extra auxiliary information correlated with the target variable, available from census or from administrative archives, allow improvement of the small area estimates produced. The estimates produced with these methods are biased but more efficient than direct ones.

### 3.1 Indirect estimation from design-based perspective

For practitioners who are unfamiliar with small area estimation, the transition from direct estimates of national-level, or major sub-national levels, parameters to indirect estimates at detailed aggregations levels is an important one, because it has much to say for one’s understanding, evaluation as well as communication of small area estimates.

In the practice of official statistics, survey-based direct estimates rely generally on the so-called sample weights. Design-based consistency or, even more ideally, unbiasedness is an essential concern for the estimator and the associated inferences (such as standard errors and confidence intervals). Indirect small area estimates, however, imply that approximate design-unbiasedness is no longer attainable, except for areas with sufficiently large sample sizes. But the design-based properties of the indirect small area estimates can be appreciated in terms of the reduced variances against the increased biases, e.g., through the mean squared error (MSE). We shall refer to this as the design-based smoothing perspective of indirect estimation. Notice that an indirect estimator can be consistent in the traditional asymptotic setting where both the within-domain sample size and the population size tend to infinity. But this would merely imply that the small area estimation problem vanishes on its own asymptotically. Rather, the appropriate asymptotic setting for small area estimation is to let the number of small areas tend to infinity, while keeping the within-domain sample sizes and population...
sizes finite. Clearly, then, an indirect estimator is inconsistent asymptotically.

Indirect small area estimation therefore differs from the situations, such as treatment of non response, seasonal adjustments, estimation based on cut-off sampling, etc., where the design-based inferential framework is either insufficient by itself or undefined. The smoothing perspective allows one to maintain a design-based approach to small area estimation and inference, just as in the case of traditional direct estimation at the national level. And, as we shall explain in Section 2, it is generally possible to obtain indirect small area estimates that are ‘smoother’ than the direct area-specific estimates, using the same data that are available for the national-level estimates.

One must however acquire a ‘new’ understanding of the properties of the indirect estimates. Variance is by nature purely random. But bias can have certain recognizable and expected patterns in small area estimation. It may therefore be worthwhile sometimes to ‘trade’ bias against variance for the given purpose. For instance, a small area with a parameter value that is very outlying compared to the mass of these values tends to have its indirect estimate ‘pulled’ more towards the centre. This implies e.g. that, on average across the small areas, the estimates that are higher than the population average value tend to have negative bias, whereas those that are lower tend to have positive bias. Provided the parameter values change ‘smoothly’ from one time point to the next, the indirect small area change estimates may benefit from the bias-variance trade-off compared to the direct change estimates. Similarly for cross-sectional between-area comparisons such as the ranks derived from the indirect small area estimates instead of the direct estimates.

### 3.2 Role of models in small area estimation

From the model-based point of view of inference, small area estimation presents no new challenges conceptually. But one needs more sophisticated models that are able to account for the heterogeneity across the small areas. Moreover, given the validity of the design-based inferential framework as explained above, one must be prepared to subject the model-based small-area estimates to design-based evaluations.

For instance, the linear regression models are often used for direct model-based estimation at the national level (e.g. Valliant et al, 2001). In theory, at least, it is possible to apply the same model separately inside each area of interest to thereby derive the corresponding direct model-based estimate. However, in the context of small area estimation, the lack (or absence) of within-area sample data implies that these direct estimates are unacceptable in general, just like the design-based direct estimates. The linear mixed models with random effects in addition to fixed effects may be developed to address this problem (Rao, 2003), in which case the model-based estimates become indirect because the model parameters
are estimated from the data across the small areas. Thus, it is not the use of model per se that makes the estimates acceptable, but rather indirect estimation by means of models that are more sophisticated than what suffices for direct estimation.

Since the optimality of the model-based estimates holds only under the assumed model, one must be careful when it comes to the comparison between estimates derived under alternative models. For instance, the small area estimates derived under a linear regression model will appear unrealistically precise when these are evaluated under the same model: they can have very small variances because there are only a few global parameters that need to be estimated, and they have no bias under the assumed model. But the linear regression model may not be suitable at all because it takes no account of the heterogeneity in the data, so that a more realistic evaluation of the linear-regression estimates may be obtained under the alternative linear mixed model.

More generally, however, the alternative models may not be nested with each other, in a way which makes it possible to identify uniquely the most comprehensive one among them that encompasses all the others. For instance, one may have a linear mixed model, denoted by A, and an alternative quantile regression model, denoted by B. The estimators derived under model A may be better than those derived under model B, when the evaluation is carried out under model A; whereas the model-B estimators may outperform the model-A estimators, when both are evaluated under model B. In such cases, model-based comparisons may appear somewhat inconclusive.

As long as the design-based framework is valid, it is possible to evaluate the design-based bias and variance of any model-based estimator, usually by means of Monte Carlo simulations. This provides therefore a common ground for comparing alternative model-based estimators. Moreover, it is valid to compare these biases and variances with those of the indirect estimates derived under the design-based perspective. In this way, models are not ends in themselves, but become tools for ‘smoothing’. And it may outperform purely design-based indirect small area estimates in terms of the trade-off between design-based bias and variance, without one necessarily having to adopt the model-based inferential framework. Such an evaluation can help to convince the adherents of the design-based perspective to accept model-based small area estimates.

## 4 Standardized three-stage process

In this Section we outline a standardized three-stage process in the situation where (i) there is in existence a sample survey from which statistics at the national level, or major sub-national levels, are regularly produced by means of design-based direct estimation, and (ii) one would like to investigate whether it is
acceptable to produce statistics at some lower aggregation levels, possibly using small area estimation techniques. In the figure 4.1 the standardized three-stage process is displayed.

The flowchart showing the steps of the process involving the needs and priorities, the choice of the small areas and their hierarchy structure, the analysis of available additional information, the basic estimators (direct, synthetic and composite) and the quality evaluation of the produced estimates. This standardized process is explained in more detail in the following sections:

(I) the first stage of clarification for the identification and prioritization of the needs and uses of small area estimates, the survey of the relevant data that are available, and the choice of evaluation criteria of the obtained small area estimates;

(II) the second stage of basic smoothing where the triplet of direct, synthetic and composite estimates are calculated from a design-based perspective. No change of the inferential framework or additional data is needed compared to the existing regular survey. The triplet can always be produced;

(III) the third stage of enhancement that is needed if the results of basic design-based smoothing are not acceptable. Quality assessment of the triplet of design-based small area estimates should identify the weaknesses for improvement. The general means is to utilize additional information though modelling, targeted at the perceived short-comings in the basic smoothing results.

4.1 Clarification: Needs, data, evaluation criteria

The following user needs have often been mentioned in the literature:

- policy and program formulation and evaluation
- allocation of funds
- local government and business planning

Given the existence of a regular survey, the primary motivations for the NSI to investigate the potentials of small area estimates may include:

- Inter-census updating of census statistics, perhaps at an aggregation level higher than that in the census, but certainly much lower than otherwise common in survey sampling.
Figure 4.1 Overall flow chart of the three stages
- Monitoring of key socio-economic indicators, such as the Employment and Unemployment rates based on the Labour Force Survey (LFS), household income, crime experience, etc. at lower aggregation levels.

- Extending the traditional scope of official statistics, such as poverty mapping, welfare conditions, various environmental indicators, etc..

From a methodological point of view, however, the following distinctions regarding the nature of the small area statistics are important:

- Cross-sectional totals (or means) vs. their changes over time. As mentioned above, the bias of indirect estimation may have different consequences here. Moreover, timely correlations in the population and sample data have a bearing in the case of changes, but maybe less so in the case of totals.

- Area-specific best prediction vs. ensemble small area characteristics. When the interest is area-specific, one is concerned with the accuracy the estimate for each small area on its own. Whereas, if the interest is some ensemble characteristics, the small area estimates must be considered in relation to each other. Examples of ensemble characteristics include the difference between the maximum and minimum small area parameter values, the distribution of parameter values across the small areas, the rank of the small areas according to their respective parameter values, etc..

It may be noticed that the practitioners who are only familiar with the tradition of producing national-level estimates often find it difficult to prioritize among the different objectives of the small area estimates and hence, the preference and balance among their statistical properties. It is nevertheless important to be aware that the bias of the indirect small area estimates, as explained above, implies that

- the area-specific best prediction approach to the derivation of small area estimates, which is the most typical in practice, may not be suitable if the small area ensemble characteristics are of primary interest,

- special adjustments of the estimation method are usually needed in order to strike a better balance between cross-sectional totals and changes over time, or that between area-specific statistics and small area ensemble characteristics.

In the present context, the data and metadata for small area estimation include

- the statistical variables of interest from the existing sample survey;
• the auxiliary variables used in the existing estimation method for the national, or major sub-national, statistics — it is important to make clear whether these auxiliary variables are available at the small area level, such that the existing estimation method can be applied, either exactly or in close resemblance, inside each small area in order to produce the corresponding direct estimates;

• additional covariates (from previous census and surveys, administrative sources, etc.) that may help to explain the statistical variations in the target variables;

• frame data, either exact for the target population or in proximity, that can be used to set up realistic Monte Carlo simulation studies;

• the most important metadata to be clarified is the hierarchy among aggregation levels: from the national level to the small areas of interest, and possibly domains within the small areas.

It is good practice to make the repository of data as thoroughly as possible. Moreover, there may be an interaction towards the clarification/prioritization of the needs/uses of the small area estimates. It also helps to give an idea of the accuracy that may be expected of the small area estimates in relation to that of the existing statistics.

It is generally unrealistic to expect the small area estimates to have the level of accuracy that one is accustomed to in the case of national-level estimates. Examples of precision sizes/thresholds used in different contexts by different institutions are provided below. They are not meant to be prescriptive but rather to give some feasible benchmarks when defining precision thresholds:

• A coefficient of variation of 5% or lower means a satisfying level of reliability for estimates, while a coefficient of variation higher than 5% means lower reliability (Ardilly, P., 2006);

• In the ICT household survey, the estimated standard error shall not exceed 2 percentage points for the overall proportions and shall not exceed 5 percentage points for the proportions relating to the different subgroups of the population, where these subgroups constitute at least 10% of the total population in the scope of the survey (Eurostat, 2010).

• In EU-SILC, a methodological document presents the use of the "compromise power allocation" method to allocate the EU sample size (which should not exceed 80,000-100,000 sample households) to countries. The main household income measure is the poverty rate and varies roughly in the range 5-25%. At national level, taking a proportion (percentage) of 15%
as the basis for computations, a simple random sampling of 5,000 households is required (except perhaps for the smallest countries) to estimate this with 1 percentage point error (the absolute margin of error) (95% confidence interval). This corresponds to a size of around 0.5 percentage points of the absolute standard error.  

• The European Health Interview Survey (EHIS) methodological sampling guidelines (Eurostat, 2009) present the allocation of the EU sample size (270,000 individuals) to countries, by using the “compromise power allocation” method. This sample size was derived from the consideration that an average of 10,000 or 7,500 individuals per country would allow to meet a good precision (for a sample size of 7,500 and a percentage of 10%, the absolute standard error equals 0.3 percentage points. The national effective sample sizes have been computed by retaining as the most critical indicator the percentage of people with activity severely hampered. The indicator was selected because of low prevalence in some Member States which could lead to precision problems for some subgroups. For the national effective samples, corresponding errors in absolute percentage points (standard error in absolute terms) were then computed. At national level, the absolute standard error varies from 0.1 to 0.4 percentage points. This corresponds to an absolute margin of error of between 0.2 and 0.8 percentage points.

• in ISTAT, according to a rule of thumb, coefficients of variation should not exceed 15% for domains and 18% for small domains; when they do, this serves also as indication to use small area estimators. It is however underlined this is just a rule of thumb and that all domains are not equivalent, because the domains are associated with the percentage of the population they represent and this population can vary.

• Statistics Canada applies the following guidelines on LFS data reliability (Statistics Canada, 2010, pp. 30-31):

  – if the coefficient of variation (CV) ≤ 16.5%, then there are no release restrictions;
  – if 16.5% < CV ≤ 33.3%, then the data should be accompanied by warning (release with caveats);
  – if CV > 33.3%, then the data are not recommended for release.

• In the context of the Labour Market, some criteria for small area estimates have been pointed out (see chapter II, section D of WP3 final Report):

the ONS dissemination policy established (by 2004) a CV<20% for a small area estimate to be considered as publishable;

– in US, for monthly employment and unemployment estimates at state level, the sample is too small to produce reliable direct estimates (CV varies from 8% to 16%). At national level, CV=2% for unemployment is acceptable;

As explained earlier, the bias of the indirect small area estimates implies that it is insufficient to consider CV on its own. Area-specific MSE of the small area estimator is the most common measure of uncertainty in practice. Less often is the use of confidence intervals, or prediction intervals from the model-based point of view. It is also rare to find uncertainty measures of the ensemble characteristics of the small area estimates.

Apart from the choice of the summary measure of uncertainty, it is important to consider and select a set of diagnostics and checks that may help

1. to better understand the data that one is dealing with;

2. to assess the estimation methods and/or the underlying model assumptions;

3. to form a more wholesome picture of the quality (strengths as well as weaknesses) of the obtained small area estimates.

Finally, one should consider the feasibility of a realistic Monte Carlo simulation study based on a constructed “target population”, i.e. from the design-based point of view. It is not necessary that the constructed population faithfully reflect every aspect of the reality. But it should be realistic with respect to the key uses / needs of the small area estimates, as well as their corresponding desirable statistical properties, as explained above. Such Monte Carlo simulation studies, when feasible, provide often the most assuring evidences for the choice among the alternative methods and/or models.

To summarize, the various diagnostics and uncertainty measures are helpful in any case for comparisons between alternative estimators, in order to identify the best available method. But it is difficult to give a set of explicit and absolute general conditions that the estimation method and its results must meet in order to be considered acceptable for dissemination, just as in the case of official statistics at the national level. The decision is more likely to be made on a case-to-case basis with regard to a fit-for-purpose assessment, and the ‘acceptance’ margin is unlikely to be uniform across all the small areas of interest.
4.2 Basic smoothing: Triplet of direct, synthetic and composite estimates

Direct estimates are derived separately inside each small area based only on the data from the given area. In theory, provided auxiliary information is used for the regular national-level estimates and is available at the small-area level of interest, it is possible to apply the existing estimation approach inside each small area. In reality, however, the lack of data can cause problems. Firstly, if there are no sample data at all from a particular small area, then obviously no direct estimate can be produced, with or without auxiliary information. But the estimation method may also fail because the direct sample data are too few (or sparse). For instance, the post-stratified estimator may break down due to empty within-area sample post-strata; similarly, the calibration estimator may be infeasible because of ‘empty’ within-area sample margins.

As explained earlier, the direct small area estimates are not expected to be acceptable in general. (Otherwise, one does not really have a small area estimation problem.) The next option is indirect “synthetic estimation” (Rao, 2003, Section 4.2). Essentially, this amounts to replace the direct estimates with regression estimates, where the regression coefficients are estimated based on the data in a larger area (or domain) to which the small area of concern belongs. Simple examples below illustrate the idea.

- The broad area ratio estimator of a small area total is the product of the within-area population size and an estimated mean of a ‘broader’ area to which the small area belongs. The estimator is feasible without any additional auxiliary information.

- The “simple” ratio-synthetic estimator of a small area total is the product of the within-area auxiliary total and an estimated ratio between the target variable and the auxiliary variable, from a ‘broader’ area to which the small area belongs.

- The count-synthetic estimator is based on area-specific post-stratification, where the post-stratum means are taken from a ‘broader’ area to which the small area belongs.

- The “combined” ratio-synthetic estimator is obtained if a “simple” ratio-synthetic estimator is applied inside each post-stratum. The auxiliary categorical information for post-stratification is combined with an additional auxiliary variable.

The indirect synthetic estimates can have much smaller variances compared to the direct estimates. Nevertheless, they rarely provide satisfactory solutions to
the small area estimation problem, because they tend to “smooth away” the between-area variation of interest. That is, they can have unacceptably large biases. Moreover, from a smoothing perspective and at least in theory, it is always possible to further improve the bias-variance trade-off by means of composite estimation.

A small area composite estimate is given as a weighted average of the corresponding direct estimate and a synthetic estimate of choice. The idea is to give more “weight” to the direct estimate if it is reliable, whereas less “weight” is assigned to it otherwise. Indeed, the “composite weights” can be derived to minimize the area-specific MSE. As a result, the composite estimator can be assumed to have a reduced variance but an increased bias compared to the direct estimator, whereas it has a reduced bias but increased variance compared to the synthetic estimator.

In practice, however, one needs to allow for extra uncertainty associated with the estimation of the optimal “weights”. A tricky issue here is that the variance of the direct estimator can not be estimated with great precision, if the variance estimator is to be unbiased (or merely consistent), because any unbiased estimation is unreliable in general due to the lack of direct data. Two alternatives may be considered.

1. For repeated regular surveys, one may consider fixing the composite weights over a certain time period and, thereby, eliminate the extra estimation uncertainty altogether. These fixed weights can be calculated based on the historical data. The key is to ensure that they do not introduce too much additional MSE to the composite estimator compared to the optimal weights.

2. A simple choice is to define the composite weights directly based on the within-area sample size. For instance, one might assume the variance of the direct estimator to be inversely proportional to the within-area sample size, and the variance of the synthetic estimator to be inversely proportional to the broad-area sample size. Next, one may specify a conjectured bias of the synthetic estimator in a smooth fashion, e.g. in percentage relation to the standard error of the synthetic estimator. In this way, the composite weights become a smooth function of the relevant sub-sample sample sizes, no matter how much the target values actually vary in the population.

In one way or another, each synthetic estimator can give rise to a composite estimator, given the choice of the direct estimates. Together, they form a triplet of small area estimates that can always be produced based on the existing data, for all the small areas where the direct estimates exist, without the need to change the design-based inferential framework. We shall refer to the triplet as the basic smoothing approach.
It may be noticed that, based on the same available data, there are usually several different ways of constructing the triplet. For instance, in count-synthetic estimation, one uses relevant auxiliary information for post-stratification. Using the same information, one can construct a composite estimator based on the direct within-area estimator and the corresponding count-synthetic estimator. Or, one can first construct a composite estimator for each area-specific post-stratum mean, and then combine these to form an area-specific estimator. So there is scope for exploration.

The composite estimates are sometimes known as shrinkage estimates, because by construction all the direct estimates are pulled towards the corresponding synthetic estimate of a broader area. A consequence of this is that, together, the composite estimates that are derived by minimizing the area-specific MSE generally display less between-area variation than they should. This is referred to as the over-shrinkage problem. Alternative estimates are available that achieve better ensemble properties, such as the constrained empirical Bayes or “triple-goal” approach, and the simultaneous estimation approach from the frequentist perspective.

Composite estimation is in general easier to implement compared to explicit model-based estimation, especially as the model grows more and more complicated. Moreover, when the composite weights depend only on the sub-sample sizes, it is possible to derive composite estimates for a large number of target variables at the same time. In contrast, a model applies only to one variable, or perhaps very few variables, at a time. It is usually impractical to build models for all the statistical variables that are collected in the sample, neither at the national level nor at the small-area level.

4.3 Enhancement: Modelling to address outstanding issues

The basic smoothing may not be satisfactory for one or several reasons, e.g.

- There may be too many empty sample domains, where the synthetic estimator causes unacceptable over-smoothing.
- Even the ‘best’ composite estimator may have too large bias and/or MSE in general.
- The optimal composite weights are numerically too unstable as explained above, while the alternative practical choices of composite estimator are not good enough.
- The small area parameters may have skewed non-normal distribution across the areas, such that the composite estimates may fail for areas at either of
the two ends of the distribution. An example is binary means close to zero or one.

Explicit model-based approach may help to address the perceived short-comings, e.g.

- Additional covariates and/or correlations can be included in a model. For instance, the last census population totals (or means) can be used as the explanatory variables for the current population parameters, whereas the information can not be naturally utilized in design-based estimation unless the census-values of the current sample units can be identified. The incorporation of spatial and/or temporal correlations is another example. Further reductions of both bias and variance may therefore be possible with the model-based estimates. But additional model assumptions are necessary.

- The main numerical instability of the composite estimator is associated with the direct (unbiased or consistent) estimation of the variance of the direct small area estimator. Models specified for area-level sample statistics face a similar issue, if the sampling design effect needs to be taken into account. Being completely model-based, the unit-level models can provide an alternative in such cases.

- There are large classes of generalized linear and non-linear models that can be useful for handling non-normally distributed small area parameters.

It is beyond the task of the current guideline to provide a comprehensive coverage of statistical modelling. Instead we present in Figure 4.2 a diagram highlighting the options that may be available, i.e. the decisions that need to be taken.

- In many ways the first, or most basic, choice is whether to specify the model at the unit or area level. The choice depends on the nature and the availability of the data at the two levels. But the sampling design can also play a part. There may be a strong design effect for reasons such as stratification, multistage sample selection and/or clustering. An area-level model needs to take this into account in order to describe appropriately the variation in the area-level sample statistics. Meanwhile, a strong design effect may nevertheless be considered as non-informative from a model-based perspective, given appropriate auxiliary information. For example, the design effects due to stratification by age and sex can be handled in a unit-level model by including age and sex as explanatory variables. Or, the household clustering effects for the LFS employment status may be ignored provided the model makes use of good administrative employment data at the individual level, even if the administrative data are not used in the sampling design at all.
Next is the choice between linear and generalized linear (or non-linear) models. In theory, generalized linear models are preferable for categorical data. In practice, however, linear models are computationally much easier, and often yield similar results. For example, the employment status may be modelled as binary data using a logistic regression model, but it is often viable to apply a linear model at unit or perhaps area level, just like implicit models underlying design-based survey weighting are usually linear, regardless the type of data.

- The conceptual relationship between (fixed effects) regression models and (random effects) mixed models is analogous to that between synthetic and composite estimation. In small area estimation, the mixed models is always more appropriate than the corresponding regression model, where the latter simply excludes all the random effects of the former, because it allows for heterogeneity across the small areas. From the smoothing perspective, however, the questions are of an empirical nature: (i) is the regression model good enough? (ii) does the extra computational effort of the mixed model pay off? In particular, going from a generalized linear model to a generalized linear mixed model can be much more complicated technically than going from a linear regression model to a linear mixed model.

- Finally, there is sometimes a question of multivariate modelling. Multivariate modelling is more efficient (or appropriate) when there are multiple...
target variables from each small area, and that these are either correlated with (or mutually restrictive of) each other. For instance, the average household incomes of different household types from the same small area may be positively correlated. Or, a set of counts may sum up to a known total in each small area, such as the number of persons in different household types or the number of persons with the three different labour market statuses. Again, from the smoothing perspective, the question is practically of an empirical nature, just like the choice between regression and mixed models.

In summary, it is not necessary to start the small area estimation process with modelling. Rather, explicit modelling could be considered after the basic smoothing results have been obtained and examined. Modelling should be targeted to address the perceived short-comings of the basic smoothing results in the given situation. This is because, from the smoothing perspective of small area estimation, the goal of modelling is not to build the most plausible theoretical construct to explain the data, but to find a more powerful and, hopefully, acceptable tool for prediction.

5 Sources and selection of additional information

Small area estimation methods which enhance the reliability of estimates carry this out by ‘borrowing strength’. Borrowing strength means that estimates are no longer dependent solely on survey data confined to the small area itself. Smaller areas are assumed to behave in the same manner as larger areas for which reliable estimates can be made. In such a situation, if sample survey data alone is used, then the application of small area estimation methods is, in effect, simply an area smoothing. The area specific survey results can be used to vary this to a small degree but such an effect is marginal. While such estimates can be shown to be collectively ‘good’ (e.g. lack of overall bias), they achieve nothing in demonstrating local variability. In the extreme it can result in estimating all areas to the same national mean value!

For instance, consider a country comprising one hundred small areas say Area1 to Area100. A sample survey has been conducted that shows, nationally, 15% of all households live in overcrowded conditions. This is a reliable national direct estimate but the sample size in each of the areas is too small for such a reliable estimate to be made for each individual area. The extreme, as noted above, would be to assume that all areas behave as nationally and give the estimate of 15% to all one hundred areas. Maybe there exist a few large regions also for which reliable direct estimates could be made. In this case each small area could be
given the same estimate as the region in which it falls. While, better than the previous eventuality, this still is of little use in indicating real local variation. To avoid such a situation arising, small area methods apply additional assumed conditions. These can be spatial relationships between neighbouring small areas, correlations between area direct estimates observed over time or relationships between the variable of interest and additional data variables. This third is the most common implementation and described here.

The introduction of additional data, often termed auxiliary data, enables the assumption of homogeneity of behaviour among areas to be retained but made conditional on indicators defined by the additional data.

Returning to the example described above, we now introduce some known census data. We know that, nationally 58% of all households own (or are buying) their houses, 35% rent and 7% live with parents. The fact that this is census data means that such percentages are also known for each of the hundred small areas. They are likely to vary considerably. Of course it may be that house tenure and whether a household is overcrowded are entirely unrelated. In this case good knowledge of the former cannot help in understanding the latter. Whether a relation exists and forming a view of its strength can be gained through looking at their correlation. Furthermore, a regression-like model can be fitted where being overcrowded is the dependent variable and housing tenure the independent. Such a model could be formed either by using households as unit records or by small area using survey proportions of overcrowded households and area proportions by tenure. As long as the fitting of such a model shows a degree of dependence, then it can be used in the generation of estimates for each small area. Homogeneity of behaviour is maintained, but it is now made conditional on the values of an explanatory variable.

The principle then is: (i) fit the model using survey and auxiliary data; (ii) examine the fitted model for significance (using techniques such as high R2 and coefficients significantly different from zero); (iii) take the auxiliary data values for each area and use the model to generate estimates for each area.

Taking this approach, gives a set of model-based instead of sample-based small area estimates. These depart from the traditional approach and unlike the sample-based estimates may not be unbiased for particular areas. This is a loss due to the use (in effect) of the whole of the survey data in estimating each area. However there can be an important gain - that of increased precision. In concept this is due to the strength of the modelled relationship. If the variability of the sample points around the regression slope is small, then using the model to generate estimates will similarly give these estimates with small standard errors. Of course this is only for a good fitting model. A model with little explanatory power could be used for generating estimates but would give these estimates with much larger standard errors thus negating any advantage over direct sample estimates!

The example given above uses a single explanatory variable with just a few cat-
egories and is the simplest use of auxiliary data. In fact it could be conducted more simply as it is likely that reliable estimates could be made for each tenure category. However the real potential benefit of using auxiliary data is when several such variables are available and in particular when they are continuous rather than categorical with just a few values.

There is one other advantage gained from the use of auxiliary data. This is the ability to estimate even for areas for which there is no sample data. As long as it can be accepted that non-sampled areas do not differ in behaviour from sampled areas, then the auxiliary data for such areas can be input to the model to give the required estimates.

So summing up, suitable auxiliary data need to be identified. For the model to produce reasonable estimates there should be a strong relationship between the auxiliary data and the population of interest. A strong relationship allows easy identification of potential auxiliary variables. It also aids in explaining the model to users, in particular, what is driving the small area estimates. Potential relationships can be analysed through scatter plots, correlations or simple models.

So important is it in fact that, until an investigation has been made for suitable auxiliary variables, it may not be possible to guarantee in advance that suitable small area estimates can be made.

5.1 Selecting auxiliary data

a) Variables for initial consideration

It is a good idea to consider a wide range of additional data sources. Administrative data and censuses are perhaps the best sources as their area based values will not suffer from sampling error (though consideration should be made of their quality as they could suffer from many kinds of non-sampling error). The use of area based additional data from other surveys is possible but would make the calculation of precision estimates more complex. Administrative and census data are rarely available at individual level for confidentiality reasons, however survey data from other variables in the same survey data record can often be used. For the purposes of estimation rather than model fitting though, it will be necessary that equivalent area values are available for any individual level auxiliary variable used.

Auxiliary data should not contain missing values as these can bias the model or cause model failure. Missing values should be accounted for before modelling, for example, using imputation. Area level variables with too many zero values should also be avoided as these may have a negative effect on model fit. Any shortcomings that may arise from the data should be understood.

In the process of consideration, it is important, where available, to include variables which are felt theoretically or by repute to affect the variable of interest.
This could involve a trawl for availability and consulting subject matter experts. Careful examination of the correlation between the target variable and auxiliary variables is required. It is also important that detailed knowledge is available of the nature of the source and concepts of the additional data such as the following:

1. Availability for all areas for which estimation required.
2. Population scope of the data;
3. Definitions of variables / concepts used;
4. Purpose of data collection;
5. Reference period – concurrent to survey period is ideal but often not available e.g. census data may be from several years earlier – this is usually acceptable;
6. Methodology of data collection;
7. Survey design used if data is from sample survey;
8. Quality of framework used for unit selection;
9. Extent of missing data and whether any imputation used;
10. Classifications used;
11. Editing or data validation process used.

Auxiliary variables that provide demographic information such as sex and age are always useful because many social variables have some relationship to such demographic data. Another reason for using demographic information is where the population size or demographic composition varies considerably across areas.

b) Reducing the set

An initial set of possible auxiliary data could be large and cover much common grounds. In common with many other modelling fields, the ideal of parsimony of explanatory variables is important. So we should strive to fit a model with as high an explanatory power as possible but with relatively few auxiliary variables. Reducing the initial set can be based on a mixture of data investigative and automated statistical procedures. Bivariate correlations should be calculated between pairs of auxiliary variables. Any very high correlations (greater than 0.9) should be noted and it should be ensured that both variables do not occur together in a model accepted for estimation. This is to avoid the instability associated
with near multi-collinearity of explanatory variables. Perhaps at this stage one of any such pair can be eliminated but care should be taken as although highly correlated, one may nevertheless perform better in a model with the variable of interest.

This may then leave some variables with a high correlation with the variable of interest, some with little or no correlation and the remainder with some effect but not apparently strong. It would be right at this stage to eliminate those with little or no correlation but not the middle group.

Attention could then be given to those variables with the highest correlations to the variable of interest. A variety of models can be attempted with individual variables. The model fits should be studied and the most promising noted. Sometimes it is possible that a transformation of the variable will give better explanatory power so these can also be attempted.

Following on from this, the more promising models can be augmented with additional variables and further improvements observed. Automated procedures such a stepwise methods can be used to reduce burden if software allows, but overdependence on these procedures should be avoided. Such a procedure of adding variables can be continued then as long as model improvements are observed. It should be noted that when modelling is utilised, some model specific errors are introduced. One is model specification error - has the right regression model been chosen for the available data? Key auxiliary variables may not be available or be of poor quality. The error structure or clustering effects may be incorrectly represented by the chosen model. For example, significant measurement error in the observed auxiliary variables or clustering within geographic areas may not be correctly accounted for by the model. Model diagnostics are investigatory procedures undertaken after a model fit which can guard against such inappropriateness. They are described in detail in section 7 but in essence are used to determine its fitness for use in estimation. A model which does not pass on relevant diagnostics should be discarded. Ultimately, success will be determined by the acceptance of a model with good explanatory power. Ultimately success will be determined by the acceptance of a model with good explanatory power including a reasonably small set of auxiliary variables.

5.1.1 Area types and auxiliary data

It is interesting to consider the type of areas for which small area estimation can be conducted. Here we are referring to geographical small areas rather than non-geographical subpopulation domains.
5.1.2 Boundary systems

Boundary systems can be of different natures. Each country has its administrative boundaries, political boundaries and postal system boundaries. These are de jure boundary systems created for operational and community representation purposes. There can also be other de facto boundary systems which relate to natural and ecological phenomena such as river deltas and mountain ranges or to man-made obstacles such as major railway lines and motorways. The effects of barriers such as these can often be felt over an amorphous area. The natural boundaries will particularly be of importance for weather systems e.g. rainfall, and pollution measures. Industries such as water services need to organise their activities by such areas and therefore draw their own boundaries to suit. Also such features can determine daily transport arrangements, consequently defined ‘travel to work areas’ can also form their own boundaries. These alternative boundary systems may have little consistency with the de jure boundary systems mentioned.

The administrative and political boundaries are usually organised in a hierarchical manner commonly with a small number of regions at the highest level and in descending order: sub-regional areas, major cities, towns or parts of cities, local neighbourhoods. Within EU countries, an establishment has been made of a classification system for the top three levels of a hierarchy, namely Nomenclature of Statistical Territorial Units (NUTS) with the levels from the highest downwards named NUTS1, NUTS2 and NUTS3. These cater for populations down to a NUTS3 minimum of 800,000. This lowest level of NUTS3 is likely to be of interest in a small area context since it is unlikely that sample surveys can give sufficient precision for direct estimates at this level. However calls for estimates are also likely at smaller sized areas. By extension, EU countries have defined two classifications below NUTS3, these are Local Administrative Units levels 1 and 2 (LAU1 and LAU2) and Urban Audit which is city based with classifications of core city, “Larger urban zone” and “Sub-city districts”. While the first system follows down the administrative boundary hierarchy, the urban audit system caters for the more de-facto influence of the cities. More importantly though, in some countries, notably the UK, LAUs can change periodically.

Calls for small area estimates are likely for both these lower level types of boundary classification and also for the alternative boundary systems based on natural and man-made barriers. Sometimes users may be interested in estimation at more than one boundary level.
5.1.3 Auxiliary data at boundary systems

This leads to availability of possible auxiliary data for such areas. Ideally, the data should be available for the areas at which estimation is required or for nesting smaller areas. However, this need not be absolutely necessary. If areas are part of a hierarchy of boundary systems then, as long as some well correlating data is available at the level required or below, there is value in including other good data available only at a higher level. This will be especially the case if estimates are required at higher levels of the hierarchy as well.

Census variables can be particularly attractive for use as auxiliary variables since they are often available at very small administrative area levels. The variables cover many aspects of population and household characteristics in fields of housing, health and industry and economic circumstance. Thus some are likely to have good correlation with survey variables of interest in these fields. They do suffer from a disadvantage that they are static between two censuses, usually ten years, but in practice this is not usually of major concern.

Data available from administrative systems is by its nature a by-product of their main function. It is rarely available at the lowest area level that census data is. Depending on its nature it can be available at one of the administrative levels or at an area particular to its field of operation. However, when the nature of the survey variable of interest is comparable, it can be very valuable. One example of such information could be counts of numbers registered as work-seekers or claimants of non-employment benefits when seeking to estimate unemployment from a Labour Force Survey. Another would be counts of social benefit claimants when seeking to estimate poverty rates from a social survey.

Moving away from hierarchical boundary systems to those which do not exhibit consistent nesting can cause difficulty. Sometimes a potentially good auxiliary variable is not available at the exact geography. However, maybe, such data is available at a very small level of another hierarchy. You can consider whether there is an approximate nesting of these very small areas within the required area. If so then the data could be used in the usual manner. This technique can be particularly useful with census data as its lowest publication level is usually very small. Alternatively, if a small level nesting cannot be found, data recasting techniques have been developed which can attempt to translate data on one boundary system to that of another. Data recasting techniques are described in Heady and Ralphs (2005).

5.1.4 Auxiliary data and covariates

The terms auxiliary data/variables and covariates are used throughout discussions of small area estimators. An explanation of context is given here. In section 4.1, a distinction has been made between auxiliary variables and additional covari-
ates. Auxiliary variables are termed to be variables which are used in existing estimation methods for national or major sub-national statistics. That is they are variables which can help to improve the precision of direct estimates. Additional covariates are considered as further variables which may help to explain the statistical variations in the target variables when considering small area estimation. In practice, though, it is difficult to preserve this distinction. In this section 5, the term 'auxiliary' has been used to encompass either such type of variable. The following section 6 defines and describes estimators and how they make use of any auxiliary data. The term covariate is then used to describe auxiliary variables which are fitted into the explicit models of section 6.3, i.e. for those accepted into models used for estimation. This is to emphasise how their variation in an explicit model directly affects the estimate. While the looseness of terminology is regrettable, 'covariate' either in the section 4.1 sense or as used in section 6 does demonstrate the need for further explanatory effects in small area estimation required above that needed in standard estimation methods.

6 Main SAE estimators

There are two main types of small area estimators: the direct estimators which use values of the variable of interest only from the time period of interest and only from units in the domain of interest and indirect estimators which borrow strength from different time periods or from other domains. Direct estimators may or may not also utilise auxiliary data. Their main problem is that when the sample size in the areas of interest is too small the estimates are too unstable. Indirect estimators which cover synthetic and composite estimators usually depend on covariates for which population values are known for the domain and the time period of interest. The relationship among target variables and covariates is formulated through a model (which may be implicit) and from whom the estimator formula can be obtained. In the following sections the main direct and indirect SAE estimators are briefly reviewed.

6.1 Direct estimation

Direct estimators are based only on survey data from within that area. Direct estimators can be obtained with or without auxiliary data. The basic survey estimator, the Horvitz-Thomson estimator, is the sum (or mean as appropriate) of the weighted responses within the area of interest. When no auxiliary data are available and the population size in a domain/area is unknown the direct survey estimator for a domain/area total is unbiased but has low precision in smaller areas. Precision can be improved by using estimators which include auxiliary data. Two of such types of direct estimators are described here Horvitz-Thompson
estimator and Generalised Regression Estimator. For the total mean value the Horvitz Thompson estimator (HT) is given by:

$$\hat{Y}_{d,HT} = \sum_{i \in d} w_i y_i$$

where the inclusion probabilities $\pi_i$ and the sampling design weights $w_i = 1/\pi_i$ are fixed and known quantities.

This estimator needs only the knowledge of the sampling weights and no auxiliary information is needed, furthermore it is unbiased under any sampling design and its variance is entirely determined by the choice of sampling design. The drawback is that it is unreliable if domain sample size is not large enough. Of course, no estimates are possible for no-sampled areas.

The Generalised Regression Estimator (GREG) improves the precision the HT estimator using population totals of auxiliary variables at domain level and covers a wide range of estimators. It is expressed by

$$\hat{Y}_{d,GREG} = \sum_{i \in d} w_i y_i + \left( X_d - \sum_i w_i x_i \right)^T \hat{\beta}_d$$

where $X_d = (X_{d,1}, \ldots, X_{d,k})^T$ is a vector of $k$ population totals of covariates, and $\hat{\beta}_d$ is weighted estimates of the vector regression coefficients

$$\hat{\beta}_d = \left( \sum_{i \in d} w_i x_i x_i^T \right)^{-1} \sum_i w_i x_i y_i$$

The GREG estimator is design consistent and approximately design unbiased and its variance is small when the model fits well for all domains. Moreover it ensures consistency with the known auxiliary totals $X_d = (X_{d,1}, \ldots, X_{d,k})^T$ in the sense $\sum_{i \in d} w_i x_i = X_d$.

GREG estimator covers different types of direct estimators applied in the NSIs surveys such as ratio estimator, raking ratio and poststratified-ratio estimator. Nevertheless, the variance of the GREG estimator can become large in domains where the sample size is small and when relationship among target variable and auxiliary information is weak it can produce bad or even unacceptable estimates.

A modified version of GREG-type estimator borrows strength from the overall data belonging to a broad area for the estimation of regression coefficients, without increasing the effective sample in small areas. In this way the estimate of regression coefficients is realible when just the whole sample size is large enough, the expression of the estimator is

$$\hat{Y}_{d,MOD} = \sum_i w_{id} y_{id} + \left( X_d - \sum_i w_{id} x_{id} \right)^T \hat{\beta}$$
Being $\hat{\beta}$ is the WLS estimate of regression coefficients, given by

$$\hat{\beta} = \left( \sum_{i,d} w_{id}x_{id}x_{id}^T \right)^{-1} \sum_{i,d} w_{id}x_{id}y_{id}$$

The modified direct estimator is approximately design unbiased as the overall sample size increases, even if domain sample sizes are small. Moreover, this estimator borrows strength for estimating the regression coefficients but it does not increase the effective sample size, unlike indirect estimators (Rao, 2003).

6.1.1 Some particular direct estimators

- **Domain area ratio estimator**

The estimator, $\hat{Y}_{d,\text{ratio}}$, for small area d is given by

$$\hat{Y}_{d,\text{ratio}} = \frac{N_d \hat{Y}_d}{\hat{N}_d}$$

where $\hat{Y}_d$ is the direct survey estimate (Horvitz-Thomson) of the target variable for the small area d, while $\hat{N}_d$ is the direct survey estimate (Horvitz-Thomson) of the population size of the small area d (summing survey weights). $N_d$ is the known population size of the small area d.

- **Domain area regression estimator**

An alternative estimator can be used where the value of an auxiliary variable is available for each small area from a separate source and is also measured in the survey.

The estimator, $\hat{Y}_{d,\text{REG}}$, for small area d is given by the following expression:

$$\hat{Y}_{d,\text{REG}} = X_d \frac{\hat{Y}_d}{\hat{X}_d}$$

where $\hat{Y}_d$ is the direct survey estimate of the target variable for the small area d, $X_d$ is the direct survey estimate of the auxiliary variable at the small area d, and $\hat{X}_d$ is the known value from another source of the auxiliary variable at the small area d.

For the case of these estimators using auxiliary data (or population sizes), the estimator is of a ratio form with the target variable sample estimate being adjusted by an appropriate ratio. In this sense, the assumption is that
if the survey underestimates the additional variable it will underestimate
the target variable by the same proportion. If auxiliary data is available at
the area level then the adjustment proportions can be area specific.

6.2 Indirect estimators

In the context of small area estimation, direct estimates are often not adequate
because their variances can be unacceptably large when the domain sample sizes
are small. In this case one should consider the application of indirect estimators,
for which the area estimate does not depend uniquely on the survey responses
within that area. On the contrary, the responses coming from a larger area,
namely the Broad Area, are used to obtain estimates. There are basically two
types of indirect estimators: the synthetic and the composite estimators which
can be derived under a design-based approach or assuming an explicit area or a
unit level model. In the latter case, according to the nature of the target variable,
the relationship among variable of interest and covariates can be formulated by
a linear or a generalized linear model for synthetic-type estimators and by means
of linear or generalized mixed models for composite type ones. Based on these
types of models the small area indirect estimates can be obtained by means of
Empirical Best Linear Unbiased Prediction (EBLUP), Empirical Bayes (EB) or
Hierarchical Bayes (HB) approaches to inference.

The term synthetic estimator is used when a reliable direct estimator for a BA,
including the small areas of interest, is used to obtain an estimate for the small
areas under the assumption that all the small areas included in the BA have the
same characteristics. As a special case of synthetic estimator, one can use the
area-specific expansion of the BA mean if no covariates are available (broad area
ratio estimator). A ratio-synthetic estimator is obtained when a single covariate
is available, a ratio-estimator from the BA is instead utilized. In particular, a
count-synthetic estimator is obtained when the area-specific post-stratification is
applied, where the post-stratum means are estimated across the areas. Whereas
one obtains a ‘combined’ ratio-synthetic estimator, if a ‘simple’ ratio-synthetic
estimator is applied inside each post-stratum, i.e. instead of expansion of the
post-stratum mean. Thus, these special cases of regression-synthetic estimation
can be characterised by the combined choices of post-stratification (or not) and
single covariate (or not). The synthetic estimator is essentially a biased estimator
but when the small area does not have a strong individual behaviour it will be
efficient with a small mean square error.

The composite estimator is used as the immediate reaction when the direct es-
timates are rejected because of their large variances and the synthetic estimates
are unacceptable because of their large (design-based) biases. It is an estima-
tor built for balancing the properties of the direct and synthetic estimator in
function of precision of direct estimates. Then, when the sample size is large
enough direct estimator is generally more valuable, otherwise, when the sample size is small more importance is given to the synthetic estimator. A composite estimate is essentially given as a weighted average of the area-specific direct estimate and a suitable synthetic estimate. As a result, composite estimate can be assumed to have a reduced variance but an increased bias compared to the direct estimate, whereas it has a reduced bias but increased variance compared to the synthetic estimate. However, the composite estimates may be acceptable based on a MSE-criterion. Two observations are worth noting. Firstly, auxiliary information which can be built into the hierarchy of areas (or domains) may always be explored by means of composite estimation, before an explicit model-based approach. For instance, in count-synthetic estimation one uses relevant auxiliary information for post-stratification. Using the same information, one can construct a composite estimator based on the direct within-area estimator and the corresponding count-synthetic estimator. Or, one can first construct a composite estimator for each area-specific post-stratum mean, and then combine them to form an area estimator. Secondly, composite estimates are sometimes known as the shrinkage estimates, because by construction all the direct estimates are pulled towards the corresponding synthetic estimate of a larger area (or domain). A consequence of it is that, as an ensemble, the composite estimates that are derived by minimizing the area-specific MSE generally display less between-area variation than they should. This is referred to as the over-shrinkage problem.

6.2.1 Synthetic estimators

The basis for indirect estimation lies with what is known as synthetic estimation. Synthetic estimation uses both survey and auxiliary data from outside as well as within the domain/area of interest. An estimator is called synthetic estimator if a reliable direct estimator for a broad area, covering several small areas, is used to derive an indirect estimator for a small area under the assumption that the small areas have the same characteristics as the large area. There is a trade off between bias and precision in the direct and synthetic estimators, the direct estimators have low or no bias but possibly low precision; the synthetic estimator is biased but has a higher precision. A choice between the two alternatives may depend on domain sample sizes. A key assumption is that the behaviour of the target variable in all areas with respect to the auxiliary data variable is the same – so that regressions for each area of the target variable against the auxiliary variable would give slopes very similar to the slope of regression in the broad area/whole population. The bias of the synthetic estimator depends on how true this assumption is. In the situation where the only available additional information is the population area sizes, broad area ratio estimator of small area total is given
\[
\hat{Y}_{d,BARE} = N_d \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i} = \frac{N_d \hat{Y}_{BA}}{N_{BA}}.
\]

If domain-specific auxiliary information is available in the form of known totals \(X_d^T\), then the regression-synthetic estimator \(\hat{Y}_{d,BARE} = X_d^T \hat{\beta}\) can be used as an estimator of domain mean, where \(\hat{\beta}\) is derived from any model has been used. These estimators can be applied for estimation when sample data are not available for the domain of interest, since the only information that is needed is the local covariate totals or means and the value of \(\hat{\beta}\), which is based on data from the broad area, e.g. the entire region, or country, covered by the survey.

This estimator has a straightforward formula and then it is very easy to calculate; it can be applied for domain of interest even when there is no unit from this domain in the sample. Moreover, when the assumption that small areas have the same characteristics as the broad area is fulfilled, it is unbiased. Otherwise, if some small areas into the broad area have specific characteristics, the estimates may be strongly biased.

Some specific synthetic estimators are briefly described below.

- **simple ratio-synthetic estimator**

  The following synthetic estimator uses a broad area survey estimate \(\hat{Y}_{BA}\) but can be used when the value of a single auxiliary variable, \(X\), is observed in the survey and its total is known from another source for each small area.

  The estimator, \(\hat{Y}_{d,RSE}\), is given by:

  \[
  \hat{Y}_{d,RSE} = X_d \frac{\hat{Y}_{BA}}{\hat{X}_{BA}}
  \]

  where \(X_d\) is the known value of the auxiliary variable for the small area \(d\) and \(\hat{X}_{BA}\) is the direct survey estimate of the total of the auxiliary variable at the broad area.

  Essentially \(\hat{Y}_{BA}/\hat{X}_{BA}\) acts as the slope of an implicit regression of the target variable \(Y\) and the auxiliary variable \(X\) which is then applied to the value of \(X\) for each of the small areas.

- **count-synthetic estimator**

  The synthetic estimator \(\hat{Y}_{d,RSE}\) takes the good direct survey estimate at a large (broad) area as a basis. However, it is possible that good direct estimators for the BA exist for a cross-classification of respondents – perhaps by sex and age groups. If population sizes or auxiliary variable totals values are known for these cross classifications for the small areas of interest, then equivalent synthetic estimators can be constructed. These are known
as post-stratified estimators (where the classification counts are the post-strata).

When population sizes are known, the count-synthetic estimator, $\hat{Y}_{d,CSE}$, is given by the following expression (remark: the BA suffix is removed for simplification in the following formula, one can think to estimates at national level – this simplification will be used whenever a cross-classification is introduced for the estimators given):

$$
\hat{Y}_{d,CSE} = \sum_g N_{dg} \frac{\hat{Y}_g}{\hat{N}_g}
$$

where $g$ represents the cross-classifications of post-strata, e.g. $g = 1$ to $6$ for three age groups by male and female; $\hat{Y}_g$ is the direct survey national estimate (e.g. the Horvitz-Thomson estimate) of the target variable for cross-classification cell $g$; $\hat{N}_g$ is the direct survey national estimate of the national population size for cross-classification cell $g$; and $N_{dg}$ is the known population size for cross-classification cell $g$ of the small area $d$.

Alternatively, when the value of a single auxiliary variable, $X$, is known at the cross classification groups at each small area and it is also measured in the survey, the combined ratio-synthetic estimator, $\hat{Y}_{d,CRSE}$, can be given by

$$
\hat{Y}_{d,CRSE} = \sum_g X_{dg} \frac{\hat{Y}_g}{\hat{X}_g}
$$

where additionally, $\hat{X}_g$ is the direct survey estimate of the auxiliary variable for cross classification cell $g$ and $X_{dg}$ is the known value of the auxiliary variable for cross classification cell $g$ of the small area $d$.

As it is shown, various formulation of synthetic estimation can be considered, however, all these formulation are based on taking the more reliable Horvitz-Thomson direct estimator for the broad area or domain (often, the whole population) and using it to derive an estimator for the small area/domain. The synthetic estimator is biased but has a smaller variance as survey data for the large area is used and hence the sample size is greater.

6.2.2 Composite estimates

According to Ghosh and Rao (1994), composite estimation is a natural way to balance the potential bias of a synthetic estimator against the instability of a
direct estimator by choosing an appropriate weight. The composite estimator of population total for a small area \( d \) can be defined as

\[
\hat{Y}_{dC} = \phi_d \hat{Y}_{dD} + (1 - \phi_d) \hat{Y}_{dS}
\]

where \( \hat{Y}_{dD} \) and \( \hat{Y}_{dS} \) are the direct and the synthetic estimator respectively, while \( \phi_d \) is a suitable weight \((0 \leq \phi_d \leq 1)\) for domain \( d \).

The expression of the composite estimator is given by a convex combination of direct and synthetic estimator, the weight of the first one grows as the sample size increases whereas decreases as the sample size gets smaller. Many of the estimators, both design or model based, have this basic form. The estimates calculated with a composite estimator are less biased than synthetic estimates. A suitable choice of the weights under a design based approach can be done in different ways (see Rao, 2003, for a comprehensive list of the methods). For instance, the weights can be obtained by minimising the MSE of the composite estimator, \( \hat{Y}_{dC} \), with respect to \( \phi_d \), under the assumption that the covariance between direct and synthetic estimator is small respect to the MSE of \( \hat{Y}_{dS} \). The optimal weights are given by:

\[
\phi_d \approx \frac{MSE(\hat{Y}_{dS})}{MSE(\hat{Y}_{dD}) + MSE(\hat{Y}_{dS})}.
\]

Rao (2003, section 4.3) provides a number of different type of composite estimator derived under a design based approach, including the Sample Size Dependent Estimator (SSD) and James-Stein Method as well as examples of their applications.

### 6.3 Model based estimators

Model based small area estimation is based on explicit models that relate survey data or survey direct estimates from small areas/domains to auxiliary data. The modelling approach fits a model to survey data and assumes that this model can be used to make predictions for out of sample units. The modelling approach can be used for continuous, binary or count data.

Different types of models are available and a decision needs to be made on which type of model is best suited to the available data. Classification of models are:

1. area versus unit level models
2. synthetic versus random effects models
3. linear versus generalised linear models
To assess the choice of model, a variety of models need to be statistically compared. This can often be subjective and will depend on sound statistical judgment and experience. It is also necessary to measure the goodness of fit - how well the model fits the data. Established methods are, however, available for measuring model goodness of fit. Different features can be incorporated into the model to increase a regression model’s power. Specific characteristics of the data can be taken advantage of to enhance the reliability of small area predictions. Which features one chooses to use depends on the strength of the enhancement given the quality of the data. Complicated models may be more difficult to fit and interpretation of results is not always straightforward. An area level model requires that area level auxiliary data are available for sampled and non-sampled areas/domains. The direct survey estimates need to be available for sampled areas/domains. In a unit level model individual survey data are required together with either unit level auxiliary data for each population element or at least population means of auxiliary variables for each small area/domain. The extra information available at the unit level provides additional explanatory power so unit level models are generally preferred to area level models. However, it may not be possible to use a full unit level model because of the difficulty in obtaining and matching unit level auxiliary data to the survey respondents. In general unit level models make use of the response variable data and auxiliary data at the finest level of detail available. Full unit level models use unit level auxiliary data and exploit the relationship between the unit’s status and the unit level auxiliary variables. Often the unit is an individual and auxiliary data can include information on a person’s age, sex, employment status, ethnicity etc. The unit level auxiliary data needs to be available for every small area/domain. If unit level survey data is not available or relevant, then an area level model is appropriate. Area level models require strong auxiliary data at the area level, otherwise the model constructed will have a weak association with the true results. This may even lead to the production of incorrect estimates. A key factor in deciding between a unit level and an area level model is whether most of the variation occurs at the unit level or more broadly at the area level. When there is a substantially larger variability at unit level, a unit level model will more efficiently partition the estimate of the variability between the levels thus giving more accurate area precision estimates. A word of warning is necessary here though. In the full unit model (where auxiliary variables are also at unit level) the inferred relationship between the survey variable of interest and the auxiliary variables can be different from that which would be inferred if area level values of auxiliary variables are used. This can exhibit itself in different values for the estimates of the regression coefficients sometimes even with different signs. It is an ecological effect, what is known as the ecological fallacy and its converse the atomistic fallacy. The ecological fallacy describes the fault of inferring individual behaviour from aggregate data relationships. It is well known in epidemiological studies that aggregate data studies
should not be used for this purpose. The atomistic fallacy involves the fault of assuming aggregate behaviour mirrors individual behaviour. In the small area estimation context it is this error which is of concern, for whether the relationship is inferred at individual or at area level, it is applied at area level in order determine the small area estimates. This leaves one in a quandary. Do you use only area level auxiliary variables even if unit level ones are available and will help to explain within area variability? Or do you use the unit level variables and risk incorrect area level estimates? One solution is to use both - area level auxiliary variables can appear as they are, unit level variables shown as divergences from the area level value. This will allow separate coefficients to be estimated for the area and unit levels of the same auxiliary variable. It will indicate whether an ecological effect is present - if not, the values of the coefficients in each pair of area and unit level variables will be similar. Either way the area estimates will be determined by using the area level variables and coefficients only. This is because the unit level auxiliary variables, being included as offsets from area level values, aggregate to zero. However, the inclusion of the unit level in the model affects the conditional estimation of the area level coefficient thus it is not equivalent to a model omitting the unit level variable. See Heady and Hennell (2000) for detailed discussion of the ecological effect.

On comparing synthetic and random effects models, we concentrate on the case of having a normally distributed continuous response variable and hence talk about a standard normal linear model. The synthetic regression models are also known as fixed effects models. If auxiliary data is available in the form of known area totals or means, then the regression synthetic estimator can be used as an estimator of the population of interest.

First the following model is fitted on the dataset of area terms

\[
\text{Area level response}_d = K + \beta_1 \left( \text{explanatory variable}_1 \right)_d + \beta_2 \left( \text{explanatory variable}_2 \right)_d + \cdots + \text{error term}_d
\]

where the suffix \(d\) refers to the small area of interest and \(K\) is constant term. The explanatory variables (often termed covariates) are obtained from the auxiliary data and the \(\beta\)'s are corresponding slope coefficients. Measurement error, omitted explanatory variables, unobserved individual heterogeneity, sampling error etc. are captured by the error term.

The regression synthetic estimator is obtained by applying the fixed effect terms to the area values of the explanatory variables as follows:

\[
\text{Regression estimates}_d = \hat{K} + \hat{\beta}_1 \left( \text{explanatory variable}_1 \right)_d + \hat{\beta}_2 \left( \text{explanatory variable}_2 \right)_d + \cdots
\]

The regression synthetic estimator is efficient providing that strong individual area effects (after taking account of explanatory variables) are not exhibited. Its
precision is dependent on a combination of the estimated standard errors of the coefficients and the constant K (the fixed effects) and of the estimated variability of the error term. However without strong individual area effects, the variability of the error term is small and often neglected.

In the case of the unit level model, it is fitted on the unit level response data in a similar manner:

\[
\text{Unit level response}_{id} = K + \beta_1 \left( \text{explanatory variable}_{1id} \right) + \beta_2 \left( \text{explanatory variable}_{2id} \right) + \cdots + \text{error term}_{id}
\]

Here i refers to the unit/responder within area d.

As with the area level model, the error term captures measurement error, omitted explanatory variables, unobserved individual heterogeneity, sampling error etc. The unit level regression synthetic estimate is then determined by applying the fixed effect terms to all population units in the area (where the auxiliary variable values are known). The area level regression synthetic estimate is then the aggregate (sum or mean as appropriate) of the unit estimates. Alternatively, it is obtained by applying the fixed effect terms to known auxiliary variable values for the area. Its precision is similarly dependent on the precision estimates of the fixed effects neglecting the small estimated variability of (in this case) the area mean of the error term.

Random effects regression models differ from synthetic regression models by allowing for between area variations. This is usually achieved by including area-specific random effects that account for between area variation beyond that explained by auxiliary variables included in the model. The general area level model is given by:

\[
\text{Area level response}_d = K + \beta_1 \left( \text{explanatory variable}_{1d} \right) + \beta_2 \left( \text{explanatory variable}_{2d} \right) + \cdots + \text{area random term}_d + \text{error term}_d
\]

So for each area d, there is a relationship between the area level response and values of explanatory variables. Associated with each area is also a random variable that reflects the variability within the area. Essentially the previous area error term has been separated into this random term and the remaining error term. It is common to assume that these area random variables are all independent and identically distributed with mean zero and a common variance. The above formulation is a standard linear model which includes design-induced random variables (error terms) and model based random variables (area random effects). The small area estimate is then determined by applying the fixed effects terms to all population units in each area. In addition values are generated for the area random term (but not for the error term) and this is added on to the
fixed part. In this case strong individual area effects mean that the variability of the area random term cannot be neglected in determining precision which is therefore based on the precision of the fixed effects and the estimated variability of the area random term.

Unit level random effects models can be used if unit level response data is available. When unit level auxiliary data are also available the model is given by:

\[ Unit\ level\ response_{id} = K + \beta_1 (explanatory\ variable_{1id}) + \beta_2 (explanatory\ variable_{2id}) + \cdots + area\ random\ term_{id} + error_{id} \]

So for each unit \( i \) in area \( d \), there is a relationship between the unit level response and value of explanatory variables. Again area random variables are included to reflect between area variations. In practice, unit level auxiliary data is not usually available and area means are commonly used instead.

The synthetic regression model assumes that small areas/domains have similar characteristics after taking into account any auxiliary data. If auxiliary data cannot explain area specific characteristics then a random effects models is more appropriate. It should be noted that, although more complex models may provide better estimates, they are usually harder to implement and interpret.

If the continuous data is normally distributed then a linear model can be used. A transformation of the data may be required before modelling to make the data normally distributed – scatter plots of the data should be analysed to check the assumption.

If the target variable is not continuous or normally distributed then a generalised linear model may be applied. These are a class of models such that, after application of an appropriate data transformation (the link function), the expectation of the transformed variable is modelled by a linear model. An example is the logistic transformation with binary data where the log of the odds of an event occurrence follows a linear model. Another is the Poisson model for count data. Unlike continuous data, count data only has discrete values ranging from zero to positive infinity. The Poisson model takes account of the discrete nature of the data and is usually the best model to use. The Poisson distribution assumes that the variance is equal to the mean and departure from this assumption is very common in real data. Adjustments to the Poisson model can be made to incorporate situations where the variance does not equal the mean (usually overdispersion).

A Poisson model with log transformation is the standard generalised linear model for count data.

With unit level data the target variable of interest is often binary. The logistic or probit model is more appropriate in this case. The logistic transformation is used to account for the binary nature of a Bernoulli or a Binomial distribution. It transforms the Bernoulli/Binomial distribution to a normal distribution. These
models give the probability that the unit will be in the population, for example, the probability that a household is in poverty. A Bernoulli model assigns each unit its own probability and these probabilities are allowed to vary from unit to unit. In a Binomial model all units within a group (e.g. age/sex) within each small area are assumed to have the same probability. Strong subject matter knowledge will help in deciding whether to use a Bernoulli or Binomial model.

So the choice of a linear or generalised linear model will depend on what the target variable of interest is. Discussion with the users is usually required to agree upon a suitable target variable. Often, it is not possible to produce estimates for the exact target variable that a user might want (because of insufficient data, too complex a model) and a simpler or more appropriate option may have to be settled for. For example, grouping a continuous response variable and fitting either a binary or multinomial model. This is often the case when the level of spread is too high. For example, modelling income for those in the top 10% of incomes – a binary model with income equal to high earner or otherwise may be more appropriate.

Quality of the estimates needs to be always assessed, usually using the mean squared error. Moreover the whole SAE estimation process needs to be assessed, so that other diagnostics e.g. a measure of bias by comparing modelled estimates with the direct survey estimates, a measure of the degree of overlap between the modelled estimates and direct survey estimates. It is also worth that the small area/domain estimates add up to the direct survey estimates produced at wider level of aggregation. Anyway, quality assessment methods are described in more detail in section 7.

### 6.3.1 SAE models and approaches to inference

The linear mixed models play an important role in SAE context. The random effects are intended to reduce the extra-variability not explained by fixed effects. Standard small area models generally consider only i.i.d. area random effects, whereas more realistic and efficient models might include further structured random effects, e.g. relating to the structure to meaningful components, such as time for repeated surveys or spatial autocorrelated random effects.

The general representation of a linear mixed model is:

$$y = X\beta + Zu + e$$

where $y$ is a vector of responses, $X$ is a known covariates matrix, $\beta$ is the regression coefficient vector of $X$, $Z$ is a known structure matrix of the area random effects $u$, which is a specific random effect vector, while $e$ is the random errors vector associated with sampling errors or the variation of individual of unit level. The basic assumptions are that the random area effects and the random errors
are i.i.d. with zero means and finite variances. In many applications the area random effects are assumed to be uncorrelated, but they may have a more complex structure allowing spatial and time autocorrelation structures (See EURAREA Consortium, 2004).

Different inferential approaches can be applied in order to derive the predictor and the correspondent measure of the precision of the small area estimates. These can be broadly classified into predictive and Bayesian method of inference. The predictive approach is widely used in SAE context, under the assumption that the variance components are known, the Best Linear Unbiased Predictor (BLUP) is obtained in order to estimate the target parameter by predicting the unknown non sampled quantities. The final estimator belongs to the class of linear, model-unbiased estimators and it is obtained minimizing a square loss function – i.e. the predictive mean square error (more details can be found in Searle et al, 1992).

Anyway, since variance components are usually unknown, Empirical BLUP (EBLUP) is obtained by plugging the estimation of variance components into the BLUP estimator. The estimation for the parameters $\sigma^2_u$ and $\beta$ can be obtained recursively, using Maximum Likelihood ML or Restricted Maximum Likelihood (REML) (see Cressie, N., 1992), or Method of moments. The inference into the Bayesian approach is instead based on the posterior probability function. The predictor of the target parameter is given by the mean of its posterior distribution. Anyway, whereas in the Empirical Bayes (EB) method the unknown model parameters are estimated from the marginal distribution of the data and then plugged into the Bayes predictor formula; in the Hierarchical bayes approach (HB), the fixed effect parameter $\beta$ and the variance components are considered random and a prior distribution of these parameters is assumed. The Bayes Theorem is used to derive the joint posterior distribution of the small area quantities of interest. (more detail on Bayesian methods into SAE framework see Ghosh and Rao, 1994, Pfefermann, 2002, Rao, 2003).

Although EBLUP, EB and HB methods can provide the same result under specific assumptions, (Rao, 2003), the Bayesian approach has several advantages. Different types of target variables, such as binary or count data, and more complex random effects structures, such as time or spatial correlation, can be easily taken into account. Models to smooth the survey sample variance estimates can be considered, as well. Another advantage of HB is that uncertainty about model parameters is directly taken into account into the posterior distribution of the small area estimates.

A different way for comparing approaches of inference is linked with the complexity of the methods. In fact, even if the theory of EBLUP estimator is very easy, the inference becomes more difficult as soon as the complexity of the model increases. Conversely, Bayesian methods allow to deal with high dimensional and sophisticated models, even if the underlying theory is commonly perceived as being more difficult than EBLUP ones. The model in the HB approach is specified
in successive steps which are easy to understand, although the entire model fitting process can be rather complicated. For this reason, in order to approximate the posterior distribution and hence the posterior means and variances of the small area quantities of interest are competed using the MCMC methods. The small area applications are based on two types of models that can be viewed as special cases of the general mixed linear model. In fact, if the vectors $y$ and $X$ involved into the general modelling are refer to small domain an area level model can be specified, whereas a basic unit model is defined when these vectors refer to the units belonging to the small area of interest. These two types of the models will be briefly described in the following subsections. Moreover, for dealing with categorical data, SAE methods based on Generalized Linear Mixed Models (GLMM) need to be considered. These models allow to deal with discrete responses that are very common in practice (see for more detail Jiang 1999 and Ghosh et al., 1998).

Small area estimation methods based on GLMM models have been mainly developed using a Bayesian approach to inference. For instance MacGibbon and Tomberlin (1989) uses an EB methods to estimate small area proportions under a logistic mixed model; Maiti (1998) applies an HB methods for the estimation of mortality rates into the disease mapping context. Moura and Migon (2001) extend the mixed effects logistic model of MacGibbon and Tomberlin (1989) by introducing a further component to account for spatial correlation in the binary response data. Moreover, a semiparametric model, based on nonparametric regression that allows to combine small area random effects with a smooth, nonparametrically specified trend are described in Opsomer et al., 2008. By using penalized splines as a representation for the nonparametric trend, they express the nonparametric small area estimation problem as a mixed effects model regression. This model can be easily extended to handle bivariate smoothing, allowing for instance a different way of treating the spatial correlation structure among areas. Finally, in order to account for the multivariate nature of the variable of interest and to obtain more efficient estimators, multivariate models for dealing with multiple responses are developed. Datta et al., 1996 have considered a multivariate area level model, which allows a reduction of the MSE of the small area estimates by exploiting the correlations with the other response variables; whereas Molina et. al. (2007) have considered the multinomial nature to explicitly model employment status. A generalization of their model can be found in Scealy (2009). Other examples on how multivariate categorical variables can be treated is the SPREE estimator, which can be considered as an extension of the classical synthetic estimator (Purcell e Kish (1980)) and its generalizations (Zhang and Chambers, 2004) in where linear (mixed) models on the parameters defining the interactions among variables of the cross classification is introduced.
6.3.2 Basic area level mixed model

In this type of model, the auxiliary information is at small area level, so if $\theta_d$ is the parameter to be estimated for each domain $d$, a linear relationship between the parameter of interest and a set of covariates whose values are known for each domain of interest is assumed, that is

$$
\theta_d = X_d^T \beta + u_d,
$$

where $X_d$ is the vector of covariates for domain $d$ and $u_d$ ($d=1,...,D$) is the domain random effect assumed to be distributed with mean zero and variance $\sigma_u^2$. The area random effects account for the variability among areas, not explained by the auxiliary variables in the model.

Moreover, a design unbiased direct estimators $\hat{\theta}_d$ for each area and the corresponding variance are supposed to be available (even if not necessarily for all the domains) and it is assumed the following sampling model

$$
\hat{\theta}_d = \theta_d + e_d
$$

where $e_d$ is the sampling error associated with the direct estimators of each small area $d$. It is assumed that $(E(e_d|\theta_d) = 0)$, being the direct estimators unbiased, and that $V(e_d|\theta_d) = \varphi_d$, where the sampling variances $\varphi_d$ are supposed to be known.

Combining the two previous equations a linear mixed model is obtained and it can be formulated as follows:

$$
\hat{\theta}_d = X_d^T \beta + u_d + e_d
$$

Normality for both $e_d$ and $u_d$ are commonly assumed for the MSE estimation, even if this assumption is not necessary for estimating of the parameters.

Note that, to avoid identification problems, the sampling variances are assumed to be known. Nevertheless, under the hypothesis of homoscedasticity of the random error $e$, the variance can be estimated either from unit level data or through a generalized variance function.

The fixed effects parameters $\beta$ and the variance, $\sigma_u^2$, of the random effects are generally unknown. In order to compute the estimates, under a model based small area approach, predictive approach to inference or empirical and hierarchical Bayesian methods can be used. These three methods are detailed explained in Rao (2003). For instance, the empirical best linear unbiased estimator (EBLUP) is given by

$$
\hat{\theta}_{d,\text{EBLUP,AREA}} = \gamma_d \hat{\theta}_d + (1 - \gamma_d) X_d^T \hat{\beta}
$$

where
\[ \gamma_d = \frac{\hat{\sigma}^2_u}{\hat{\sigma}^2_u + \varphi_d} \]

is the weight of the direct estimator and \( \hat{\beta} \) is the weighted least square (WLS) estimator of the regression coefficient vector \( \beta \). The weights are provided by a diagonal matrix whose generic element is given by \( 1/(\hat{\sigma}^2_u + \varphi_d) \).

If only the regression component is considered, a synthetic estimator is obtained:

\[ \hat{\theta}_{SYNTH\_AREA}^d = X^T_d \hat{\beta} \]

This estimator is built only on the basis of the relationship between target variable and the covariates and does not exploit the direct information. Of course, for domains with no data only synthetic estimates can be computed. The basic model assumes that the distributions of area random effects is symmetric, while sometimes skewness may be present. In this case, if transformation of variables do not reduce skewness, advanced methods may be employed (see e.g. H. Chandra and R. Chambers, 2007). Moreover, assumptions of normality with known variance might be unsustainable for areas with small sample size. The method is applied by U.S. Census Bureau’s Small Area Income and Poverty Estimates (SAIPE) program for poverty estimates since 1993, for U.S. school districts since 1995, and for the Census undercount by Statistics Canada. More examples can be found in Fay & Harriot (1979), Dick, J.P. (1995).

There are several extensions of the basic area model allowing to take into account time and spatial correlation patterns, as well as of correlated sampling errors, which should be considered when repeated surveys with rotated samples are involved. Most of NSIs surveys are conducted continuously in time, hence SAE methods which borrow strength from data collected in the past. Area level mixed models with different correlation structures of area and time random effects have been considered in EURAREA (2004). In addition, use of complex spatial and temporal structures can be found in several other papers. For instance, Tiller (1991) uses a time series approach for modelling the true unemployment rate, specifying an ARMA model for the formulation of the time correlation of sampling errors of the estimates of these rates. Moreover, Rao and Yu (1994) have generalized the area mixed model by using an AR model that allows for the combination of cross-sectional data with information observed in preceding waves of the survey.

A different approach for formalization of the correlation between the parameters with respect to area and time based on state-space model is considered in Singh et al,1994. In particular, within this framework, Peiffermann and Tiller (2006) have also considered benchmark constraints, so that the small area estimates sums to the direct estimates at an aggregated leve. They have developed an algorithm for state-space models, that explicitly takes into account the rotating panel design.

With reference to spatial autocorrelation among areas, Pratesi and Salvati (2008) have proposed a spatial EBLUP estimator based on area linear mixed model, in which the random area effect of each area is correlated with the random effects of its neighbours and modelled as Simultaneously Auto-Regressive (SAR) process. A further extension of the basic mixed area level model has been proposed to cope with the unmatched sampling and linking models, that happen when the target variable is not a linear function of population total or mean. In this case the assumption that $E(e_d | \theta_d) = 0$ may not be valid for areas with small sample sizes (You and Rao, 2002).

Multivariate mixed area level models have been developed for a multivariate response. This type of model leads to more efficient estimators for a small area quantity of interest than the univariate Fay-Herriot model, reducing also the MSE of the small area estimates by considering the correlations with the other variables (see Datta et al.,1996; Rao,2003). Within area level mixed model framework, alternative models can be specified in order to consider that the variable of interests are counts (i.e. number of poor households or number of employed people) or a proportion of these quantities respect to the population. In these cases, as alternative to the classical standard area-level estimators are based on normal distributions, more suitable distributional assumptions, like the Poisson or the binomial, can be considered. Into Bayesian approach several hierarchical models have been proposed, like the Normal-Poisson-logNormal or the Gamma-Poisson-logNormal models (see Trevisani, Torelli, 2007).

### 6.3.3 Basic unit level mixed model

The unit level mixed model can be used when unit-specific auxiliary variables are available in each small area. The area-specific random effect terms are considered in order to take into account the between area variation, through the correlation among units within a small area. The basic unit level linear mixed model formulated by Battese et al. 1988 can be expressed as follows:

$$ y_{di} = x_{di}^T \beta + u_d + e_{di}, $$

$$ u_d \sim iid N(0, \sigma_u^2), \quad e_{di} \sim iid N(0, \sigma_e^2) \quad \forall i = 1, \ldots, N_d \quad d = 1, \ldots, D. $$

Assuming simple random sampling has been involved at design stage, the model, which holds for the population units, is valid also for the sample units. For this, the model for the sample units can be written, using a matrix formulation, as:
\[ y_s = x_s \beta + z_s u + e_s, \]

where \( y_s \) is \( n \)-dimensional vector of the observed values for the variable \( y \), \( x_s \) is the \((n \times p)\)-dimensional matrix of the covariate values observed for the sampling units, \( e_s \) is the \( n \)-dimensional error vector, \( z_s \) is the \((n \times D)\)-dimensional incidence matrix of the sampling units in the small areas, and \( u \) is the \( D \)-dimensional vector of area random effects.

In order to obtain the small area estimates based on the above model, a predictive, empirical or hierarchical Bayesian approach can be employed (See Rao 2003, for more detail). For instance, within the predictive framework, the Best Linear Unbiased Predictor (BLUP) is obtained by minimizing the quadratic loss in the linear unbiased estimator class. The BLUP estimator depends on the variance components \( \sigma_u^2 \) and \( \sigma_e^2 \), that are usually unknown, then their estimates need to be computed. The variance components and fixed effects parameters can be estimated in different ways, for example by means of Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) (Cressie, 1992) methods.

Once the parameters of the model have been estimated, the Empirical Best Linear Unbiased Predictor (EBLUP) based on unit level linear mixed model is a composite type estimator. Letting aside the finite population correction factor, it is given by

\[
\hat{\theta}_{EBLUP,A}^d = \gamma_d \left[ \bar{y}_d + \left( \bar{X}_d^T \hat{\beta} - \bar{x}_d^T \hat{\beta} \right) \right] + (1 - \gamma_d) \bar{X}_d^T \hat{\beta},
\]

where

\[
\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_d}
\]

is the weight of the direct component, \( \bar{y}_d \) and \( \bar{x}_d^T \) are the sampling means of the target variable and the covariates in the small area \( d \) respectively; \( \bar{X}_d \) is the vector of population mean values for the \( p \) covariates, and \( \hat{\beta} \), \( \hat{\sigma}_u^2 \), \( \hat{\sigma}_e^2 \) are the estimates of the parameters of the unit level linear mixed model assumed.

There are several extensions of basic unit level model above described. Since the basic model does not consider that sample data are usually collected with a complex sample design, some methodological developments have been directed to specify more complex models, that can consider the features of the sampling design. For instance, Stukel and Rao (1999) proposed a two-fold nested error regression model for data collected from a stratified two-stage sampling. Moreover, when an informative design is used the inclusion probabilities of sampling units depend on the values of the target variable, then the model which holds for the sample data is different from the model assumed for the population data. Ignoring the sampling process in these cases, can produce severe bias in the predictor of the variable of interest. Following Pfeffermann and Sverchkov (2007),
this effect can be taken into account either using all the design variables used for the sample selection as covariates into the model or using the sample weights as surrogates of these variables. Anyway, the first option is untenable when either the design variables are unknown at the inference stage or when there are too many of them. In the last case, the model is over-parametrized. In any case, the significance of coefficients of regression associated to these variables can be used in order to evaluate the informativeness of the design. The second option is instead untenable because the sample weights are not available for non sampled areas or non sampled units. For this, they propose under a unit level mixed model an estimator that takes into account the sampling weights distribution within the sampled areas. Moreover, a Pseudo EBLUP estimator was proposed by Prasad and Rao (1999), starting from the basic unit linear mixed model. This model is weighted through normalized weights, achieving in this way a survey-weighted aggregated area level model.

Mixed models are based also on assumption on symmetry (or normality) of the random effects, which may not holds in practice, even after transformation of variables to reduce skewness. A relaxation of this assumption on the random effects models can be achieved by employing M-quantile models, recently proposed by Chambers and Tzavidis (2006) and Tzavidis et al. (2010). The M-quantile regression is a quantile-like generalization of regression based on influence functions, i.e. it is a generalisation of the ordinary quantile of order q. Hence, M-quantile model for small area estimation may protect against departures from assumptions of the unit-level nested error regression model. To extend to the small area estimation the M-quantile regression, Chambers and Tzavidis (2006) have suggested that the conditional variability across the population of interest can be characterised by the M-quantile coefficients of the population units, i.e. the grouping produces similar M-quantiles. The area level M-quantile coefficients can be defined as an area average of the individual coefficients. These M-quantile coefficients are use to predict the mean of target variable for nonsampled units. Anyway, this estimator is biased. Hence a bias correction by applying the Chambers-Dunstan distribution function was proposed by Tzavidis and Chambers (2007).

Moreover, the linear unit level mixed models are in theory applicable only for continuous observations, then some enhancement models has been considered in order to deal with categorical dependent variables. In that case, Generalized Linear Mixed Models (GLMM) can be considered (Jiang and Lahiri, 2006). Within the class of GLMM, logistic regression models with mixed effects are commonly used for estimating small-area proportions: Dempster and Tomberlin (1980) have proposed an empirical Bayes method based on a logistic regression model for the estimation of the census undercount for local areas. Respect to the EBLUP estimator based on linear model, that needs only the availability of the population means or total for each area, the logistic estimator needs the knowledge of the values of auxiliary variables for each unit belonging to the target population. In
order to cope this problem all possible profiles, built on the basis of auxiliary information, need to be considered (Malec et al., 1997). The units belong to these profiles have common probability level.

For repeated sample surveys, some extensions aimed to introduce time random area effects can be also considered. Unit linear models with i.i.d. area and time effects either independent or autocorrelated have been considered in Eurarea (2004).

Instead, with the aim to consider the spatial autocorrelation among areas, a unit level model with spatially correlated area effect has been proposed in Eurarea, 2004. The spatial correlation is introduced through the variance-covariance matrix of the random effects as a function of the distance between areas, in a way that the spatial correlation among areas decreases as the distance among small area increases. The spatial correlation can be modelled also by means of a SAR model. In this case, the random area effects of each area is correlated with its neighbours (see Chandra et al., 2007b and Pratesi and Salvati, 2008).

Finally, multivariate nested error regression model has been proposed in order to estimate more then one small area parameters of interest simultaneously. This type of model, applied in Datta et al (1999), allows to take into account the correlation among the characteristics under study observed in the sample units. Multinomial models are considered in Molina et. al. (2007), whose model has been further extended by considering categorical specific random effects (Scealy, 2009).

7 Diagnostics and accuracy assessment

Given that small area estimation relies on the use of models, either implicit or explicit, the validation of the methodology is essential. On one hand, from the external point of view, users and academic reviews should be carried out in order to guarantee the plausibility and utility of the estimates and the appropriateness of the methodology respectively. Indirect estimates should also be compared against related data sources, like census values from the past, or related indicators (see chapter V, section L “Previous experiences in UK” in WP3 Final Report), studying the level of agreement between them. On the other hand, internal validation is mainly based on the analysis of certain properties of the estimates and also on diagnostic measures, specially within the explicit model-based approach. Some of them are outlined in the next sections.

7.1 Design-bias exploration

A common diagnostic tool to check for overall bias (implemented in WP4) is based on fitting a regression line to the scatter plot of the indirect estimates vs. direct
estimates, and comparing it to $Y=X$ (see Figure 7.1), which provides both a visual illustration of bias, and a parametric significance test. On one hand, from the visual point of view, close examination of the plot could reveal systematic patterns in the data, which might require further investigation. On the other hand the results of the regression, and the significance test ($H_0 : \beta_0 = 0; \beta_0 = 1$) provide figures that can be used to compare the behaviour of several indirect estimators. Some issues that need to be addressed, like the presence of heteroskedasticity or influential observations, and examples of its use can be found in sections IV-D (“Model diagnostics) and V-L (“Previous experiences in UK”) in WP3 Final Report, and also in the case studies developed by France, Italy and Netherlands. In the case studies developed by Spain and Switzerland there is an adaptation to the case of Monte Carlo simulations where the value of the target parameter is known.

![Figure 7.1 Example of bias scatterplot](image)

**7.2 Over-shrinkage and other ensemble properties**

The degree of shrinkage can be empirically observed also in the regression of indirect estimates vs. direct estimates, specially when the difference between the slopes of the fitted line and $Y=X$ is large, as can be seen in Figure 7.2, where more shrinkage is apparent for the set of estimates in the left-hand plot than the ones in the right-hand plot.

A numeric measure of shrinkage based also on the comparison between direct and indirect estimates is the *Average Shrinkage Indicator (ASHR)* (in percentage):
Figure 7.2 Example of over-shrinkage

$$ASHR (\hat{\theta}^{\text{INDIRECT}}) = \frac{1}{D} \sum_{d=1}^{D} \left| \frac{\hat{\theta}^{\text{INDIRECT}}_{d} - \hat{\theta}^{\text{DIRECT}}_{d}}{\hat{\theta}^{\text{DIRECT}}_{d}} \right| \cdot 100$$

Sometimes the shrinkage of the indirect estimates is justified, as it is discussed in chapter 6 of the guide developed by the Australian Bureau of Statistics (see reference [55] in WP3 Final Report). If the relative standard error for each of the areas whose direct estimates fall outside the range of the indirect estimates is significantly higher than those that do fall within the range of the indirect estimates then the smoothing, i.e. the shrinkage, is warranted. A graphical representation can be done by plotting the absolute difference between each direct estimate and their overall mean, against their relative standard error. The range limits of the indirect estimates can be marked off on the graph to identify the direct estimates outside this range, and observe their error. Another possibility is to plot the absolute difference between direct and indirect estimates for each area against their error.

In the case study developed by Norway, a shrinkage-adjusted alternative to a classical composite estimator is proposed, justified under a constrained empirical Bayes approach:

$$\hat{\theta}^{A}_{d} = \phi_{d}^{\eta} \hat{\theta}^{\text{DIRECT}}_{d} + (1 - \phi_{d})^{\eta} \hat{\theta}^{\text{SYNTHETIC}}_{d} \quad \text{for} \quad 0 < \phi, \eta < 1 \quad (\eta = \frac{1}{2} \text{ is used})$$

Another alternative is proposed by Zhang (see references [131] and [133] in WP3 Final Report): he shows that the empirical best linear unbiased predictor (EBLUP) suffers from over-shrinkage as expected, and proposes a simultaneous estimator which has better ensemble properties from the frequentist perspective. Also a finite population bootstrap is proposed to derive the predictive intervals of small area population parameters.
Nevertheless the ensemble properties of the estimators can usually be analysed within the Monte Carlo simulation framework only. Some summary statistics which measure how an estimator performs as an ensemble estimator have been applied by Zhang (also in references [131] and [133] in WP3 Final Report):

1. **Averaged squared distributional error (ASDE):**

\[
ASDE(\hat{\theta}) = \frac{1}{D} \sum_{i=1}^{D} (\hat{\theta}(i) - \theta(i))^2
\]

where \(\hat{\theta}(i)\) and \(\theta(i)\) are the \(i\)-th order statistic of \(\{\hat{\theta}_d\}_{d=1}^{D}\) and \(\{\theta_d\}_{d=1}^{D}\) respectively, and \(D\) is the number of domains considered. The summands can also be measured relatively (divided by \(\theta(i)\)).

2. **Absolute relative distributional error (ARDE):**

\[
ARDE(\hat{\theta}) = \frac{1}{D} \sum_{i=1}^{D} \left| \frac{\hat{\theta}(i)}{\theta(i)} - 1 \right|
\]

3. **Relative error of the range estimator (RE):**

\[
RE(\hat{\theta}) = \frac{\max_i (\hat{\theta}_i) - \min_i (\hat{\theta}_i)}{\max_i (\theta_i) - \min_i (\theta_i)} - 1
\]

All the measures proposed can be used to analyse the behaviour of a single indirect estimator, but are also useful in comparative assessment, so among several competing estimators, the one with the minimum value of these measures is preferred.

### 7.3 Model diagnostics

Model validation is crucial within the model-based approach. The objective is to lessen the chances of introducing design-bias into the small area estimates due to poor model specification. Model building is a complex task which is essentially an iterative process that involves:

- Model selection from a set of plausible models.
- Model fitting to adjust the selected model.
- Model diagnostic to check the adjusted model.
If the last step shows deficiencies of the model, then a new model should be selected, fitted and checked for improvement. The iterative process will continue until a model is satisfactory.

Then, the model assessment is related to model diagnostic but, giving that it is very helpful to compare models for model choice, it is also related to model selection and fitting in some way.

According to the approach considered, appropriate statistical methods are applied to compare several models and to examine the underlying assumptions and features of the model evaluated. The large number of available possible models and approaches results in a wide collection of tools for model assessment which are contained in Rao’s book. Nonetheless, it is worthwhile to mention some of the tools available which have been used in the case studies.

When the availability of auxiliary information is high, a preliminary stage to choose one or several subsets of covariates is necessary. Stepwise selection has been used in the cases developed by France, Poland, and Netherlands, where in the latter a dimension reduction technique (Principal Components Analysis) has been used subsequently, in order to avoid the risk of over fitting, present when the number of covariates is high compared to the number of areas.

The use of auxiliary variables based on sample surveys, which are observed with sampling errors, will also pose problems, except when their variability is approximately constant over the areas, as it is discussed also in the Dutch case study. Within the context of social surveys, socio-demographic variables like sex, age, educational level, or household type are frequently considered as covariates because their predictive power. For each target variable a proper definition of groups of age or cross-classification of these variables could significantly improve the estimates. For this purpose, an interesting example of the use of two alternative methods, Classification and Regression Trees (CART), and a nonparametric regression method, penalized splines, can be found in the Italian case study. CART methods have been also used in the past to test the robustness of the models (see chapter V, section G “Previous experiences in Italy” in WP3 Final Report).

In the final stage of the selection process several criteria have been used, like log likelihood, AIC, BIC, and two measures implemented in WP4, conditional Akaike Information Critetria (cAIC), and Cross Validation (CV).

As model selection measure, cAIC is well-suited for small area estimation. It is relevant to inferences regarding the clusters, or areas, in the context of linear mixed. The criterion is

\[ cAIC = -2 \log p(\beta, \nu) + 2p_{eff} \]

where \( p(\beta, \nu) \) is the conditional likelihood for fixed and random effects vectors evaluated at their estimated values, and \( y \) is the data. The effective number of degrees of freedom is essentially given by the trace of the hat matrix \( H, \hat{y} = Hy \)

52
CV methods allow to test the robustness of the models, quantifying their predictive power by leaving out one or more observations when fitting the models, and subsequently assessing the model predictions for the left-out observation(s). It can be quantified in alternative ways, for instance averaging the prediction errors, or calculated as:

\[
CV = \sum_{d=1}^{D} \omega_d \left( \hat{\theta}_d^{DIRECT} - \hat{\theta}_d^{MODEL(-d)} \right)^2
\]

where \( \hat{\theta}_d^{MODEL(-d)} \) is the indirect estimate, obtained from the data excluding the \( d \)-th area and \( \omega_d \) are adjustable weights.

Examples and further details can be found in the Dutch case study and in chapter IV, sections D (“Model diagnostics”) and chapter V, section L (“Previous experiences in UK”) in WP3 Final Report.

Other measures applicable to linear mixed models, useful in comparative assessment, which have been used at the ONS are:

- % between area variability explained = \( \left( 1 - \frac{\sigma^2_u \text{ (full model)}}{\sigma^2_u \text{ (null model)}} \right) \times 100 \): measures how much unexplained area level variability remains in the model compared with what exists in a model with no explanatory covariates (consisting of just the intercept);

- Percentage of between area variability explained by one covariate in the presence of all covariates:

\[
\left( 1 - \frac{\sigma^2_u \text{ (full model)}}{\sigma^2_u \text{ (model excluding 1 covariate)}} \right) \times 100
\]

where \( \sigma^2_u \) is the variance of the area random effects.

Once one or several models has been selected, it is necessary to assess the fitting quality of the model(s). Some of the tools available are:

- Residual analysis to check if the model assumptions are fulfilled:
  - Histograms, Q-Q plots and box-plots to study their distribution (see the French, Norwegian and Dutch case studies).
  - Plots of model-based estimates vs. residuals to reveal patterns and identify residuals which require further investigation, as for example in the Swiss case study where the plots are also made for subsets of domains separately, according to a classification variable (geographical).
Maps of residuals to check for spatial randomness in the case of geographic small areas (used in the Swiss case study (see figure 7.4): the detection of patterns in the maps which cannot be explained by sample sizes would suggest either that the inclusion of area effects (fixed or random) in a basic non mixed model, or the use of spatial models if the base model is already mixed, could improve the quality of the estimates.
• If the models include area effects, the same tools can be used to analyse their distribution (see the Italian and French case studies).

• Diagnostics based on the comparison between direct and model-based estimates. Some of them, developed at ONS (see reference [2] in WP3 Final Report) have been also implemented in WP4 (and used in the Spanish and Italian case studies):

  – Goodness of fit diagnostic: allows to detect conditional bias in the model-based estimates. A Wald statistic is calculated as sum of the squared differences between direct and model-based estimates (across areas) inversely weighted by their variances

  \[ W = \sum_d \frac{(\hat{\theta}_d^{\text{DIRECT}} - \hat{\theta}_d^{\text{MODEL}})^2}{\hat{V}(\hat{\theta}_d^{\text{DIRECT}}) + \hat{V}(\hat{\theta}_d^{\text{MODEL}})} \]

  Under the hypothesis that the model-based estimates are equal to the expected values of the direct estimates, \( W \) will have a \( \chi^2_D \) distribution where \( D \) is the number of small areas in the population, provided the sample sizes in the small areas are sufficient to justify central limit assumptions.

  – Coverage diagnostic: evaluates the validity of the model-based estimates confidence intervals, under the assumption that valid 95% intervals for the small areas can be calculated from the direct estimates. The actual overlap proportion between the confidence intervals of the direct and the model-based estimates (across areas) is calculated and compared to the Binomial distribution (\( H_0 : p = 0.95 \)). An adjustment in the critical value (\( z_{\alpha} = 1.96 \) under normality) of all the confidence intervals is necessary to ensure a nominal 95% overlap, since the degree of overlap between two independent 95% confidence intervals for the same quantity is higher than 95%:

  \[ z'_\alpha = z_\alpha \left(1 + \frac{\hat{V}(\hat{\theta}_d^{\text{MODEL}})}{\sqrt{\hat{V}(\hat{\theta}_d^{\text{DIRECT}})}}\right)^{-1} \left(1 + \frac{\hat{V}(\hat{\theta}_d^{\text{MODEL}})}{\sqrt{\hat{V}(\hat{\theta}_d^{\text{DIRECT}})}}\right) \]

  Plots of confidence intervals can also be used to assess the global degree of improvement in the estimates, as can be seen in Figure 7.5, where the intervals for the estimates on the right-hand plot are wider than the ones in the left-hand plot, except in the final part of the graphs.
(right-hand side), provided that the MSE estimates are unbiased, as it is discussed in the Italian case study.

- Calibration diagnostic: provides some evidence regarding spatial bias/autocorrelation of model-based estimates. The amount of scaling required to calibrate the model-based estimates to ensure they sum to direct estimates at appropriate levels of aggregation (what is referred as “benchmarking”, described later in this section) can show whether any particular larger domain is estimated worse than any other (e.g.: rural areas vs urban areas, certain group of age and sex). An important issue is deciding the calibration level, because since the aggregated direct estimates are themselves subject to sampling variation, it is inappropriate to calibrate at a too low level of aggregation. An application and discussion about this can be found also in the French case study, where in addition, a summary measure is proposed (“sigma” indicator) to assess the importance of the disagreement between direct and model-based estimates at aggregated levels. Given a set of aggregated levels whose direct estimates are reliable enough, it is calculated as sum of the percentages representing the gap, without the sign. The lower the value of “sigma” is, the better the estimation procedure is.

### 7.4 Benchmarking

It is well-known that small area estimation needs explicit, or at least implicit use of models. These model-based estimates can differ significantly from the direct estimates, specially for areas with very small sample sizes. One problem arisen from small area estimation is that the aggregation of the model-based estimates do not usually agree with the direct estimates for a larger area. Benchmarking can be considered as a solution. It can also protect against potential model
misspecification and overshrinkage.
A simple way to achieve benchmarking is making a ratio adjustment:

$$\hat{\theta}_d (\text{mod}) = \hat{\theta}_d \left( \frac{\hat{\theta}}{\sum \hat{\theta}_d} \right)$$

where $\hat{\theta}$ is a direct estimator of the larger area which is an aggregation of the small areas. The modified estimators will be somewhat less efficient than the original, optimal estimators, but they avoid possible aggregation bias.

Under a hierarchical Bayesian framework with unmatched area level models, the posterior mean squared error (PMSE) is a measure of the variability of the benchmarked HB estimators obtained by a ratio adjustment. The PMSE can be computed as the sum of the posterior variance of the parameter of interest given the data and the squared difference between the benchmarked and the unbenchmarked small area estimates.

Another solution is to construct the “optimal” predictor under a model subjected to the benchmark constraint. This new approach has been investigated both from a Bayesian point of view as the frequentist but especially with time series models. In this context, several benchmarking procedures have been experimented as, for example, taking into account the stochastic nature of the constraint due to the sampling error of the direct survey estimate or, by contrast, treat it as a fixed value; also, it is possible to apply a double-benchmark approach. First, there is an adjustment of model estimates for an aggregation level 1 to the reliable direct national estimates and, in a second step, the second level 2 estimates (more disaggregated than level 1) are benchmarked to their respective benchmarked level 1 estimates obtained in the first step (accounting for the stochastic nature of the benchmarks). Some authors propose an improvement based on benchmarking the estimates to the aggregation of survey direct estimates within a group of “homogenous” small areas with respect to the variable of interest. The MSE of the benchmarked predictor is usually estimated by direct formulas or resampling methods such as the bootstrap.

One of the most valuable experience in benchmarking due to the complexity is provided by the statistical office of United Kingdom for small area estimates of unemployment and unemployment rates in Great Britain (see WP3 Final Report, chapter V, section L). This is a case in which estimates are requested for two non-nested geographies.

### 7.5 Maps

When the small areas are geographic, the production of maps for the estimates can be a useful tool to validate the results, specially from an external point of view, because users would detect unexpected spatial patterns in the model-based
estimates which could lead to further improvements in the model. Examples (from the Italian and Dutch case studies) can be seen in Figure 7.6. It is relevant to note that as the utility of maps lies in the colouring scale, the presence of extreme estimates will distort it, making more difficult the detection of spatial patterns, as can be seen in the right-hand bottom map, where the estimates for the majority of the areas have midrange to low colours, as opposed to the estimates (of a different variable) in the left-hand map.

7.6 MSE estimation

One of the most relevant measures of the quality of an estimator is the Mean Squared Error (MSE), so it is important that any small area estimate is accompanied by an estimate of its MSE which is usually calculated by analytic or resampling methods.

For design-based basic smoothing methods and in the case of design-based or approximately design-based estimators, the estimation of the MSE is reduced to the variance estimation problem by using the direct formula in the case of the Horvitz-Thompson estimator or a Taylor linearization variance estimator in the case of the GREG estimator. However, if the sample size in the small area is very small, this variance estimation is very unstable. For design-based estimators as composite ones, it is more difficult to estimate the MSE because of the bias esti-

On the other hand, the model-based estimators for SAE include a variety of techniques such as synthetic estimators, composite estimators, EBLUP and various Bayesian techniques along with a variety of models (area level or unit level models, GLMM, time series and cross-sectional models for example) that they form a very broad class of estimators as it has been established in section 3.2. In relation to this kind of estimators, the MSE is not building as the sum of variance and bias as from the design-based point of view of inference since the inference is with respect to the underlying model. The MSE estimation process runs parallel to the building of the “optimal” predictor for which we want to estimate the error. For example, the MSE of the EBLUP can be approximated by the sum of several components that reflect the uncertainty due to model fit, the ignorance of the fixed effects and the estimation of the variance components. The unbiased estimation of each component depends on the standard variance components estimators applied (method of moments, REML or ML). In the case of the EB predictors, a similar expression is also valid for the MSE, but the components are usually estimated by resampling methods such as the jackknife. The jackknife estimator of the MSE is nearly unbiased in the sense of having bias of lower order than $D^{-1}$, for large $D$ (total number of small areas). Also, for the EB predictor, there are alternative measures using the HB approach and the posterior variance. In the hierarchical Bayes approach the posterior mean is used as the HB predictor and the posterior variance as a measure of the precision of the HB estimator.

A required reading is the monograph by Rao (2003) which provides a comprehensive treatment of the MSE estimation problem for most of the model-based methods. Other good discussion of model-based small-area estimators and the estimate of the MSE is included in the paper written by Jiang and Lahiri (2006) (see reference [51] in WP3 Final Report). Given that, as the use of resampling methods for MSE estimation is more recently in model-based methods, a brief and instructive appraisal of the developments in this field is provided by Rao (2007) (see reference [94] in WP3 Final Report). Moreover, a good summary of new developments is included in WP3 Final Report, chapter IV, section B. Finally, a remark about fitting unit-level models and the method of sample selection because, in model-based inference methods, the MSE is calculated with respect to a model and one might conclude that the sampling design and weights are irrelevant as long as the model holds. However, on some circumstances, this is not true and the model holds for the population values but not for the sample data; then ignoring the effects of the sampling scheme biases the predictors and increases their MSE (the sampling design is defined as ”informative” that
has already mentioned in section 4). The effect of this kind of designs can be controlled by including all the design variables in the model as covariates but, if they are many, fitting and validation of such models may be impossible. A good reference for the reader are the paper of Pfeffermann and Sverchkov (2007) (reference [13] in WP3 Final Report) or the Valliant’s book (2001) titled ”Finite Population Sampling and Inference”.

7.7 Monte Carlo simulation

Monte Carlo simulations are a good tool to reproduce a survey framework to allow us to carry out the comparative assessment in an empirical way than otherwise can not. Speaking in a general way, these simulations or experiments comprise the following steps:

1. Construction of a data set with individual data reproducing the population (Census or administrative register, pseudo-population or artificial population).

2. A sample is drawn from this individual data set replicating the sampling scheme applied in the real-life as close as possible.

3. The small area estimation methodologies are applied to the sample data selected in the previous step.

4. Steps 2 and 3 are repeated K times with independent applications.

5. Evaluation measures are calculated.

In relation to the step 1 the first model to resemble the real world, Census or administrative registers, is the most deterministic of the three because all elements are fixed. A pseudo-population consists in a mixture of deterministic and stochastic models as, for example, when each record of a single sample survey (or more than one sample survey when are cumulated together more survey occasions of the same survey) is duplicated as many times as its corresponding sampling weight. Finally an artificial population is completely stochastic and is obtained generating data from a model.

The number of repeated samples, K, drawn from the completely known finite population is variable. Usually the minimum value is K=100 and the maximum is K=1,000, with an average value of K=500. In some special cases where it is not very laborious to sample, the value of K is 10,000.

Unfortunately, very few published simulation studies provide enough details to allow readers fully understand all the process, giving details of how the study is performed, analysed and reported, including the definition of specific objectives in
the study, determining the procedures used to generate the data sets, the number of simulations performed and the evaluation measures calculated to assess the different methods.

In Small Area Estimation, a common reference is the EURAREA project (Enhancing Small Area Estimation Techniques to meet European Needs) and, if we talk about Monte Carlo simulations, the reference volume 1. The section B of this volume is dedicated to the simulation experiments carried out within the project framework to evaluate the performance of both the estimators themselves and of the confidence intervals that are generated as part of the estimation process. The measures used to assess the performance of small area estimation method in EURAREA cover three key areas:

- The level of agreement between the estimates and the true values for each small area and whether there is any systematic pattern (bias) to the errors that result from the estimation process.
- The coverage level of the confidence intervals derived by each method.
- How well the distribution of the small area estimates produced by each method approximates the true distribution of area values.

For the first one the evaluation measures are typically the following ones: let \( \hat{\vartheta}_d(k) \) be the estimate of the parameter \( \vartheta_d \) for the small area \( d \) in the \( k \)-th replicated sample

1. The Relative Bias for the small area \( d \) is: 
\[
RB_d = \frac{1}{K} \sum \left( \frac{\hat{\vartheta}_d(k)}{\vartheta_d} - 1 \right) \times 100
\]
The absolute value of the RB is much more informative since it avoids zero when the estimator bias exists and is symmetrical about zero.

2. In order to summarize the results for all the areas, the simplest measure to be considered is the average taken over all of them. This measure is named the average of absolute values of RB and it is defined as: 
\[
AARB = \frac{1}{D} \sum_{d=1}^{D} A RB_d
\]

3. The Relative Root Mean Squared Error for small area \( d \): 
\[
RRMSE_d = \frac{100}{\vartheta_d} \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left( \hat{\vartheta}_d(k) - \vartheta_d \right)^2}
\]
The statistician should expect a small value to ensure the good behaviour of the estimator. However, if the population parameter to be estimated is too small, the value will be large even if the estimator error (in absolute terms) is small. For this reason and in these cases, the RRMSE is replaced with the EMSE as defined below.
4. The average of the values of RRMSE: $ARRMSE = \frac{1}{D} \sum_{d=1}^{D} RRMSE_d$

5. The Square Root of the Empirical Mean Squared Error for the small area $d$:

$$EMSE_d = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left( \hat{\vartheta}_d(k) - \vartheta_d \right)^2}$$

In order to estimate the distribution of the small area values, a second group of overall evaluation criteria can be obtained considering all areas:

1. The maximum of the absolute values of RB: $MARB = \max_d |RB_d|$

2. The maximum of the values of RRMSE: $MRRMSE = \max_d |RRMSE_d|$

Another way of assessment is evaluating the coverage of the confidence intervals derived using the small area estimates, and measure the proportion of simulations in which the confidence interval contains the true area value. This indicator is called Coverage Rate and is usually based on the estimated confidence intervals at the 95th percentile. Formally, for each replicated sample the confidence interval is calculated and let

$$I_M(k) = \begin{cases} 1 & \text{if } Y_M \in C.I \\ 0 & \text{otherwise} \end{cases}$$

then the Coverage Rate for the small area $d$ is equal to

$$COV\left(\hat{Y}_M\right) = 100 \frac{1}{K} \sum_{k=1}^{K} I_M(k).$$

As it has been said above the purpose of the simulation exercises is to evaluate the performance of the estimators themselves. However, the simulations also provide us with the opportunity for an empirical investigation of all the methodological issues which together help to explain the performance of them.

For example, sometimes, in order to compare all the selected models, the measures mentioned are calculated along with the Absolute Relative Bias (ARB), which is the absolute value of RB and the Relative Improvement in the Mean Squared Error (RELIMP) of one model over another model:

$$RELIMP\left(\hat{\vartheta}_d\right) = \frac{MSE_{MODEL1}(\hat{\vartheta}_d) - MSE_{MODEL2}(\hat{\vartheta}_d)}{MSE_{MODEL1}(\hat{\vartheta}_d)}$$

62
Another measure of accuracy is the Median Absolute Relative Error (MdARE) defined as:

\[ MdARE = \text{Median}_k \left\{ \left| \hat{\vartheta}_d(k) - \vartheta_d \right| / \vartheta_d \right\} \]

The RRMSE and MdARE give the same message about accuracy when the response variable is continuous, but there might be differences in the conclusions drawn from the two measures when the variable is binary.

8 Software tools

In this section, suggestions in order to help users when applying SAE methods in large scale surveys will be given. To this aim, references to the case studies developed in the WP5 will be useful for underlining some practical issues related to the use of software for the analysis on real survey data. This section is organized following the flow-chart included in these guidelines and the general process that should drive the choice of the most appropriate SAE method. Specifically it is intended as a brief review of the available software for the application of small area estimation methods. Recommendations on the use of some specific software tools will be given, as well. In pursuing these objectives a detailed description of codes or procedures will inevitably be left aside. For a more extensive overview of the software tools the reader can refer to the WP4 report.

8.1 Design based estimation

Basic smoothing estimates can be easily computed using one of the many options available to perform design based estimation. Design based estimates, i.e. estimates resulting from applying direct estimators, generalized regression estimators (GREG) or calibration estimators, can be easily computed applying the correspondent procedures available in all the main statistical packages and described in the WP4 report. For instance PROC SURVEYMEANS in SAS and the functions included in the R library survey provide practical tools to deal with design based estimation. Furthermore, many statistical offices developed their own software tools to perform calibration process, e.g., CLAN, GENSEESE, or variance estimation, e.g., POULPE.

Whenever direct estimates cannot be considered reliable enough, synthetic and composite estimators can be applied borrowing strength from data belonging to a larger area. The synthetic regression estimator based on standard regression models can be easily computed by using the main statistical software reviewed in WP4.
8.2 Model based estimation

Dedicated software to perform all the small area estimation process is available for model based estimation, besides some general procedures for fitting models underlying the main SAE methods. For instance, the standard regression model on which the synthetic regression estimator is based, can be fitted very easily inside the main statistical packages. Once the parameter estimation is attained, some extra computations are needed to obtain small area predicted values and MSE estimates.

8.2.1 Ad hoc software tools

Among dedicated software surely is worth mentioning the SAS macros produced by the EURAREA project. In fact, as resulting from the analysis carried out in WP4, SAS procedures and EURAREA codes are the most used software tool to perform small area estimation among the NSIs participating in the project. One of software tool provided allows to perform small area estimation using standard methods based on linear mixed models, i.e. synthetic and EBLUP estimators based on unit and area level models. All the estimates are accompanied by MSE estimates and confidence intervals. Furthermore, additional SAS macros are available for downloading: a macro for area level small area estimation with time varying area effect; the EBLUPGREG code including several estimators from the GREG estimator to the EBLUP based on linear mixed models, which may contain spatial and time correlation structures; a SAS code implementing spatial EBLUP based on unit level mixed models with spatially correlated area random effects; codes for applying the Structure Preserving Estimator (SPREE) and its generalizations (Generalized Linear Structure Models, GLSM, and Generalized Linear Structure Mixed Model, GLSM); SAS codes allowing the introduction of sampling weights in unit level models. Extended documentation including training examples is available for each EURAREA software.

Nevertheless, users may encounter difficulties in defining properly the requested input and understanding the output. Furthermore, lack of accuracy may occur for the area level EBLUP since the variances of direct estimates do not take into account sampling design effects. Besides, some changes in the code are needed if one needs to apply only one estimator at a time. Moreover, it may be desirable a different specification of the auxiliary information for each predictor. This may be particularly important when users can only apply area level models because no unit level data set is available, or if they want to reduce the number of covariates included in area level models with respect to unit level models.

In the R framework, functions included in the package SAE2 allow the users to compute unit and area level EBLUP estimates. In addition, the package contains the function SEBLUP.area which permits to compute small area estimates based
on the Fay-Herriot model with spatially correlated area effects according to a simultaneously autoregressive (SAR) specification. The model is fitted using either maximum likelihood (ML) or restricted maximum likelihood (REML). MSE estimates are also provided. The functions contained in the package do not deal with the case of empty small areas, i.e., not sampled areas. Some lack of generality are also encountered in using \texttt{SEBLUP.area}.

Specific R functions have been developed within WP4. R routines for basic unit and area level linear mixed models were developed, besides specific functions for spatial unit level mixed models and logistic mixed models. Moreover tools to perform model selection and model evaluation are also provided (see next section). Several R functions are provided in the web-site of the SAMPLE project. These functions allow to fit several unit and area level models, including models with uncorrelated or spatially correlated area random effects, and models with temporal random effects, either independent or correlated temporal effects. Routines for the M-quantile estimator are also provided, as well as codes for the prediction of poverty measures based on unit level models. Furthermore for poverty and social exclusion indicators estimation specific R functions for small area estimation have been released by the AMELI project and can be found in the web-page of the project.

If users wish to perform Bayesian small area estimation, several procedures implementing MCMC techniques are available in the more common statistical software for linear or generalized linear mixed models. The common way to perform Bayesian analysis is to use the dedicated software WinBUGS, for Windows, or the counterpart OpenBUGS, for Unix or Linux. As already reported in WP4, WinBUGS codes for applying SAE methods are available on the Bias project website. They allow to perform small area estimation based on Bayesian specification of area and unit level linear models with independent or spatially dependent random area effects. Some extra code lines are needed to compute predicted values for empty areas. Several WinBUGS codes are also available in literature to apply Bayesian techniques to SAE problems. For a review see what reported in the WP4 report. Details about how to perform Bayesian SAE with statistical software, as R, SAS, SPSS, Stata, or MLwiN are reported in the WP4 report.

### 8.2.2 Time consuming efficiency

Time consuming can be a serious issue when dealing with large data sets. A comparison of the performances of some tools available in R and SAS for SAE has been evaluated in WP4. The aim of the comparative analysis was to evaluate in terms of computer intensity the performances of functions and dedicated software available in R and SAS. In details R function \texttt{lme} has been compared to PROC MIXED available in SAS, while the R function produced by this project to compute small area estimates using a standard unit level linear mixed model,
namely \texttt{mixed.unit.SAE}, has been matched up with the EURAREA SAS macros to compute small area estimates with the standard methods. For more details about the features of each function see the description available in the WP4 report. The results showed that the R function \texttt{lme} is to be preferred to PROC MIXED in terms of computer time consuming. As regards the comparison of the function \texttt{mixed.unit.SAE} and the EURAREA macros, the former resulted to outperform the SAS counterpart when the number of sampled areas is not too large, i.e., less than 250. On the contrary, the latter is more efficient when the number of areas is quite large, that is when the number of sampled areas is larger than 500. Improvements of the R function \texttt{mixed.unit.SAE} are under study to increase the performances when dealing with a large number of small areas.

When moving to the Bayesian framework several packages are available in R for hierarchical models. Although these packages have been developed for different applications, they can be used for SAE as well. Since these packages are tailored to a specific class of models, they are usually faster than BUGS software. It is worth mentioning that a package for hierarchical Bayesian SAE will be developed by Statistics Netherlands. The package will implement numerical integration over the posterior density without relying on more time-consuming MCMC methods.

8.2.3 Model selection and model evaluation

In this section references to specialized software for the quality assessment measures, which are described in WP3 are given. Users can refer both to the guideline section devoted to this topic and to the WP3 report for more details on these techniques. Preliminary analysis for model choice and selection is crucial for a good formalization of the estimator. Most of the existing functions to fit linear mixed models or generalized linear mixed models display the values for some of the most used model selection criteria, for instance AIC and BIC (see section 7.3). No specific code is still available for the computation of conditional AIC. On the contrary, the R functions for unit and area level LMMs produced in WP4 display also the values for conditional AIC and cross validation, which are more proper criteria for the selection of mixed models. In fact, they allow a more extensive model comparison analysis. For instance using the cross validation criterion, users can compare unit and area level mixed models, or fixed and mixed models. Besides R functions to evaluate model bias have been produced. The model diagnostics implemented in the codes are the model bias diagnostic proposed by Brown et al. (2001) and based on the comparison between model based and direct estimates.
9 Overview of the case studies

This summary gives a short overview of the case studies conducted by the partners in the ESSnet on Small Area Estimation. The experience in the field of SAE is different between the partners. Italy, Spain, the Netherlands, Norway and Poland have been involved in SAE research for a long time, whereas France and Switzerland have almost no experience in this domain. Spain, Italy and the Netherlands already started to consider and study SAE in the early 2000s. SAS and R are the software packages used by the partners for the case studies.

Based on the Health Survey of 2005, Italy was interested in the estimation at the Health Districts level of counts of people having at least one specialist visit done, obese people, and woman aged between 50 and 69 who had at least one mammography done. The covariates are demographic statistics and given by the labour force survey like educational attainment and household description. They investigated the EBLUP based on a linear mixed model (LMM) and a generalized linear mixed model (GLMM) namely the logistic mixed model. Both area and unit level models were considered. The Italian study shows that the small area estimation methods seem to perform well. The synthetic and EBLUP estimates give opposite results, but the EBLUP estimates are preferred. The case study focused on some nodes of the flow charts of the three stages and that of model based SAE. First of all great attention was devoted to the model selection phase. Standard criteria such as BIC and AIC were used to choose the most proper set of auxiliary variables. Furthermore two alternative methods to define the mid points for the definition of classes for the continuous variable age were considered: a semi-parametric method based on the spline functions, and the classification regression trees. Residuals and the predicted area random effects were graphically investigated to detect spatial patterns not included in the model specification, and to verify the normal distribution assumption respectively. The distributions of the residuals and the random effects seem close to normal, but the distribution of the random effects display a spatial autocorrelation. Finally bias diagnostic criteria were applied to evaluate the properties of the estimates produced by using the alternative SAE methods.

The aim of the French case study is to test other methods and the possibility of a quality assessment of the estimation of the unemployment in France. The French Labour Force Survey from 2007 is used to carry out the analyses. The Zones d’Emploi are the areas. Specific economic indicators are used as auxiliary data and come from the Census, the DEFM (Demandes d’Emploi en Fin de Mois), from the typology of geographical units, the RMI (Revenu Minimum d’Insertion) database and finally from the ZUS (areas considered as sensible). Fay-Herriot and Poisson models are considered in this study. In the case of the Fay-Herriot model - fitted using small areas with more than 50 respondents - they find that in about one half of the ZE, the contribution of the direct estimator from the LFS
is more than 20% in the final EBLUP. There is a positive correlation between the dimension of the ZE and the contribution of the direct estimator. On the other hand, the ”median” Zone d’emploi has a mean square error divided by a factor approximately equal to 4 compared to the direct estimation. In the case of the Poisson model, the mixed approach necessitates the inclusion of an overdispersion parameter. Except if the set of domains is restricted, getting rid of the most specific areas, the variance of the random local effect is not significant. It appears difficult to define an optimum set of auxiliary variables: when one restricts this set, the traditional quality criteria decrease but still some non-significant variables remain, and the overall small area estimation becomes (a little) closer to the national direct estimate.

The statistical aim of the Norwegian study is the estimation of domain mortality rates for the calculation of theoretical life expectancy at the Municipality level. The mortality data are compiled from the death records in the population registers over a five-year period from 2004 to 2008. The basic approach is composite estimation across the Municipalities within each age-sex group. The composite estimator can be motivated by a Poisson-Gamma model that is commonly used in disease mapping, or under a less restrictive semi-parametric set-up. To avoid the jerky parameter estimates among neighbouring age-sex groups, they also developed a variance component model across the age-sex groups while allowing for clustering random effects within each age-sex group. In addition, adjustment for over-shrinkage is found to be necessary. Despite the use of register instead of sample survey data, the approach follows readily the standardized 3-stage process outlined in the Guideline of WP6, where the basic composite estimates provide the basic smoothing results, and the variance component model and the over-shrinkage adjustment are developed to address the notable shortcomings. Empirical results with considerable reductions of the mean squared errors (i.e. compared to the direct estimates) and plausible between-Municipality variation of the calculated life expectancies can be produced under the proposed approach.

Based on the Labour force Survey of 2008, the objective of the Polish analysis is the estimation of percentage of unemployed persons in the population at the NUTS3 level (the 66 regional domains). The auxiliary information is provided by registers, like register of Unemployment, Vital Statistics Register and Tax Register. They consider synthetic small area estimators based on a linear mixed model (LMM) and on a linear mixed model at area level with pooled sample estimates of the within-area variance. They also applied the EBLUP based on a linear mixed model defined at unit-level at area level. The point estimates of the different estimators are computed and compared with each other and with the registered unemployment for the Register of Unemployment.

By means of simulation experiments, Spain is interested in the estimation of the census estimator of the basic structure of the population at the municipality level. They based their study on the 2011 Census of Population and Housing. They eval-
uated the Census estimator, the ratio-synthetic estimator and the count-synthetic estimator. They also considered different performance criteria like percentage of relative bias, percentage of relative root mean squared. Based on the coefficient of variation defined by the RMSE, the census estimator accuracy is acceptable for the majority of the municipalities. They also find that the census estimator is approximately design-unbiased. The synthetic estimates showed a significant bias, which was mitigated refining the broad areas through cluster analysis. Some of the model diagnostics implemented in WP3 were also applied, obtaining results which, after some considerations, were consistent with those obtained from the performance criteria.

Based on a simulation study too, Switzerland tests small area estimation based on the Census from 2000. The areas are defined by the 2896 communes. The Census provides also the auxiliary information (age, sex, nationality, civil status and NUTS2). They considered the EBLUP defined at unit-level and area-level, the synthetic estimator defined at unit and area level and the You-Rao estimator defined at unit-level. They also estimated the empirical best predictor at area-level model and the binomial predictor at unit level, both based on a logistic mixed model. For all the estimators the RMSE decreases with the commune size. The RMSEs among the model based estimators were very similar. The estimators based on binomial random intercept model do not outperform the other small area estimators. Compared to the GREG, the model based estimators showed in general considerably lower RMSEs. The gain in RMSE is due to low standard deviations that widely compensate the observed design bias of model based estimators. The bias based on single samples is investigated by means of regression technique. This technique gives a good indication on the size of the shrinkage effect.

Germany conducted a preliminary analysis of the possibilities of SAE methods to provide labour statistics at NUTS 3 level, essentially covering stage I of the process described in section 4 of this report. Variables to be estimated (number of unemployed and employed people) were derived from the micro census, covering about 1% of the population. To limit the scope, the study was restricted to one state (North Rhine-Westphalia). Register data about employment is available at area level, from the German Federal Employment Agency. While these register variables differ significantly from the target variables, the correlation is strong. Hence, these variables are likely to have good predictive power and may be suitable to build SAE models with. Analysis of direct estimates for the small areas would be the next step if this study was continued. In particular, the between area variance would be of interest, and the effect of using synthetic estimates instead of direct estimates (this is step II, see section 4).

Finally, the Netherlands use a crime victimization survey, the National Safety Monitor, from 2008 and 2009 to carry out their analysis. 25 Police Districts are the small areas. Five target variables are analyzed: the perceived nuisance (mea-
sured on a ten point scale), victim of poverty and of violent crime, satisfaction with the police and feelings of unsafety (given in percentages). Auxiliary information is demographic information like gender, age, level of urbanization and mean house prices but also information from Police records on reported crimes and offences, information on the NSM from preceding periods, and estimates on crime victimization from another survey carried out in parallel. Statistics Netherlands have based the small area estimation on the basic area level model (Fay-Herriot), which is a linear mixed model. Two estimators are considered, the EBLUP and the HB (Hierarchical Bayesian). Particular attention is paid to selecting the most suitable covariates based on the cAIC criterion. It is investigated whether Principal Component Analysis offers benefits in terms of reducing the dimension of the space of potential covariates. The selected models lead to estimates with margins reduced by approx. 40% compared to the direct estimates. An outstanding issue is model choice in repeated surveys: can the model be altered from one edition of a survey to the next, or must the model be kept constant but perhaps suboptimal? The case study includes some discussion on the use of covariates that are measured with error, such as outcomes from another survey.

10 Conclusions

The aim of this report is to propose general criteria for setting up the process for production of official domain estimates through the use of indirect estimators. The importance of these methods, known in literature as SAE techniques, is due to the increasing demand for more detailed and reliable statistical information. As a consequence, in the last decades, many methodological developments to improve the basic models have been proposed and then validated through several experimental studies. Moreover dedicated codes for SAE have been produced, making the application of these methodologies computationally possible. For their aim, SAE techniques allow to produce estimates for a disaggregation of areas which are in between of the usual planned administrative domains, planned of large scale sampling surveys, and the complete enumeration of units collected in census or available from administrative registers. Therefore, with these methods, updated detailed statistical information can be produced, keeping under control both budget constrains and non sampling errors. If this is the main purpose of SAE methods, the main drawback is that these type of estimators do not assure the design consistency and unbiasedness of the estimates. Furthermore, whereas the direct estimators allow to produce estimates and sampling variances for a large set of target parameters, by means of use of final sampling weights, SAE techniques usually demand for a study and fit of models for each of variables of interest. For these reasons, SAE techniques have specific aspects not in usually use for the production of official statistical information. These guidelines should
be then viewed as an attempt to fill this gap promoting in this way the application of SAE methods in NSIs. The SAE estimators are mainly obtained by means of indirect methods, which often are derived assuming a super-population model. We think that one of the main aspect when SAE methods need to be used is the process that it should be followed. The main steps of any estimation process are

1. definition of needs and uses;

2. identification of available information coming from survey data (target and auxiliary variables), frame data (area or domain population means of auxiliary variables) and metadata information;

All SAE methods borrow strength from external data (i.e. from other areas, from other survey occasions and/or covariates), this usually produce a smoothed estimates with a degree of smoothing strongly dependent on the implicit or explicit model assumed. When small area estimates need to be produced, the statistician may easily apply the basic estimators (synthetic or composite in Figure 4.1), which, in some situation, are able to provide a satisfactory answer to the user needs in terms of information and reliability of the required estimates. Otherwise at the cost of a major complexity in estimation process, the statistician may apply more enhanced models, that reflect realistic aspects of the target variable (non-linearity, e.g. logistic for binary response) or more specific relationships (temporal correlation for repeated surveys or specific formalization for taking into account the complexity of the sampling design). Then when dealing with model based SAE techniques, it is crucial the model choice step. Here it is worthwhile remarking that difficulties in applying "composite" estimator (model-based) may occur when too much area level variability is present in the data. In fact, given the sample data, in some applications the estimation of between area variation measured by the random area effect variance may result to be equal to zero. Hence, the "composite" estimates are equal to the synthetic type estimates. In order to overcome this problem, when the exiting between area variation is not identified by sample data, the area random effect variance component may be deduced by the last available census data (see Rao, 2003, pp 134; National research council, 2000, pp. 51). Another fundamental step of the SAE process is the supplement of the resulting estimates with a quality assessment checks. That is usually performed by:

- **internal evaluation** "to establish that a model is performing well in terms of its assumptions" (National research council, 2000, pp.54-57). Basically, regression diagnostics with graphical plots for checking model assumption and comparisons with direct estimators are usually recommended.

- **external evaluation** "involve comparison of estimates from a model with target or true values that where not used to develop the model" (National
research council, 2000, pp. 57-61). This second evaluation is difficult to apply because it requires external sources on target variable of a good quality, i.e. census data can be used as target using the sample data collected in the year of the census.

Note that for the model-based SAE techniques, bias and accuracy are usually evaluated on the basis of the superpopulation assumed. Anyway, design sampling properties could be assessed on pseudo-population whenever it is possible. Furthermore, for repeated surveys it is not convenient to repeat the choice of the model for every occasion of the survey. In fact, different choices of the model may have an impact on the comparison between waves, as well and it may be costly in terms of resources and time. Anyway, it is essential to update the fit of the model for each survey occasion, as well as the assessment of the quality of the produced estimates over time. This mean that when starting a SAE program, if more occasions of the survey are available, the model may be studied over time and when the SAE program has been settled, same model could be applied if no signal from diagnostic suggests us that fundamental changes in relationship among target variable and the auxiliary variable occurred.

Standard estimates are usually completed with an evaluation of variability i.e variance, coefficient of variation or confidence intervals. When indirect estimates are disseminated, MSE is used as measure of variability, which includes variance and bias. Anyway, users should be aware that the proposed MSE for the model based estimators is different from the variance or CV of the direct estimators, which is based on the randomization of samples. Nonetheless, under specific conditions it has been shown that the MSE tracks the design MSE (see Rao 2003). Finally, before disseminating SAE estimates, the data provider should make users aware of all the aspects mentioned in these guidelines, highlighting the differences between the approach used and the standard direct estimator, usually used for calculation of estimates disseminated by NSI’s. The auxiliary variables that have been employed should be described as well as should be well described the kind of model that has been assumed. Furthermore, the most relevant diagnostics procedures that have been applied to ensure the quality of the estimates produced should be highlighted. As pointed out in Australian Bureau of Statistics (2005) in section 2, before the dissemination step it is important to assess whether the SAE estimates meets the user expectation. In particular, when some funds need to be allocated among areas on the basis of SAE estimates, a sensitive analysis should be performed in order to determine how the final decisions are sensitive respect to change of SAE estimates and with reference also to precision level of the produced estimates. The sensitivity analysis is useful not only to test the robustness of the results but is also useful for verifying the absence of errors in the specification of the model, changing reasonably the assumptions underlying the model used. In SAIPE program (National research council,2000) at paragraph 6 is showed a
sensitivity analysis carried out for investigating the impact of different scenarios when funds need to be allocated. Moreover, when estimates are disseminated all the different type of errors that may occur applying SAE methods should be well described to users. Finally, confidentiality issues can arise since the dissemination of small area estimates might increase the risk of disclosure at individual level (Ichim, 2011).

References


79


