ISSN 1680-8789

SERIES

STATISTICS 97

Disaggregating data in household surveys

Using small area estimation methodologies

Isabel Molina

UNITED NATIONS

ECLAC

UNFPA

Thank you for your interest in this ECLAC publication

Please register if you would like to receive information on our editorial products and activities. When you register, you may specify your particular areas of interest and you will gain access to our products in other formats.

Register

UNITED NATIONS

ECLAC

www.cepal.org/en/publications

facebook.com/publicacionesdelacepal

www.cepal.org/apps

SERIES

STATISTICS 97

Disaggregating data in household surveys

Using small area estimation methodologies

Isabel Molina

UNITED NATIONS

ECLAC

UNFPA

The Spanish version of this document was produced by Isabel Molina, Researcher in the Department of Statistics, Carlos III University in Madrid, in collaboration with Andrés Gutiérrez, Regional Expert on Social Statistics, and Álvaro Fuentes, Statistician, both of the Social Statistics Unit of the Statistics Division of the Economic Commission for Latin America and the Caribbean (ECLAC), as part of the activities related with the tenth tranche of the United Nations Development Account. This translation into English was made possible by the financial support of the United Nations Population Fund, for the text to be used as reference material for the capacity-building activities performed under the thirteenth tranche of the United Nations Development Account.

The views expressed in this document, an unofficial translation of a Spanish original which did not undergo formal editing, are those of the authors and do not necessarily reflect the views of the Organization or the countries it represents.

United Nations publication
ISSN: 1680-8789 (electronic version)
ISSN: 1994-7364 (print version)
LC/TS.2018/82/Rev.1
Distribution: L

Printed at United Nations, Santiago
S.22-00167

This publication should be cited as: I. Molina, "Disaggregating data in household surveys: using small area estimation methodologies", *Statistics series*, No. 97 (LC/TS.2018/82/Rev.1), Santiago, Economic Commission for Latin America and the Caribbean (ECLAC), 2022.

Index

Tables

Figures

Summary

Household surveys are widely used as a tool for obtaining information on people's socio-economic status and well-being. However, the accuracy of household survey estimates decreases significantly when it comes to making inferences for population groups who represent disaggregations for which the survey was not designed. It is possible, in this context, to use estimation processes that combine information from household surveys with existing auxiliary information at population level, such as censuses or administrative records.

This paper offers a methodological guide to the combination of survey statistical techniques with probabilistic models in order to produce disaggregations for interest groups, known as small area estimation (SAE) techniques.

A description of the problem of disaggregation when there is insufficient data is followed by a discussion of three techniques for achieving the proposed objective. Firstly, there is a review of the direct estimators (adapted directly from the surveys), which have the advantage of being free of bias, albeit with low accuracy, when disaggregation is applied. Next, there is an analysis of some so-called indirect estimators, whose functional form is similar to that of the direct estimators, except that they rely on auxiliary population information to improve accuracy. Following this, probabilistic models are introduced to improve the statistical properties of estimators of interest. Modelling can be done at two levels: at the level of the individuals of interest (households or persons) or at the level of disaggregation categories (subgroups of interest). This discussion of the theory includes illustrations and examples which are supported by R statistical software.

Finally, a practical application is offered for some of the methods reviewed and conclusions are drawn regarding the feasibility of their use in this specific problem.

Introduction

Household surveys provide fundamental information for measuring the living conditions of a country's population and are an essential tool for defining and monitoring public policies in several areas. As a source they make it possible to generate accurate and unbiased information at a national level and for the disaggregations considered in the survey design.

There is a growing demand for information for specific population groups and smaller geographical areas. For example, the global framework of indicators for monitoring the Sustainable Development Goals states that information should be disaggregated not only geographically (in subregions of interest such as provinces, municipalities, or districts), but also by income group, sex, age, race, ethnic origin, immigration status and disability status. However, the reliability of inferences drawn from the indicators decreases as the sample size decreases, so it is generally not possible to achieve the desired levels of disaggregation with a suitable accuracy.

Thus, in the last decade, there has been a rise in the concept of data disaggregation, meaning numerical information that has been collected from different sources, or measured by means of multiple variables or even different units of observation and that is compiled in an aggregated and summarised form. The purpose of this aggregation is to present society with estimates of interest that have good statistical properties which can be used to extract information and even formulate public policies in each of the subgroups of interest.

This document serves as a guide for the disaggregation of statistical data related to the living conditions of individuals, whether geographically (at a regional level) or by population subgroups. Chapter I begins by describing the problem of disaggregation of statistical data (section I.A); specifically, it describes exactly in which situations this problem occurs and defines the terms and concepts that are commonly used and that will also appear throughout this paper. Section I.B goes on to establish up to what level it is suitable to disaggregate the statistical data, given that, due to the decrease in sample sizes, as the direct estimates are disaggregated, the sampling errors increase, thus rendering these estimates too volatile and therefore unreliable. For example, let us consider a population divided successively at different levels; Spain, for example, is divided into autonomous communities which in turn are divided into provinces; these are divided into regions, and finally the regions are divided into municipalities. In the European Union as a whole, the common nomenclature NUTS (Nomenclature of Territorial Units for Statistics) is used and countries (NUTS 0) are divided into regions called NUTS 1, NUTS 2, etc. Section I.B provides indications on the maximum level of disaggregation of direct estimates at which indirect estimates would come into play. The latter are much more reliable because they use different data sources to borrow information from all areas. The extent to which it is advisable to use these indirect estimators is also discussed, as it is advisable to contain the possible bias of these estimators. Thus, recommendations are made concerning cases where it would be prudent not to produce any estimate. In any case, the survey can be redesigned to give a more comprehensive coverage of the domains for which statistical data is required. It should be borne in mind that, at a local level, the information or knowledge possessed by local communities could contradict the data provided. It is, therefore, essential to establish to what extent it is advisable to disaggregate the estimates so that the data produced is of sufficient quality and realistic, without straying too far from local knowledge. Finally, section I.C contains a review of the various methodologies that provide indirect estimates which go beyond the limits of disaggregation of direct estimates. Specifically, a review is made of the basic indirect estimators, which include synthetic and composite estimators, and model-based estimators, which are perhaps most widely used for obtaining reliable estimates at highly disaggregated levels. “Model-assisted" estimators, which use a working model but do not require goodness-of-fit of the model in order to maintain their unbiasedness, are dealt with in Chapter III along with direct methods, as they have good theoretical properties for large sample size areas.

Chapter II reviews various indicators dealing with individual quality of life; specifically, measurements of poverty and inequality. A family of poverty measurements, called the FGT family, is defined in more detail and will be used to illustrate the various procedures in the following chapters. A description of each procedure will explain how it would be applied to the estimation of indicators of this family and, for some of them, examples will be made using the sae (Molina and Marhuenda, 2015) R package, which the reader can copy.

Chapter III then gives a detailed overview of the usual direct estimators. Basic direct estimators such as the Horvitz-Thompson and Hájek estimators (section III.A) are included, as well as model-assisted estimators; specifically, generalised regression estimators and calibration estimators (section III.B), together with estimators of their sampling errors. The computation of direct estimators in R is illustrated by means of two examples.

Chapter IV reviews some basic indirect estimators such as the synthetic post-stratified estimator (section IV.A), synthetic area-level (section IV.B) and individual-level (section IV.C) regression estimators, and composite estimators (section IV.D). These estimators are included only because they provide a simple illustration of the ideas underlying the more sophisticated methods included later in Chapter V. Again, two examples are included that demonstrate the calculation of synthetic and composite post-stratified estimators.

The model-based methods in Chapter V are significantly more realistic than the basic indirect methods and are better suited for use in real applications as they provide potentially less-biased estimates. Model-based methods include estimators based on the most popular area-level model (section V.A) and those based on the basic individual-level model (section V.B). There is a demonstration of how to obtain these estimators in R using three examples. This section includes the ELL method, traditionally used by the World Bank to estimate poverty and/or inequality indicators (section V.B), since in principle this method considers the basic model at the level of the individual. However, as we shall see, this method is essentially synthetic, and should therefore perhaps be included in Chapter IV which deals with synthetic estimators. There is also a description of the EB method (section V.D), which estimates general indicators in the same way as the ELL method but improves on this method by considering that there is heterogeneity between areas and, consequently, produces more accurate estimates. The HB procedure in section V.E obtains estimates very similar to the EB method but with lower computational cost in the case of large populations, especially when it comes to providing the error measurements (mean squared error) of these estimates. Finally, section V.F outlines specific methods for estimating indicators that take the form of proportions or means of binary variables. Although in principle other methods can be used to estimate these indicators, such as those in sections V.A or V.B, these can generally provide estimates outside the natural area of a ratio. In some cases, the estimates obtained by different methods may differ only slightly.

On the other hand, some of the methods described are only applicable to linear indicators, i.e., they are additive in the values of the variable of interest for the units of the area, as means or totals. Other methods, such as the individual-level model-based methods ELL, EB and HB in sections V.C, V.D and V.E, are designed to be able to estimate general indicators defined as a function of the values of a continuous variable (e.g., income) in the units in the area; values for which a model is assumed. Methods based on area-level models are, in principle, applicable to many types of indicators, as long as the necessary assumptions are verified, but in practice it is difficult to verify such assumptions (such as the unbiasedness of direct estimators) for non-linear indicators. Therefore, in principle, they are more suitable for the estimation of area means or totals. In any case, after each method is described, there is a summary which specifies the indicators to which it might be applicable, the necessary data requirements other than observations of the variable of interest obtained from a survey, and the pros and cons of each method compared to methods that would be applicable to the same type of indicators.

It should be noted that it is not possible to give a detailed description of all existing methodologies due to limited space. A detailed description of some of the most widely studied methods having the right properties is included. These will help in the understanding of more complex methods. Some of the extensions of these main methods are cited, redirecting the reader to the corresponding bibliography should further information be required. Methods with unknown theoretical properties are not included, even though they may be promising. Nor are there details for procedures that require excessive mathematical formulation, such as the estimation of the mean squared error of the estimators in Chapter IV and section V.B. In both cases, the reader is redirected to the bibliography where this material can be found. In the case of the basic indirect estimators in Chapter IV, another reason for not including such material is that there are no known reliable estimators of the mean squared error of the estimators that are also different for each area. There are estimators that are excessively unstable but different for each area, or stable but the same for all areas, but not both at the same time. So, this is an unresolved problem.

An important clarification should be made regarding the approach used for evaluating the quality of an estimator. There are three alternative approaches for evaluating the properties of estimators, but often each type of estimator is evaluated using only a measurement calculated in respect of the natural approach for this estimator. The direct and basic indirect estimators are evaluated in respect of the distribution of the sample design, i.e., in respect of all possible samples that can be drawn from the population using the particular survey sample design. In this case, the values of the variable of interest in the units of the population are considered to be fixed values and only the units that are selected for the sample vary (according to a randomised procedure). A good estimator is, thus, the one that has a good average performance for all possible samples, with the values of the variable in the population units fixed.

On the other hand, model-based estimation methods are evaluated in respect of the distribution triggered by the considered model, conditionally on the observed sample. In other words, the values of the variable of interest in the individuals of the population are considered to be random and are generated by a model known as a superpopulation model. According to this approach, the census of our variable is a possible realisation of a random vector that follows a model (or probability distribution). The estimators are evaluated with regard to all the possible censuses generated by the model in question. In other words, a good estimator would be one that performs well on average for the infinite possible censuses of values of the variable of interest generated by the model, leaving the individuals that appear in the sample constant (although their values of the variable of interest vary, since they are extracted from a census).

Finally, Bayesian methods, such as the HB method in section V.E, are evaluated conditionally on the observations of the variable of interest in the sample (posterior distribution). In other words, an estimator will be evaluated with regard to the distribution of the conditioned indicator to the available data, rather than averaged over the possible values of that data.

There is no consensus on what is the best approach for evaluating small area estimators. The "under sample design" approach is non-parametric as it assumes no model. This means that the error measurement provided under this approach (usually the mean squared error) captures the estimation error across the possible samples, without the need for model checking. This is the approach preferred by government agency statisticians. The "under the model" approach assumes a model, but fixes the sample obtained, providing the error for the particular sample one has, rather than an average for all possible samples that could be drawn. With this approach, the measurement of error captures the uncertainty throughout the possible censuses that the model generates, i.e., through the possible realities that could occur, with the values observed in the sample also varying. Finally, the Bayesian approach considers the indicators in question to be realisations of random variables that follow a distribution, and provide measurements of error in the form of descriptives of the distribution of those indicators, conditional on the observed values of the sample, i.e., for the particular sample observations that have been obtained, rather than averaged over all their possible values.

As stated above, each estimation method is usually evaluated on the basis of its natural approach. In other words, the error measurements that accompany the estimates to assess their quality; specifically, the mean squared errors, are usually calculated with regard to the approach used for obtaining the estimates. This means that the mean squared errors of different estimators, having been obtained using different approaches, are not directly comparable. However, it is known that, if the hypotheses assumed by the considered models are verified, these mean squared errors are, in fact, comparable when averaged over a large number of areas of the same sample size. In addition, the mean squared errors under the design of model-based estimators are not easy to estimate and no acceptable estimators are known. On the other hand, by performing a prior model check to verify that the model fits the available data adequately, the estimators of the mean squared errors under the model, which are relatively stable, can be compared with the mean squared errors under the design.

I. The problem of data disaggregation (or small-area estimation)

A. Description of the problem

Official surveys conducted by National Statistical Institutes, as well as by Regional Statistical Institutes and other agencies or institutions at a supranational or international level, are designed to produce statistical data at a particular level of aggregation for either geographical or socio-economic subdivisions of the population. For example, the Socio-economic Conditions Module (*Módulo de Condiciones Socioeconómicas* (MCS)) of Mexico's National Survey of Household Income and Expenditure (*Encuesta Nacional de Ingresos y Gastos de los Hogares* (ENIGH)) is designed to provide estimates of poverty and inequality indicators at a national level for all 32 states (31 states plus Mexico City) disaggregated by rural and urban areas, every two years. However, in this country there is a requirement to produce estimates every 5 years at municipality level. This situation also occurs frequently in other countries and areas which means that, once a survey has been conducted, with sample sizes established to produce reliable estimates at a given level of aggregation, a demand for data at a more disaggregated level is subsequently produced. To do so, we want to be able to use the data from this survey without incurring additional costs due to an increase in the sample. However, in the case of Mexico, the ENIGH subsamples taken from each municipality do not allow for reliable direct estimators to be obtained in all of them and, in fact, more than half of the municipalities lack observations. This is the problem that often arises when one tries to produce statistical data for smaller subdivisions than originally planned.

To avoid this problem to a certain degree, aspects of the sample design could be improved prior to carrying out the survey. For example, it is possible to increase sample sizes in areas where this is necessary (with a corresponding increase in cost) or to distribute the total survey sample size across areas more efficiently. Although there are various mechanisms for improving the sample design and having a sufficient minimum of data in all subdivisions of the population, "the client always demands more than what has been specified at the design stage" (Fuller, 1999).

In the literature, the subdivisions for which statistical data (or estimates) is required are commonly referred to as "areas" or "domains", regardless of whether they are geographical or socio-economic demarcations. When estimating a specific indicator in one of these areas, we use the term direct estimator to describe an estimator that uses only the survey data for that area. The usual direct estimators are unbiased or virtually unbiased in respect of the distribution of the sample design, i.e., across all possible samples that can be drawn from the population using the corresponding sample design. However, if the survey was not planned to estimate at such a disaggregated level, the sample size in some of the areas may be too small, resulting in excessively large sampling errors for the direct estimators of the indicators of interest in those areas. The areas where this occurs, regardless of population size, are referred to in the literature as "small areas". Therefore, it is not the population size of the area that confers the adjective 'small' since, in many cases, areas of large population size (e.g., states in the USA) are considered as 'small areas' if direct estimates of sufficient quality are not available. Specifically, the term "small area" refers to areas, in which the direct estimator of the indicator of interest is inefficient due to the insufficient number of observations obtained (or surveys conducted) in that area. For example, when it comes to producing estimates of poverty and inequality indicators based on the Socio-economic Conditions Module of Mexico's ENIGH, municipalities would be considered small areas, since the survey is not designed to obtain precise estimates for them.

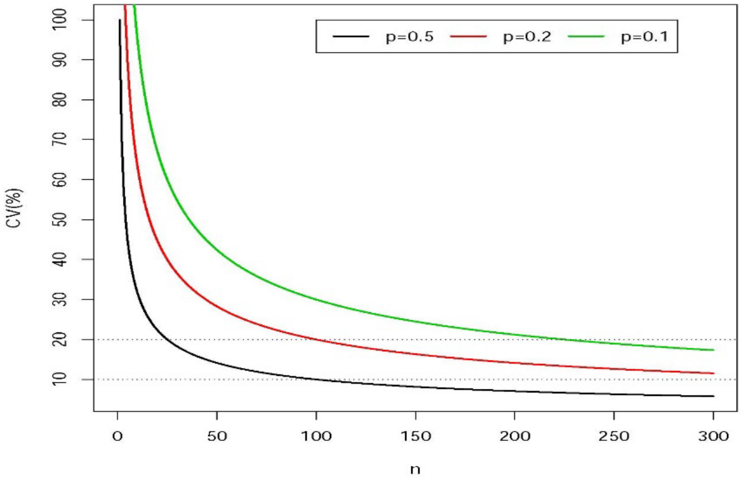
Estimates at a very detailed geographical level are often represented in the form of cartograms or maps showing the corresponding regions with different shades or colours representing different degrees of magnitude of the indicator of interest. For example, the World Bank produces disaggregated poverty or inequality maps for many countries, (see e.g., Elbers, Lanjouw and Lanjouw (2003)). These maps, as well as the specific estimates, are an essential tool for monitoring living conditions in the different regions of a country and are used by governments and international agencies to plan regional development policies. It is highly recommended to supplement estimates with measurements of estimation quality (usually sampling errors). As with estimates, these can also be plotted on a map.

B. Limitations to the disaggregation of statistical data

Although there is no formal definition, an area is referred to as "small", as stated above, when the sampling error of the direct estimator considered for the indicator of interest is not acceptable. However, there is no universal upper limit for this sampling error, above which the area where it is estimated is considered as "small". Each National Statistics Institute or international agency establishes its own limit for the relative sampling error or coefficient of variation (CV), above which statistical data is considered unreliable and therefore is not published. This data is sometimes published with some indication that it lacks the required quality. Nor is there a specific sample size below which the area is considered small, since the sampling error varies depending not only on the sample size, but also on the indicator that is being estimated and the specific estimator that is used. For example, when estimating the mean of a continuous variable (e.g., mean income) with a given maximum sampling error, a smaller sample size is often needed than for estimating the proportion of individuals possessing a given characteristic (mean of a binary variable), especially if that characteristic is very rare or very common, i.e., when the actual proportion is close to zero or one.

Figure 1 illustrates the minimum sample size needed to obtain a specific maximum CV for the sampling proportion under simple random sampling. The required sample size is seen to vary depending on the true value of the proportion to be estimated. Specifically, this graph shows how, in the case where the true proportion is $p = 0.5$, a sample size of about $n = 25$ is sufficient to ensure a CV of the sample proportion below 20%, whereas for $p = 0.2$ at least $n = 100$ units are needed, and for $p = 0.1$ more than $n = 200$ units are needed in the sample. Therefore, it is not possible to establish a minimum sample size of the areas that guarantees the desired level of efficiency for any estimator and/or any objective indicator.

Figure 1
CV of the sample proportion $\hat{p}$ according to sample size n, for each value of the true proportion p
(In percentages)



Source: Prepared by the author.

In particular, some common poverty indicators are proportions. For example, the poverty incidence, also called the rate of individuals at risk of poverty, is the proportion of individuals with incomes below the poverty line. This threshold is the value of (equivalised net) income below which an individual is considered to be at risk of poverty or exclusion. Similarly, certain types of deprivation are measured as the proportion of individuals with access to certain basic services such as health, housing, and food. As already stated, the sample size needed to obtain direct estimates of these indicators with sufficient quality is usually larger than that needed to estimate means or totals of quantitative variables.

Although there are no universal upper limits for sampling errors (and no lower limits for sample sizes) above which statistical data is of insufficient quality, some National Statistical Institutes agree that data is "unsuitable for publication" when its relative sampling error or CV exceeds 20%. Thus, for these institutions, areas for which direct estimates of a given indicator of interest have a CV greater than 20% would be considered "small" for these indicators. For example, Mexico's municipalities would be small areas when estimating poverty indicators from the ENIGH. In these areas, it would be necessary to increase survey sample sizes or use 'indirect' methods in order to produce statistical data of a sufficient quality for publication.

Not only do "indirect" estimation methods consider sample data from the domain or area of interest, but they also use sample data from other areas or domains. These estimators use information from other variables (known as auxiliary variables) that are related to the variable of interest. This relationship is considered similar for all areas and is shown by a model that links them by means of common parameters. By estimating the parameters common to all areas using all the data in the sample (overall the sample is usually large), the use of a greater amount of information provides more efficient estimators (compared to direct ones). These estimators tend to slightly compromise the bias under the design, in exchange for greatly increasing the overall efficiency of the estimator, evaluated in terms of mean squared error.

The improved efficiency of indirect estimators with respect to direct estimators is greater the smaller the sample size of the area. However, these tend to improve in most areas, including many with large sample sizes. In fact, some indirect estimators (see Chapter V) have the useful property of converging to a direct estimator as the sample size of the area increases. Indirect estimators that possess this property can, therefore, be used for all areas, regardless of whether they are "small" or not, thus reducing the importance of having a more exact or more formal definition of "small area".

In practice, however, one must determine the level of disaggregation to which it is suitable to continue using conventional direct estimators, the level at which to resort to indirect estimators, and even, if it is suitable, to produce statistical data for any possible level of disaggregation, which at the limit would be at the individual level. In virtue of the above, it is advisable to use the direct estimators at the level for which the CVs of these estimators do not exceed the limit established for any of the areas. If the aforementioned limit for any areas exceeded, it would be more advisable to use indirect estimators for the areas of that level.

It should be emphasised that it is not advisable to produce estimates for any area because, if the model cannot be verified exactly (virtually no model is exactly true), the bias under the design of indirect estimators increases as the sample size decreases. Although the mean squared error of the indirect estimators remains smaller than that of the direct estimator, it is not advisable to overly compromise the bias under the design. It, therefore, makes sense to set an upper limit for the relative absolute bias of an indirect estimator and decide not to produce data for areas where that limit is exceeded. This limit will be set according to the requirements of the data user (e.g., 10% or 5% of relative absolute bias). It is, therefore, recommended:

- To use the direct estimators for the entire population and for the higher levels of aggregation, as long as the direct estimators for all areas of that level have a CV below the established limit.
- For more disaggregated levels, we will use indirect estimators in areas for which the relative absolute bias does not exceed the pre-set maximum amount.
- Finally, for areas where the indirect estimators have relative absolute bias above the maximum amount, it would be advisable not to produce estimates or to modify the survey design in order to achieve a minimum sample size in all areas of interest.

The bias under the design of an estimator is not known as it depends on the true value of the indicator in question. In some cases, it can be approximated theoretically. Another option is to obtain it empirically using simulation experiments. These experiments can be carried out by simulating the facts as far as possible, e.g., based on a previous census or by using survey data to generate a census and drawing samples from it. These experiments have a very important additional use, which is the validation of estimation methods in situations where the true values are known. In both cases, it is possible to determine the required minimum sample size of the areas that would avoid exceeding the upper limit of the relative absolute bias of the estimator of the indicator in question, (see section V.A). Thus, when producing estimates with the actual data, indirect estimators would only be used in areas in which the sample size exceeds this minimum sample size.

C. Methodologies for overcoming the limitations of disaggregation

As previously stated, if the aim is to avoid sample increases due to the corresponding costs, or if the demand for data at a more disaggregated level has occurred after the survey has been conducted, a cost-effective way to obtain more reliable estimators for all areas of interest than direct methods is to use indirect methods. These methods do not only use the survey data for the area concerned but use data from other areas that have some similarity to the area in question. This similarity is usually illustrated by means of a model (representing a set of hypotheses). The simplest indirect estimators are based on unrealistic assumptions and may therefore have considerable bias. These include synthetic estimators, which do not take into account the heterogeneity that usually exists between areas. Well-known synthetic estimators are the synthetic post-stratified estimator and the synthetic regression estimator (Chapter III). Other classic indirect estimators are the composite estimators, which are calculated as a weighted average between a direct estimator and a synthetic estimator, and include the known sample size dependent estimator or the optimal composite estimators. The weight given to each estimator does not depend on the goodness of fit of the model assumed by the synthetic estimator. Moreover, in practice the weight of the direct estimator is usually close to one and, therefore, little information is borrowed.

Slightly more sophisticated indirect estimators, which take into account the existence of diversity between areas, are those based on regression models. There are two main groups of regression models used for estimation in small areas: area-level models and individual-level models, although it is also possible to establish models at intermediate levels of aggregation (e.g., by sex/age groups within areas). Area-level models only use aggregated data for estimation areas or domains. This type of data can, typically, be obtained with fewer restrictions since aggregation avoids confidentiality issues. The so-called Fay-Herriot (FH) models, advanced by Fay and Herriot (1979), are very widely used linear models at the area level. These models have a two-level structure. At the first level, the relationship between the indicators of interest for the areas and the available area-level auxiliary variables is considered to be constant for all areas. For example, the decrease in average earnings from being employed to being unemployed is considered, all other things equal, to be the same in all areas. Thus, all areas are connected through a linear regression model. At the second level, it is assumed that, given the true values of the indicators of interest, the direct estimators of the areas are centred on these true values and have variances that are assumed to be known. Such variances typically vary between areas due to the fact that the sample sizes of the areas are different. These models have been deservedly successful because the resulting estimators for the areas are a composite or weighted average between the direct estimators and the synthetic regression estimators. When the synthetic model does not fit the data well (i.e., the considered auxiliary variables do not sufficiently explain the heterogeneity of the indicator across the areas) or the sample size of an area is large, the FH model-based estimator gives more weight to the direct estimator, which is sufficiently accurate. Conversely, when the synthetic model fits well or the area sample size is small (imprecise direct estimator), the weight given to the synthetic regression estimator increases. In this case, efficiency is increased due to the fact that the synthetic estimator has a regression coefficient that is common to all areas and is, therefore, estimated using data from all areas. Furthermore, since the direct estimators are approximately unbiased with regard to the sample design, for areas with larger sample sizes, the estimators obtained from the Fay-Herriot model also maintain a small bias under the design. One of the difficulties with these models is how to determine the values of the direct estimator variances (or heteroscedastic variances of the model error terms). Although, as stated above, these variances are assumed to be known, in practice they are replaced by estimates. Given the small amount of data in some of the areas, the estimates of these variances are also very imprecise. There are smoothing methods such as the generalised variance function method (see Fay and Herriot, 1979) or non-parametric estimation of these variances, (see González-Manteiga et al. (2010)). The estimation of these variances adds the problem of incorporating the estimation error of these variances into the error of the final estimator.

In individual-level models, as the name suggests, the model is established for each individual in the population (superpopulation model), and therefore the fit of these models requires individual data on the response variable and auxiliary variables. The first model of this type was advanced by Battesse, Harter and Fuller (1988) and is known as the nested error model. This is a linear regression. model that, in addition to the individual model errors, includes random effects associated with the areas, which illustrate the heterogeneity between the areas that is not explained by the available auxiliary variables. These models are currently widely used when the necessary data is available, as they incorporate much more information than area-level models and the variances of the model errors do not need to be known.

The adoption of a stochastic model that generates the values of the variable of interest in the individuals of the population makes the indicators of interest random quantities. Thus, in the literature, it is common to use the term "predict" rather than "estimate" the value of the indicator of interest and "predictor" rather than "estimator". In this paper, both terms will be used synonymously. In this context, an unbiased predictor of an indicator is one whose expectation under the model matches the expectation of that indicator. When estimating linear-type indicators in the values of the variable of interest in the individuals of the population, as means or totals, the basic models that are used at area or individual level are part of the mixed linear models that include random effects on the areas of interest. Under these models, the usual indirect estimator is the best linear unbiased predictor (BLUP), which consists of the linear combination of the observed values of the response variable in the individuals in the sample, which is unbiased under the model and minimises the mean squared error. The BLUP depends on the unknown parameters of the model, which represent the common behaviour between the areas. By replacing these unknown parameters with estimators, we obtain the empirical BLUP (EBLUP). This is eventually the usual model-based estimator (or predictor) of a linear indicator in a small area.

The BLUP does not require any assumption of normality in the model. However, to estimate more general indicators than the linear ones, the best predictor is the one that minimises the mean squared error, without requiring it to be linear or unbiased. This is equal to the expectation under the model of the indicator to be estimated, conditional on the values observed in the sample. Under normality, the best predictor of a linear indicator is the BLUP. When there is no normality or when the indicator to be estimated is not linear, it may be that the expectation that defines the best predictor cannot be calculated analytically. In that case, numerical approximations of the best predictor are used. Other widely used models, for example when estimating proportions of binary variables, are generalised linear models with random effects (see Chapter V).

Let us now consider a population that is divided into domains, and these domains are in turn divided into subdomains, and we wish to estimate at one or both levels. For example, Mexico is divided into 31 states plus Mexico City and each state, in turn, is divided into a number of municipalities. More appropriate models for this situation include random effects at various levels (see, for example, Stukel and Rao, 1999 for the estimation of linear indicators or Marhuenda et al., 2018, for the estimation of general indicators). On the other hand, when there are several interrelated variables of interest, multivariate models can be used (see Fay, 1987 or Datta, Fay and Ghosh, 1991). Also, when there is temporal and/or spatial correlation, one can resort to models that include random effects that follow a temporal series process and/or a spatial process (see, for example, Pfeffermann and Burk (1990) or Rao and Yu (1992) for temporal models, Molina, Salvati and Pratesi (2008) for a spatial model and Marhuenda, Molina and Morales (2013) for a spatio-temporal model). On the other hand, Bayesian models are an alternative to frequentist models that more often than not present computational advantages, providing estimates that are practically identical to those obtained with the corresponding frequentist model as long as the prior distributions considered are non-informative (see Chapter IV). The case study by Rao and Molina (2015) gives a detailed overview of the most widely used techniques in small area estimation and carries out a thorough review of most of the work done in this field up to the date of publication.

II. Common indicators of poverty and inequality

In the literature there are countless indicators of poverty and inequality that summarise different aspects of the living conditions of a population. Indeed, based on official surveys of living conditions in various countries, the National Statistical Institutes usually produce a wide variety of indicators in order to illustrate the different measurements of poverty or inequality. The mathematical form of the particular indicator to be estimated is very important when it comes to selecting appropriate small-area estimation techniques, as not all techniques are applicable to all types of estimators.

In this chapter we will review many of the indicators that appear in the literature, as well as the indicators that are usually produced from official surveys of living conditions. Although it is not possible to include all existing indicators, some of those described in this chapter will be used to illustrate the small-area estimation techniques most commonly used for our purposes. Thus, the following chapters will review the different methods that can be used, with an indication of the types of indicators to which they are applicable.

Neri, Ballini and Betti (2005) review indicators of poverty and inequality. The most widely used poverty indicator is the poverty incidence or poverty rate, also called the at-risk-of-poverty rate, which is calculated as the proportion of individuals with (equivalised net) income below the poverty line. Another common indicator is the poverty gap, which measures the extent of poverty rather than the frequency of individuals at risk of poverty. Both these indicators are elements of a wider. family of indicators defined by Foster, Greer and Thorbecke (1984), which we will call the FGT family of indicators, and which have the advantage of being additive in individuals. The small area estimation methods which we will describe in later chapters will be illustrated by applying them to some of the indicators in this family, although it is important to note that some methods are applicable to many other indicators not included in this family. In each chapter, it will be made clear to which indicators each method is applicable.

Let U denotes the target population (e.g., the residents of a country), of size N, which is divided into D subpopulations, the areas, or domains to be estimated, of sizes $N_1, \dots, N_D$. Note that the population sizes of the areas are usually very large because, as discussed in Chapter II, the term "small area" refers to the sample size (more specifically to the sampling error of the direct estimator used) and not to the population size.

E_{di} denotes the measure of purchasing power (e.g., income or expenditure) of the individual i in the area d, $d = 1, \dots, D$. z denotes the poverty line used, below which an individual is considered to be at risk of poverty. The FGT indicator family for the area d is defined by:

$$F_{\alpha d} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{z - E_{di}}{z} \right)^{\alpha} I(E_{di} < z), \quad d = 1, \dots, D, \alpha \geq 0, \tag{1}$$

where $I(E_{di} < z)$ is an indicator function, which assumes the value 1 if $E_{di} < z$ (individual i at risk of poverty) or the value 0 otherwise. Assuming $\alpha = 0$, we obtain the poverty rate or incidence. The poverty gap is the indicator obtained assuming $\alpha = 1$.

A more complex indicator that uses both the poverty gap and poverty incidence, in addition to the Gini coefficient, is the Sen Index (Sen, 1976). On the other hand, within the indicators that do not depend on a poverty threshold but on the relative situation of individuals within the overall ranking, we can mention the Fuzzy monetary index and the Fuzzy supplementary index (see Betti et al. (2006)). Beyond the monetary dimension, it is often interesting to measure other types of constraints or deprivations that are not strictly monetary. These deprivations are usually measured as proportions of individuals who have (or do not have) access to certain services such as healthcare, housing, and education. On the other hand, indicators of inequality include the Gini Index, the generalised entropy index, or the Theil Index (see e.g., Neri, Ballini and Betti (2005)).

At the European Council of December 2001, as part of the Lisbon Strategy of 2000 for the coordination of social policies of the member states, a set of indicators of poverty and social exclusion, known as the Laeken indicators, were established. These indicators include the at-risk-of-poverty rate F_{0d}, the quintile share ratio (the ratio between the incomes of the richest 20% of the population and the poorest 20%), the relative median poverty risk gap and the Gini Index, among others.

An example of multidimensional poverty measurement is the one used by the CONEVAL (National Council for the Evaluation of Social Development Policy) in Mexico, known as the multidimensional poverty indicator, which measures the proportion of individuals with at least one from among a set of established disadvantages or deprivations, and whose income is below the welfare threshold or line. The following chapters will review some small area estimation methods which, although illustrated by estimating indicators of the FGT family, can be used in the same way to estimate a large number of indicators.

III. Direct methods for the disaggregation of poverty data

This chapter gives an overview of basic direct estimators for the mean of a variable in a domain or area, expressed as

$$\bar{Y}_d = N_d^{-1} \sum_{i=1}^{N_d} Y_{di}, \tag{2}$$

where Y_{di} denotes the value of the variable for the individual i within the area (or domain) d. Note that the FGT indicators given in (1) can also be written in the form of mean values as in (2) calling

$$F_{\alpha,di} = \left(\frac{z - E_{di}}{z}\right)^{\alpha} I(E_{di} < z),$$

whereby we obtain that $F_{\alpha d}$ is the mean of the values $Y_{di} = F_{\alpha,di}$ for the individuals in the area d, or in other words,

$$F_{\alpha d} = N_d^{-1} \sum_{i=1}^{N_d} F_{\alpha,di}. \tag{3}$$

As stated above, an estimator of an indicator in a given area qualifies as "direct" if it is calculated using only data from that area and without making use of data from any other area. These estimators are the default estimators used by National Statistical Institutes, due to their good sample design properties (such as unbiasedness) in areas with sufficient sample size. For example, direct estimators have traditionally been used to produce statistics on living conditions in Chile at national and regional levels and for a set of comunas with a representative sample, according to the Chilean National Socio-economic Characterisation Survey (Encuesta de Caracterización Socioeconómica Nacional or CASEN). From CASEN 2015 onwards, the methodology for estimation in non-representative comunas is carried out using indirect model-based methods; specifically, the Fay-Herriot method described in the introduction (see the paper on methodology for estimating poverty at the comuna level, with data from Casen 2015 from Chile's 2017 Social Observatory of the Ministry of Social Development).

In this paper, s denotes the sample of size n drawn from the population U, s_d the subsample of the area d of size n_d (which may be equal to zero) and r_d the set of out-of-sample elements from the same area, $d = 1, \dots, D$, where $\sum_{d=1}^{D} n_d = n$. Furthermore, π_{di} denotes the probability of inclusion of the individual i in the sample of the area d, $w_{di} = \pi_{di}^{-1}$ denotes the sampling weight of the same individual and $\pi_{d,ij}$ denotes the probability of inclusion of the individuals i and j in the sample of the area d. We will now give an overview of the best-known direct estimators.

A. Basic direct estimators

The unbiased estimator with respect to the sample design of the mean of the area d, $\bar{Y}_d$, is known as the Horvitz-Thompson (HT) estimator. This estimator needs to know the true size of the area N_d and the sampling weights $w_{di} = \pi_{di}^{-1}$ for the sample individuals in the area d. Assuming that these are known, the HT estimator of $\bar{Y}_d$ is

$$\hat{\bar{Y}}_d = N_d^{-1} \sum_{i \in s_d} w_{di} \, Y_{di}. \qquad (4)$$

Note that for the total area d, $Y_d = \sum_{i=1}^{N_d} Y_{di}$, the HT estimator is simply $\hat{Y}_d = \sum_{i \in s_d} w_{di} \, Y_{di}$ and does not need to know the area's population size N_d.

If $\pi_{di} > 0$ for every $i = 1, \dots, N_d$, an unbiased estimator of the variance under the HT estimator design of $\bar{Y}_d$ is expressed as

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d) = N_d^{-2} \left\{ \sum_{i \in s_d} \frac{Y_{di}^2}{\pi_{di}^2}(1 - \pi_{di}) + 2 \sum_{i \in s_d} \sum_{\substack{j \in s_d \\ j > i}} \frac{Y_{di} Y_{dj}}{\pi_{di} \pi_{dj}} \left(\frac{\pi_{d,ij} - \pi_{di}\pi_{dj}}{\pi_{d,ij}} \right) \right\}. \qquad (5)$$

It often happens that, at the estimation stage, not all information about the sample design is available apart from the sampling weights w_{di}. Since the second-order inclusion probabilities $\pi_{d,ij}$ are not available, the estimator (5) cannot be calculated. However, for sample designs with second-order inclusion probabilities verifying $\pi_{d,ij} \approx \pi_{di}\pi_{dj}$, for $j \neq i$, as for example in Poisson sampling, where equality is given, the second term of (5) becomes approximately zero. Moreover, by replacing $w_{di} = \pi_{di}^{-1}$, we obtain the following variance estimator, which does not depend on second-order inclusion probabilities

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d) = N_d^{-2} \sum_{i \in s_d} w_{di} \, (w_{di} - 1) Y_{di}^2. \qquad (6)$$

This estimator is provided by the direct() function of the R sae package, which will be used in example 1 to illustrate these procedures, when the sampling weights are included. This function assumes that no information about the sample design is available other than the sampling weights. If we have information about the sample design, there are more suitable R packages such as survey (Lumley 2017) or sampling (Tillé and Matei 2016). In addition, there are alternative approximations of variance depending on the sample design and available information, e.g., the ultimate clusters method or the *Balanced Repeated Replications* (BRR) method with Fay's correction (U.S. Bureau of Labor Statistics and U.S. Census Bureau 2006).

The HT estimator weights the individual observations Y_{di} by using the sampling weights or inverses of the sample inclusion probabilities, $w_{di} = \pi_{di}^{-1}$. This protects against situations where the probability of selecting an individual is related to the value of the variable of interest (informative sample design). Indeed, if certain types of individuals (e.g., those with lower incomes) have a higher probability of appearing in the sample, it is likely that these types of individuals will appear more frequently in the final sample, while the types of individuals less likely to appear (e.g., those with higher incomes) are likely to be scarce in the sample. This means that, if we were to estimate by giving the same weight to all the observations in the sample, as in the basic sample mean, we would have a bias (e.g., average income would be underestimated). For this reason, less weight must be given to those observations which are more likely to appear in the sample, and more weight to those which are less likely to appear.

Although this estimator is exactly unbiased with respect to the sample design, its variance under the design can be very large when the sample size of the area n_d is small. A slightly biased estimator for small n_d but with a somewhat smaller variance, and which does not need to know the size of the area N_d in order to estimate the mean $\bar{Y}_d$, is the Hájek estimator. This estimator is equal to the weighted mean in the observations in the area, using the sampling weights as weightings, i.e,

$$\hat{\bar{Y}}_d^{HA} = \hat{N}_d^{-1} \sum_{i \in s_d} w_{di}\, Y_{di}, \text{donde } \hat{N}_d = \sum_{i \in s_d} w_{di}.$$

For the total $Y_d = \sum_{i=1}^{N_d} Y_{di}$, the Hájek estimator is $\hat{Y}_d^{HA} = N_d \hat{\bar{Y}}_d^{HA}$, which does need to know the population size N_d.

Under the sample design, an estimator of the variance of the Hájek estimator, $\hat{\bar{Y}}_d^{HA}$, is obtained using the Taylor linearisation method. The resulting estimator is obtained by simply replacing Y_{di} by $\tilde{e}_{di} = Y_{di} - \hat{\bar{Y}}_d^{HA}$ in the variance estimator of the HT estimator of the total $\hat{Y}_d$ and dividing by $\hat{N}_d$; i.e.

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d) = \hat{N}_d^{-2} \left\{ \sum_{i \in s_d} \frac{(Y_{di} - \hat{\bar{Y}}_d^{HA})^2}{\pi_{di}^2} (1 - \pi_{di}) \right.$$

$$\left. + 2 \sum_{i \in s_d} \sum_{\substack{j \in s_d \\ j > i}} \frac{(Y_{di} - \hat{\bar{Y}}_d^{HA})(Y_{di} - \hat{\bar{Y}}_d^{HA})}{\pi_{di}\pi_{dj}} \left(\frac{\pi_{d,ij} - \pi_{di}\pi_{dj}}{\pi_{d,ij}} \right) \right\}, \quad (7)$$

assuming that $\pi_{di} > 0$, for each i. For designs in which $\pi_{d,ij} \approx \pi_{di}\pi_{dj}$, for $j \neq i$, as in Poisson sampling, this estimated variance is reduced to

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d) = \hat{N}_d^{-2} \sum_{i \in s_d} w_{di}\, (w_{di} - 1)(Y_{di} - \hat{\bar{Y}}_d^{HA})^2.$$

As stated above, the FGT indicators have the advantage that they can be written as a mean for the individuals in the area (see (3)). Therefore, the Horvitz-Thompson estimator of $F_{\alpha d}$ is then

$$\hat{F}_{\alpha d} = N_d^{-1} \sum_{i \in s_d} w_{di}\, F_{\alpha,di}.$$

Alternatively, the Hájek estimator of $F_{\alpha d}$ is expressed as

$$\hat{F}_{\alpha d}^{HA} = \hat{N}_d^{-1} \sum_{i \in s_d} w_{di}\, F_{\alpha,di}.$$

Note that, by aggregating the direct HT estimators of the totals Y_d for the areas of a larger region, say for the entire population, we obtain the HT estimator of the population total $\hat{Y} = \sum_{d=1}^{D} \sum_{i \in s_d} w_{di}\, Y_{di}$, i.e.

$$\sum_{d=1}^{D} \hat{Y}_d = \hat{Y}.$$

Given that the HT estimator is efficient at a higher level of aggregation such as the population level, this property, known as the benchmarking property, is recommended for estimators in the areas. However, other estimators, especially the indirect estimators that we will see in the following chapters, will not add up exactly to the direct estimator considered for the population total (which may be different from that of HT). Adjustments can be made to the estimators to force this to happen. Let $\hat{Y}_d^{EST}$ be an estimator that does not verify this property. If we want these to aggregate the HT estimator at the national level $\hat{Y}$, a common adjustment is of the ratio type, expressed as

$$\hat{Y}_d^{AEST} = \hat{Y}_d^{EST} \frac{\hat{Y}}{\sum_{d=1}^{D} \hat{Y}_d^{EST}}, \quad d = 1, \ldots, D.$$

There is a large body of literature on other types of adjustments, such as difference adjustments, and on methods specifically designed to compel computed estimators to check this property even at various levels, but they are not included in this paper for purposes of conciseness. For more information, see e.g., Ghosh and Steorts (2013) and the references quoted therein.

Below, we summarise the types of indicators to which these estimators are applicable, the data that is necessary to produce them as well as the data of the variable of interest obtained from a survey, and the advantages and disadvantages from an eminently practical point of view.

Objective indicators: additive parameters, in that they are sums of certain variables for each individual in the area. These variables can be functions of the variables of interest for the individuals (e.g., $F_{\alpha,di}$ is a function of the variable used to measure the purchasing power of the individual, E_{di}).

Data requirements:

- Sampling weights w_{di} for sample individuals in the area d.
- For the HT estimator of the mean and for the Hájek estimator of the total, area population size, N_d.

Advantages:

- The HT estimator is exactly unbiased and the Hájek estimator is approximately unbiased with respect to the sample design. Both are consistent with respect to the design when the sample size of the area n_d increases. Therefore, they perform well for areas with sufficient sample size under sample designs with unequal probabilities, including under information sampling, as long as they are calculated using the true probabilities of inclusion of individuals in the area sample.
- They do not need to assume any models or hypotheses about the variables in question Y_{di} which means they are completely non-parametric.
- They satisfy the benchmarking property: if we add up the estimated totals for all the areas in a larger region, we get the estimated total for that region which is obtained by the same method.

Disadvantages:

- They are very inefficient (i.e., they have a high sampling error) for small areas due to the small sample size.

- They cannot be calculated for unsampled areas or domains, i.e., with sample size n_d equal to zero.

Example 4.1. Direct HT estimators of poverty incidence, with R. We will demonstrate how to calculate direct HT estimators for poverty incidence, using simulated data for living conditions in Spanish provinces, included in the R data file known as incomedata from the R sae package. This dataset includes, for $n = 17119$ fictitious individuals living in the $D = 52$ Spanish provinces, the name of the province where they live (provlab), the province code (prov), the autonomous community code (ac), the age group from 1 to 5 (age), the nationality (nat, 1=if Spanish, 2=if not), the educational level (educ, from 0=under 16 to 2=third level), employment status (labor, where 0=under 16, 1=employed, 2=unemployed and 3=inactive), whether they are in each age group, from group 2 to 5 (age2 to age5), whether they have educational level 1 to 3 (educ1 to educ3), whether they have Spanish nationality, whether they are employed, unemployed or inactive, their equivalised net income (income) and the sampling weight (weight). We calculate direct HT estimators for the poverty incidence in the $D = 52$ Spanish provinces.

After installing the sae library, we load it, along with the incomedata dataset, which contains the sample data, and the sizeprov dataset, which contains the population sizes for the provinces, $\mathrm{N_d}$:

```
library(sae)
data(incomedata)
attach(incomedata)
data(sizeprov)
```

Next, we use the direct() function to obtain the direct HT estimators. First of all, we calculate the total sample size, the number of provinces and their sample sizes and extract the population sizes from the sizeprov file:

```
n<-dim(incomedata)[1]          # Total sample size
D<-length(unique(prov))        # Number of provinces (areas or domains)
nd<-as.vector(table(prov))     # Sample sizes of provinces
Nd<-sizeprov$Nd                # Population sizes of the provinces
```

We set the poverty line, which is calculated as 0.6*median (income) with the previous year's data, and construct the poor variable, which is the indicator of having income below the poverty line:

```
z<-6557.143
poor<-numeric(n)
poor[income<z]<-1
```

Finally, we calculate the direct HT estimators of the poverty incidence in the provinces (averages of the poor variable in the provinces), using the direct() function including the sampling weights given by the weight variable:

```
povinc.dir.res<-direct(y=poor,dom=prov,sweight=weight,domsize=sizeprov[,-1])
print(povinc.dir.res,row.names=F)
```

The output of this function is:

```
Domain SampSize     Direct         SD       CV
     1       96 0.25503732 0.04846645 19.003670
     2      173 0.14059242 0.03042195 21.638397
     3      539 0.20785096 0.02178689 10.481979
     4      198 0.26763976 0.04090335 15.282986
     5       58 0.05512200 0.02555426 46.359465
```

```
 6      494 0.21553890 0.02357906 10.939585
 7      634 0.09999792 0.01536517 15.365488
 8     1420 0.29812535 0.01618508  5.428952
 9      168 0.21413150 0.04473542 20.891562
10      282 0.27031324 0.03125819 11.563692
11      398 0.14887351 0.02189022 14.703904
12      118 0.17598199 0.03584882 20.370731
13      250 0.20921534 0.03279230 15.673948
14      224 0.29975708 0.03934080 13.124228
15      495 0.25347550 0.02467716  9.735520
16       92 0.26334059 0.05913385 22.455274
17      142 0.18337421 0.03710194 20.232911
18      208 0.31727340 0.04043964 12.745990
19       89 0.17908182 0.04234025 23.642966
20      285 0.23690549 0.03194779 13.485457
21      122 0.12583449 0.03202547 25.450474
22      115 0.24107606 0.04856351 20.144476
23      232 0.31294198 0.04122671 13.173916
24      218 0.18801572 0.03002634 15.970122
25      130 0.15559590 0.03872448 24.887854
26      510 0.25811811 0.02459196  9.527405
27      173 0.37718722 0.05696330 15.102129
28      944 0.18218209 0.01639018  8.996593
29      379 0.22918462 0.02735631 11.936364
30      885 0.17703167 0.01648910  9.314210
31      564 0.16190765 0.01842017 11.376958
32      129 0.22799612 0.04199465 18.419018
33      803 0.26064010 0.02093779  8.033220
34       72 0.30166074 0.07179782 23.800849
35      472 0.16651843 0.02307258 13.855869
36      448 0.18549072 0.02418887 13.040474
37      164 0.16104513 0.02998243 18.617410
38      381 0.18429619 0.02054550 11.148085
39      434 0.34244429 0.03248937  9.487491
40       58 0.22262002 0.05639965 25.334492
41      482 0.20503036 0.02122527 10.352256
42       20 0.02541207 0.02540651 99.978151
43      134 0.32035438 0.04934077 15.401934
44       72 0.27364239 0.06723440 24.570172
45      275 0.12553377 0.02131991 16.983409
46      714 0.21360678 0.02070508  9.693081
47      299 0.19292332 0.03211484 16.646429
48      524 0.21694466 0.02215645 10.212948
```

```
   49       104 0.30027442 0.06025302 20.065986
   50       564 0.10034577 0.01569138 15.637311
   51       235 0.19724796 0.03341193 16.939048
   52       180 0.19109119 0.03441016 18.007191
```

Finally, we store the estimated values in a vector and count how many provinces have a CV above 20%:

```
povinc.dir<-povinc.dir.res$Direct
povinc.dir.cv<-povinc.dir.res$CV
sum(povinc.dir.cv>20)
```

There are 15 provinces whose direct HT estimators of poverty incidence have a CV greater than 20%. Those 15 provinces would be small areas for this indicator. But, as we will see, more efficient estimators can also be found in other provinces.

B. GREG and calibration estimators

A more sophisticated estimator than the basic direct estimators described in the previous chapter, in that it uses auxiliary information, is the generalised regression estimator (GREG). This estimator requires knowing the total $\boldsymbol{X}_d = \sum_{i=1}^{N_d} \boldsymbol{x}_{di}$, or the mean $\bar{\boldsymbol{X}}_d = N_d^{-1} \sum_{i=1}^{N_d} \boldsymbol{x}_{di}$, for the area d of a vector $\boldsymbol{x}_{di}$ of values of p auxiliary variables related to Y_{di}, for the individual i within the area d. If $\hat{\bar{\boldsymbol{X}}}_d = N_d^{-1} \sum_{i \in s_d} w_{di}\, \boldsymbol{x}_{di}$ is the HT estimator of $\bar{\boldsymbol{X}}_d$, the GREG estimator of $\bar{Y}_d$ is expressed as

$$\hat{\bar{Y}}_d^{GREG} = \hat{\bar{Y}}_d + \left(\bar{\boldsymbol{X}}_d - \hat{\bar{\boldsymbol{X}}}_d\right)' \widehat{\boldsymbol{B}}_d. \qquad (8)$$

Here, $\widehat{\boldsymbol{B}}_d = \left(\sum_{i \in s_d} w_{di}\, \boldsymbol{x}_{di}\boldsymbol{x}_{di}'/c_{di}\right)^{-1} \sum_{i \in s_d} w_{di}\, \boldsymbol{x}_{di} Y_{di}/c_{di}$ is the weighted least squares estimator (using the sample design weights) of the vector of coefficients of the following linear regression assumed for the units of the area d,

$$Y_{di} = \boldsymbol{x}_{di}'\boldsymbol{\beta}_d + \epsilon_{di}, \quad i = 1, \dots, N_d, \qquad (9)$$

where the model errors ϵ_{di} are independent, with zero expectation and variance $\sigma^2 c_{di}$, being $c_{di} > 0$ constants representing the possible heteroscedasticity, $i = 1, \dots, N_d$. The constants c_{di} are determined by studying the residuals of the linear model without heteroscedasticity, i.e., with $c_{di} = 1$, $i = 1, \dots, N_d$. For example, by looking at the scatter plot of the residuals against each of the auxiliary variables, we can observe graphically whether the variance of the residuals increases with any of them. In this case, we would take as constants c_{di}, the values of this variable in the units of the area or, more generally, a function $c_{di} = f(x_{di}) > 0$, of the values of this auxiliary variable.

The GREG estimator of the mean of the area d, $\bar{Y}_d$, is approximately unbiased under the sample design regardless of whether model (9) is correct or not, since the bias of the estimator of the regression coefficient vector $\widehat{\boldsymbol{B}}_d$, as an estimator of its population version, $\boldsymbol{B}_d = (\sum_{i=1}^{N_d} \boldsymbol{x}_{di}\, \boldsymbol{x}_{di}'/ c_{di})^{-1} \sum_{i=1}^{N_d} \boldsymbol{x}_{di}\, Y_{di}/c_{di}$, is small. Thus, model (9) is often called a working model and estimators that are unbiased regardless of whether the model is verified, like (8), are called model-assisted. On the other hand, the GREG is also unbiased under the regression model (9), conditionally on the sample s. Although the GREG estimator tends to improve the efficiency of the direct estimator $\hat{\bar{Y}}_d$ if the auxiliary variables are linearly related to the dependent variable Y_{di}, this estimator only uses data from the area d and, therefore, its variance may still be large for areas with a small sample size n_d.

Note that, if we wish to use the GREG estimator for the FGT indicator of order α, which is equal to the mean of the variables $F_{\alpha,di}$ in the area, i.e. $F_{\alpha d} = N_d^{-1}\sum_{i=1}^{N_d} F_{\alpha,di}$, the improvement in efficiency with respect to the direct estimator would depend on the goodness of fit of the following regression model:

$$F_{\alpha,di} = \boldsymbol{x}_{di}'\boldsymbol{\beta}_d + \epsilon_{di}, \quad i = 1, \dots, N_d.$$

However, in the case of FGT indicators, the variables $F_{\alpha,di}$ are a complex function of the variable of interest (the measurement of purchasing power E_{di}) expressed as $F_{\alpha,di} = \{(z - E_{di})/z\}^{\alpha} I(E_{di} < z)$, $\alpha \geq 0$. It is not easy to find auxiliary variables $\boldsymbol{x}_{di}$ that are linearly related to $F_{\alpha,di}$. Therefore, this model is difficult to verify in practice and, thus, for FGT indicators, GREG estimators are of less use than for estimating the means or totals of the variables of interest (e.g., income E_{di}).

Calibration estimators are widely used in National Statistical Institutes to estimate means or totals at national level and in regions with sufficient sample size. If we calibrate at the area level, we will see that the resulting estimator is closely related to the GREG estimator. The calibration method was proposed by Deville and Särndal (1992) for estimating the total of a variable of interest using auxiliary information from p related variables. Assuming that we know the totals of the auxiliary variables in the area, $\boldsymbol{X}_d$, and also assuming that the auxiliary variables $\boldsymbol{x}_{di}$ are linearly related to Y_{di}, the calibration method consists of finding new weights h_{di}, as close as possible to the original sampling weights w_{di} in accordance with a distance $G_{di}(h_{di}, w_{di})$, such that the total $\boldsymbol{X}_d$ of the auxiliary variables is estimated exactly with these weights; i.e. without error. If the variable of interest is linearly related to these auxiliary variables and the totals of these auxiliary variables are estimated exactly, it is expected that the totals of the variable of interest will also be estimated with little error. In formal terms, when estimating the mean $\bar{Y}_d$, we look for new weights for the sample units, h_{di}, $i \in s_d$, which are the solution to the problem

$$\begin{array}{ll} \min\limits_{\{h_{di}; i\in s_d\}} & \sum\limits_{i\in s_d} G_{di}\,(h_{di}, w_{di}) \\ \text{sujeto a} & \sum\limits_{i\in s_d} h_{di}\,\boldsymbol{x}_{di} = \boldsymbol{X}_d, \end{array}$$

where $G_{di}(\cdot,\cdot)$ is a pseudo-distance. Using the pseudo chi-squared distance expressed as $G_{di}(h_{di}, w_{di}) = c_{di}(h_{di} - w_{di})^2/w_{di}$, which is probably the most popular, and solving the problem by use of the Lagrange multiplier method, the resulting weights are

$$h_{di} = w_{di}\left\{1 + \boldsymbol{x}_{di}'\left(\sum_{i\in s_d} w_{di}\,\boldsymbol{x}_{di}\boldsymbol{x}_{di}'/c_{di}\right)^{-1}\left(\boldsymbol{X}_d - \sum_{i\in s_d} w_{di}\,\boldsymbol{x}_{di}/c_{di}\right)\right\}, i \in s_d. \quad (10)$$

Note that the calibrated weights h_{di} are the result of making an adjustment to the original weights, $h_{di} = w_{di}g_{di}$, where the adjustment factor g_{di} is expressed as the term within the brackets in (10). The calibration estimator of $\bar{Y}_d$ is then obtained simply the same as with the HT estimator, but using the calibrated weights instead of the original ones, as follows:

$$\hat{\bar{Y}}_d^{CAL} = N_d^{-1}\sum_{i\in s_d} h_{di}\,Y_{di}.$$

It is easy to demonstrate that, by substituting the formula obtained for these weights given in (10) in the calibration estimator $\hat{\bar{Y}}_d^{CAL}$, we obtain exactly the GREG estimator of $\bar{Y}_d$ given in (8). Deville and Särndal (1992) propose calibration estimators based on $G_{di}(\cdot,\cdot)$ distances rather than the chi-squared distance. However, they also show that the resulting estimators, under certain regularity conditions for the distance $G_{di}(\cdot,\cdot)$, are asymptotically equivalent to the GREG and thus share the same asymptotic variance. As with the GREG estimator, for a small sample size n_d, the variance of the calibration estimators can be large.

A consistent estimator (when n_d increases) for the variance of the estimator $\hat{\bar{Y}}_d^{GREG}$ is obtained using the Taylor linearisation method. The resulting estimator is derived from replacing Y_{di} by $\tilde{e}_{di} = Y_{di} - \boldsymbol{x}_{di}{}'\widehat{\boldsymbol{B}}_d$ in the estimated variance of the HT estimator given in (5), i.e.

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d^{GREG}) = N_d^{-2}\left\{\sum_{i\in s_d}\frac{\tilde{e}_{di}^2}{\pi_{di}^2}(1-\pi_{di}) + 2\sum_{i\in s_d}\sum_{\substack{j\in s_d\\ j>i}}\frac{\tilde{e}_{di}\tilde{e}_{di}}{\pi_{di}\pi_{di}}\left(\frac{\pi_{d,ij}-\pi_{di}\pi_{di}}{\pi_{d,ij}}\right)\right\}.$$

For designs in which $\pi_{d,ij} \approx \pi_{di}\pi_{di}$ is verified, for $j \neq i$, as in Poisson sampling, this estimated variance, written as a function of $w_{di} = \pi_{di}^{-1}$, reduces to

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d^{GREG}) = N_d^{-2}\sum_{i\in s_d} w_{di}\,(w_{di}-1)\tilde{e}_{di}^2.$$

Simulation studies have shown that this estimator may underestimate the variance of the GREG. However, the estimator that results from replacing Y_{di} by $g_{di}\tilde{e}_{di}$, where g_{di} is the adjustment factor for the weights w_{di}, in the estimated variance of the HT estimator, expressed as

$$\widehat{\mathrm{var}}_\pi(\hat{\bar{Y}}_d^{GREG}) = N_d^{-2}\left\{\sum_{i\in s_d}\frac{g_{di}^2\tilde{e}_{di}^2}{\pi_{di}^2}(1-\pi_{di}) + 2\sum_{i\in s_d}\sum_{\substack{j\in s_d\\ j>i}}\frac{g_{di}\tilde{e}_{di}g_{dj}\tilde{e}_{dj}}{\pi_{di}\pi_{di}}\left(\frac{\pi_{d,ij}-\pi_{di}\pi_{di}}{\pi_{d,ij}}\right)\right\}.$$

reduces this underestimation and remains consistent when n_d increases (see Fuller (1975) or Estevao, Hidiroglou and Särndal (1995)). Moreover, such an alternative variance estimator is approximately unbiased for the variance of the GREG $\hat{\bar{Y}}_d^{GREG}$ under the model (9) conditionally on the sample s, for various sample designs.

Again, note that these estimators work for estimating totals or means of the variables of interest, but do not work for other types of parameters. For example, for the FGT indicator of order α in the area d, $F_{\alpha d} = N_d^{-1}\sum_{i=1}^{N_d} F_{\alpha,di}$, the GREG or calibration estimators would be more efficient with respect to the direct estimator if the auxiliary variables $\boldsymbol{x}_{di}$ were linearly related to $F_{\alpha,di}$,which is unlikely in practice.

The main features of these estimators are summarised below:

Target indicators: means/totals of the variables of interest.

Data requirements:

- Sampling weights w_{di} for sample individuals in the area d.
- For the estimator of the mean, population size of the area, N_d.
- Sample observations of the p auxiliary variables related to the variable of interest, obtained from the same survey from which the data of the variable of interest is obtained.
- Population totals $\boldsymbol{X}_d$ or means $\bar{\boldsymbol{X}}_d$ of the p auxiliary variables in the area.

Advantages:

- They are approximately unbiased (and consistent when n_d increases) with respect to the sample design, regardless of whether the model is verified or not. Therefore, they perform well for areas with sufficient sample size under sample designs with unequal probabilities, including information sampling.

- They do not require the model under consideration to be verified for the variables of interest Y_{di}; i.e., they are non-parametric.

Disadvantages:

- Although they may improve the basic direct estimators if the regression model is verified, they may still be inefficient for small areas due to the small sample size.
- They cannot be calculated for unsampled areas or domains, i.e., with sample size n_d equal to zero.

Example 4.2. GREG estimators of poverty incidence, with R. Continuing Example 4.1, we now show how GREG estimators of poverty incidence in the provinces could be calculated with the same data, now considering auxiliary variables; specifically constant 1, age group, educational level, and employment status.

First of all, we load the files containing the missing data: the totals of individuals in each province for each age group, for each educational level and for each employment status:

```
data(sizeprovage)
data(sizeprovedu)
data(sizeprovlab)
```

We construct the matrix with the vectors of proportions of individuals in each category and province. These will form the vector of population means $\bar{\boldsymbol{X}}_d$:

```
Nd<-sizeprov[,3]
Ndage<-as.matrix(sizeprovage[,-c(1,2)])
Ndedu<-as.matrix(sizeprovedu[,-c(1,2)])
Ndlab<-as.matrix(sizeprovlab[,-c(1,2)])

Pdage<-Ndage/Nd
Pdedu<-Ndedu/Nd
Pdlab<-Ndlab/Nd

X<-cbind(const=rep(1,D),Pdage[,3:5],Pdedu[,c(2,4)],Pdlab[,2])
```

We next create the design matrix for the linear regression, with the values of the auxiliary variables for the individuals in the sample:

```
Xtot<-model.matrix(poor~age3+age4+age5+educ1+educ3+labor1)
```

Finally, we calculate the GREG estimators for the poverty incidence (mean values of the poor variable) in each province:

```
provl<-unique(prov)          # Index of each province
p<-dim(Xtot)[2]              # Number of auxiliary variables

betad<-matrix(0,nr=D,nc=p)   # Matrix with regression coefficients
                             # for each province (in rows)
Xd.est<-matrix(0,nr=D,nc=p)  # Matrix of direct estimators of the means
                             # of the auxiliary variables for each province
```

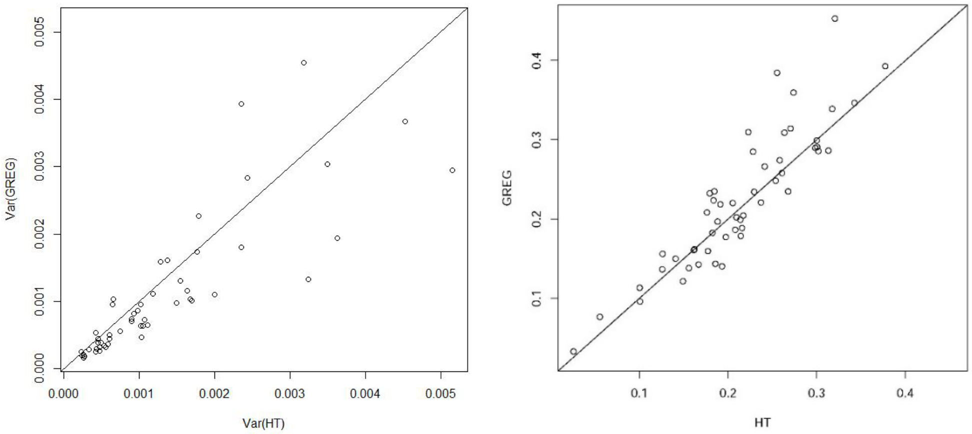
```
povinc.greg<-numeric(D)           # Vector of GREG estimators in the province
povinc.greg.var<-numeric(D)       # Vector with estimated  variances
                                  # under the design of the GREG estimators

for (d in 1:D){
  Xd<-Xtot[prov==provl[d],]       # Values of auxiliary variables
                                  # for the individuals in the province
  wd<-weight[prov==provl[d]]      # Sampling weights for the individuals
                                  # in the province
  yd<-poor[prov==provl[d]]        # Values of the variable of interest
                                  # for the individuals in the province

  # We adjust the regression for the province, with the sampling weights
  betad[d,]<-coef(summary(lm(yd~-1+Xd, weights=wd)))[,1]

  # Direct estimators of the means of the auxiliary variables in the province
  Xd.est[d,]<-colSums(diag(wd)%*%Xd)/Nd[d]

  # GREG estimator of the poverty incidence for the province
  povinc.greg[d]<-povinc.dir [d]+sum((X[d,]-Xd.est[d,])*betad[d,])

  # Estimated variance under the design of the Greg
  # estimator of the poverty incidence
  gd<-matrix(1/Nd[d]+
  +(X[d,]-Xd.est[d,])%*%solve(t(Xd)%*%diag(wd)%*%Xd)%*%t(Xd),nr=nd[d])
  ed<-yd-Xd%*%as.matrix(betad[d,],nr=p)
  povinc.greg.var[d]<-sum(wd*(wd-1)*(gd*ed)^2)
 }
# CVs of the GREG estimators
povinc.greg.cv<-100*sqrt(povinc.greg.var)/povinc.greg
```

We plot the values of the GREG estimators against those of HT, as well as their variances (or squared sampling errors):

```
M<-max(povinc.dir,povinc.greg)
m<-min(povinc.dir,povinc.greg)
plot(povinc.dir,povinc.greg,ylim=c(m,M),xlim=c(m,M),xlab="HT",ylab="GREG")
abline(a=0,b=1)
M<-max(povinc.dir.var,povinc.greg.var)
m<-min(povinc.dir.var,povinc.greg.var)
plot(povinc.dir.var, povinc.greg.var, ylim=c(m,M), xlim=c(m,M), xlab="Var(HT)", ylab="Var(GREG)")
abline(a=0,b=1)
```

We can see that the GREG estimators resemble the HT estimators, but their estimated variances are slightly smaller. This improved efficiency is achieved through the use of auxiliary information.

Figure 2
GREG estimators of the poverty incidence for the provinces versus HT estimators (left), and estimated variances of the GREG estimators versus HT estimators (right)
(In proportions)



Source: Prepared by the author.

IV. Basic indirect methods for the disaggregation of poverty data

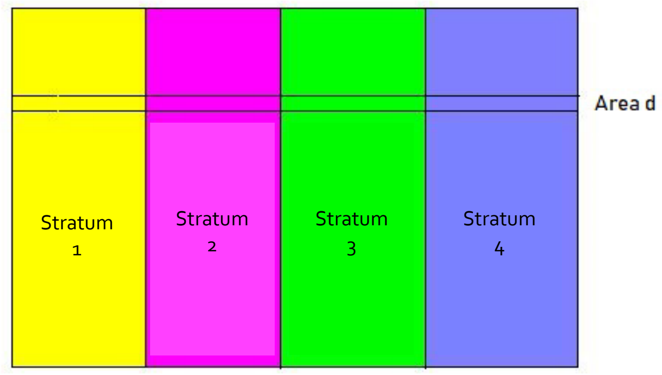
An indirect estimator for an indicator in a specific area is one that uses information from other areas by assuming some type of homogeneity between them. The use of a larger amount of information in the estimation process often leads to a decrease in the sampling error (or an increase in efficiency). Firstly, we will give an overview of synthetic estimators. A synthetic estimator is one that considers the areas to be homogeneous in that they have common parameters, without allowing any degree of heterogeneity between them. These estimators make strong assumptions that are unlikely in practice and therefore may have a large bias. Despite their potential bias, they are included in this paper for the purpose of demonstrating the intuitive idea underpinning small area estimation, which is to borrow information from other areas for the purpose of improving efficiency.

A. Synthetic post-stratified estimator

It must be emphasised once again that this estimator is rarely used in real small area estimation applications due to the fact that it is based on unrealistic assumptions; however, it is described in this paper as it provides a simple illustration of the main underlying idea of how to borrow information.

It has a qualitative variable related to the variable Y_{di}. This qualitative variable has J possible categories, which divide the population U into J groups, $U^1, \dots, U^J$ of sizes $N^1, \dots, N^J$, called post-strata, which intersect with the areas. Therefore, the area U_d of the population is equally divided into J post-strata pieces, $U_d^1, \dots, U_d^J$ of population sizes $N_d^1, \dots, N_d^J$ and with mean values $\bar{Y}_d^1, \dots, \bar{Y}_d^J$, where $\bar{Y}_d^j = \sum_{i \in U_d^j} Y_{di}/N_d^j$, $j = 1, \dots, J$ (see figure 3). For the sake of simplicity, we refer to the post-strata as strata in this figure and in the following.

Figure 3
Population divided into 4 post-strata and area d



Source: Prepared by the author.

Given that the means are additive indicators, we can decompose them into aggregations for the J strata, as follows:

$$\bar{Y}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} Y_{di} = \frac{1}{N_d}\sum_{j=1}^{J} N_d^j\ \bar{Y}_d^j. \tag{11}$$

It is assumed that individuals within each stratum behave homogeneously, regardless of the area to which they belong and, more specifically, it is assumed that

$$\bar{Y}_d^j = \bar{Y}^j, \quad j = 1, \dots, J, \tag{12}$$

where $\bar{Y}^j = \sum_{i \in U^j} Y_{di}/N^j$ is the mean of the stratum j. We can, then, take advantage of this homogeneity within strata to estimate the mean of each area by estimating the means of the strata (which must have large sample sizes). In other words, by substituting (12) in (11), we get

$$\bar{Y}_d = \frac{1}{N_d}\sum_{j=1}^{J} N_d^j\ \bar{Y}^j. \tag{13}$$

The synthetic post-stratified estimator (PS-SYN) of $\bar{Y}_d$ is obtained by estimating the means of each stratum in (13) by means of the Hájek's estimators:

$$\hat{\bar{Y}}_d^{PS-SYN} = \frac{1}{N_d}\sum_{j=1}^{J} N_d^j\ \hat{\bar{Y}}^{j,HA}.$$

The number of strata J is considered to be small, and to have a sufficient sample. Therefore, the direct estimators $\hat{\bar{Y}}^{j,HA}$ of the means in the strata $\bar{Y}^j$ have a small variance. This means that, when estimating the mean for the area d through the estimators for the strata, $\hat{\bar{Y}}^{j,HA}$, the variance is also small.

The homogeneity within each stratum is thus exploited to improve the efficiency of the estimator for the area d by using all the sample data. However, the assumption of homogeneity within each of the strata (12) is unrealistic and therefore the synthetic post-stratified estimator can have a considerable bias.

Since the bias of these estimators, rather than their variance, is not negligible and this can give a wrong picture of the quality of the estimator, it is worthwhile to obtain their mean squared error (MSE), which reflects both. For general synthetic estimators, an estimator of the MSE under the design is expressed as

$$\widehat{\mathrm{MSE}}_{\pi}(\hat{\bar{Y}}_d^{SYN}) = (\hat{\bar{Y}}_d^{SYN} - \hat{\bar{Y}}_d^{DIR})^2 - \widehat{\mathrm{var}}_{\pi}(\hat{\bar{Y}}_d^{DIR}),$$

(see Rao and Molina (2015), p.44, eq. (03/02/2016)). This estimator is very unstable as it depends on the direct estimator of the corresponding area. More stable MSE estimators have been proposed and they are based on the idea of averaging for all areas, but the resulting estimators are not area-specific; i.e., the same MSE value would be given for all areas. There are no known MSE estimators for synthetic estimators that are both stable and area-specific. This is a disadvantage of these estimators.

If we wish to use the PS-SYN estimator for an FGT indicator, it would in principle be possible, thanks to the additivity of these indicators. However, the estimator would be based on the (unrealistic) assumption that the FGT indicator remains constant within the strata, i.e.

$$F_{\alpha d}^{j} = F_{\alpha}^{j}, \quad j = 1, \dots, J,$$

if F_{α}^{j} is the FGT indicator in the stratum j. Therefore, this estimator would be more useful for estimating means or totals of a continuous variable.

These estimators can be summarised as follows:

Target indicators: means/totals of the variable of interest

Data requirements:

- Sampling weights w_{di} for all individuals in the sample.
- Area population size, N_d, and population sizes of the stratum-area intersection, $N_d^j, j = 1, \dots, J$.
- A qualitative variable (or a combination of several) observed in the same survey as, and related to, the variable of interest.

Advantages:

- If the strata have enough observations in the sample, the variance can be considerably decreased compared to a direct estimator.

Disadvantages:

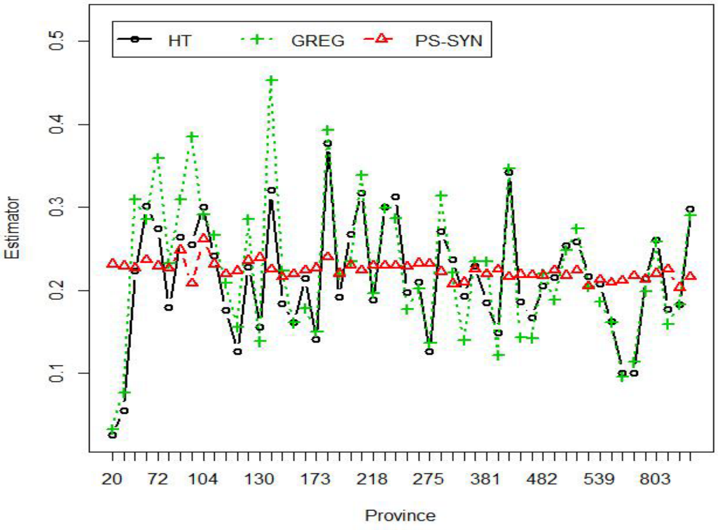
- In practice, the homogeneity hypothesis considered for the variables Y_{di} is unrealistic. If this is not verified, the resulting estimators may have considerable bias, and therefore may not reflect the facts. Besides, when sampling errors are estimated they will result in small values. However, it is rarely possible to estimate the bias adequately. Therefore, in the absence of accurate estimates of the bias, the estimators may appear to be of good quality, but this is very unlikely to be the case.
- It is not easy to find stable MSE estimators under the design.

Example 3. Synthetic post-stratified estimators of poverty incidence with R. Continuing Example 2, we can now demonstrate how to compute synthetic post-stratified estimators of poverty incidence for the provinces using educational levels (educ variable) as post-strata.

In Example 2 we had loaded the population sizes of the provinces for each educational level (sizeprovedu dataset). These sizes must be in a data frame object, where the column names must match the codes used for the categories of the post-stratum variable (educ). We therefore add the column names to the data frame with the population sizes. Next, we call on the pssynt() function, which calculates the post-stratified estimators for the poverty incidence (mean values of the poor variable) using the educ variable and we store the estimated values:

```
colnames(sizeprovedu) <- c("provlab","prov","0","1","2","3")
povinc.psedu.res<-pssynt(y=poor,sweight=weight,ps=educ,domsizebyps=sizeprovedu[,-1])
povinc.psedu<-povinc.psedu.res$PsSynthetic
```

Figure 4
HT, GREG and PS-SYN estimates of poverty incidence for each province
(In proportions)



Source: Prepared by the author.

Finally, we graphically compare the results with those obtained using the direct HT and GREG estimators for each province:

```
o<-order(nd)
M<-max(povinc.psedu,povinc.dir,povinc.greg)
m<-min(povinc.psedu,povinc.dir,povinc.greg)
k<-6
plot(1:D,povinc.dir[o],type="n",ylim=c(m,M+(M-m)/k),
     xlab="Province",ylab="Estimator",xaxt="n")
points(1:D,povinc.dir[o],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:D,povinc.greg[o],type="b",col=3,lty=3,pch=3,lwd=2)
points(1:D,povinc.psedu[o],type="b",col=2,lty=2,pch=2,lwd=2)
 axis(1, at=1:D, labels=nd[o])
legend(1,M+(M-m)/k,legend=c("HT","GREG","PS-SYN"),ncol=3,
     col=c(1,3,2),lwd=rep(2,3), lty=c(1,3,2),pch=c(1,3,2))
```

The results are shown in figure 4. We can see that the synthetic post-stratified estimators are too similar for all of the provinces, since they assume homogeneity for all individuals with the same educational level, regardless of the province to which they belong. This hypothesis is very unlikely.

B. Synthetic area-level regression estimator

Synthetic regression estimators assume a linear regression model that can be placed either at the area level or at the individual level, depending on the auxiliary information available. We begin by considering that auxiliary information is only available at the area level. $\boldsymbol{x}_d$ denotes the available vector of p auxiliary variables at the area level (e.g., the vector of means $\bar{\boldsymbol{X}}_d$ of p auxiliary variables). It is assumed that the indicator to be estimated δ_d (e.g., the area mean, $\delta_d = \bar{Y}_d$) constantly varies with respect to this aggregate data $\boldsymbol{x}_d$ for all areas, according to a linear regression model. Since the true values of the indicator in the areas are not available (they are the target parameters), direct estimators. $\hat{\delta}_d$, $d = 1, \dots, D$are considered instead. Thus, the model at the area level assumes that

$$\hat{\delta}_d = \boldsymbol{x}_d{}'\boldsymbol{\alpha} + \varepsilon_d, \quad d = 1, \dots, D, \tag{14}$$

where the error terms ε_d are assumed to be independent, with zero expectation and known variance ψ_d , $d = 1, \dots, D$. Note that since $\boldsymbol{x}_d$ is the population value and therefore has zero variance, ψ_d is the variance of the direct estimator $\hat{\delta}_d$, i.e., $\psi_d = \mathrm{var}(\hat{\delta}_d)$. In practice, these variances are estimated with microdata from the survey. The synthetic regression estimator (REG1-SYN) for the area indicator d is then expressed by the prediction of the indicator by means of the model, i.e. if $\hat{\boldsymbol{\alpha}} = (\sum_{d=1}^{D} \psi_d^{-1} \boldsymbol{x}_d \boldsymbol{x}_d{}')^{-1} \sum_{d=1}^{D} \psi_d^{-1} \boldsymbol{x}_d \hat{\delta}_d$ is the estimator of $\boldsymbol{\alpha}$ obtained by weighted least squares, the REG1-SYN estimator of δ_d is expressed as

$$\hat{\delta}_d^{REG1-SYN} = \boldsymbol{x}_d{}'\hat{\boldsymbol{\alpha}}.$$

In model (14), ε_d is the error due to the fact that we use a direct estimator $\hat{\delta}_d$ instead of the true value of the indicator δ_d, since this is unknown, and the true value δ_d is assumed to be exactly equal to the regression term, $\delta_d = \boldsymbol{x}_d{}'\boldsymbol{\alpha}$, leaving no degree of heterogeneity to the indicators of the different areas in respect of this regression. Model types which, like (14), do not incorporate area effects showing such heterogeneity, are called "synthetic models". In fact, the bias under the design of $\hat{\delta}_d^{REG1-SYN}$ for known $\boldsymbol{\alpha}$ is expressed as $\boldsymbol{x}_d{}'\boldsymbol{\alpha} - \delta_d$, which does not depend on the sample size of the area n_d; therefore, this bias does not decrease when the sample size of the area increases.

One advantage of the model-based estimators is that they allow estimation in unsampled areas; that is, with sample size equal to zero, if the corresponding auxiliary information is available. For an area d with $n_d = 0$, if we have $\boldsymbol{x}_d$, then the synthetic estimator of δ_d is likewise $\hat{\delta}_d^{REG1-SYN} = \boldsymbol{x}_d{}'\hat{\boldsymbol{\alpha}}$.

To estimate the FGT poverty indicator of order α, $\delta_d = F_{\alpha d}$, using this procedure, we need auxiliary area-level variables that verify the model at the area level

$$\hat{F}_{\alpha d} = \boldsymbol{x}_d{}'\boldsymbol{\alpha} + \varepsilon_d, \quad d = 1, \dots, D, \tag{15}$$

if $\psi_d = \mathrm{var}(\hat{F}_{\alpha d})$, $d = 1, \dots, D$ are known. Thus, the synthetic regression estimator for the FGT indicator in the area d, $F_{\alpha d}$, is expressed as

$$\hat{F}_{\alpha d}^{REG1-SYN} = \boldsymbol{x}_d{}'\hat{\boldsymbol{\alpha}},$$

where, in this case, $\hat{\boldsymbol{\alpha}} = (\sum_{d=1}^{D} \psi_d^{-1} \boldsymbol{x}_d \boldsymbol{x}_d{}')^{-1} \sum_{d=1}^{D} \psi_d^{-1} \boldsymbol{x}_d \hat{F}_{\alpha d}$.

The model (14) assumed by the REG1-SYN estimator links all the areas by means of the common regression parameter $\boldsymbol{\alpha}$. When estimating this common parameter with the direct estimators$\hat{\delta}_d$ of all the areas, we obtain an estimator with a much smaller variance than for the direct estimator. However, this model does not incorporate heterogeneity between the areas, apart from the heterogeneity explained or due to the auxiliary variables considered. In practice, it is difficult to have data for all the auxiliary variables that fully explain the variation of the indicators δ_d in the areas in which we wish to estimate. Therefore, the synthetic model (14) might not represent many of the cases that appear in practice, providing biased estimators in these cases. In addition, note that, in the most favourable case of knowing the true model (and the true value of $\boldsymbol{\alpha}$), the REG1-SYN estimator would be $\boldsymbol{x}_d'\boldsymbol{\alpha}$, with the result that the data for the variable of interest obtained from the survey for that area would not be being used. Thus, this could be considered wasteful for areas with a large sample size. Furthermore, the estimator obtained may differ greatly from the direct estimator, which would be reliable for these areas. This is a major drawback with synthetic estimators (or models). On the other hand, as mentioned in the introduction, as they are potentially biased estimators under the design, their quality should be evaluated in terms of MSE rather than variance (which will be small, leading one to think, wrongly, that the estimator is of good quality); however, there are no known estimators of the MSE under the design that are stable whilst also different for each area.

These estimators can be summarised as follows:

Target indicators: general parameters.

Data requirements:

- Aggregate data (e.g., population means) of the p auxiliary variables considered in the areas, $\boldsymbol{x}_d$, $d = 1, \dots, D$.

Advantages:

- Variance can decrease considerably in comparison with a direct estimator.
- It can be estimated in unsampled areas.

Disadvantages:

- The synthetic regression model considered does not represent those cases in which not all the auxiliary variables that explain the heterogeneity between areas are available. Therefore, in these cases, the resulting estimators may have a substantial bias.
- It is necessary to analyse the model thoroughly (e.g., by means of the residuals), as the bias of these estimators depends on the goodness of fit of the model. In particular, it is very important to check if there is an area effect, since this model does not consider it.
- If the model is known, the variable of interest data for that area would not be used.
- It does not extend to the direct estimator when the sample size increases.
- There are no known estimators of the MSE under the design that are stable whilst also different for each area.
- They require readjustment to check the benchmarking property whereby the sum of the estimated totals in the areas of a larger region matches the direct estimator for that area.

C. Synthetic regression estimator at individual level

We now consider that individual-level data (or microdata) is available for the p auxiliary variables in the survey, $\boldsymbol{x}_{di}$, $i \in s_d$, $d = 1, \dots, D$. In this case, a synthetic regression estimator for the indicator of interest can be obtained by assuming an individual-level linear regression model for Y_{di}. $\boldsymbol{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$ denotes the vector of values of the variable in question for the individuals in the area d.

The indicator to be estimated in the area d is a function of this vector, i.e., $\delta_d = \delta_d(\boldsymbol{y}_d)$. The basic synthetic regression model considers that the variables Y_{di} for all individuals in the population follow the linear regression model

$$Y_{di} = \boldsymbol{x}_{di}'\boldsymbol{\beta} + \varepsilon_{di}, \quad i = 1, \dots, N_d, d = 1, \dots, D, \tag{16}$$

where the errors ε_{di} are independent, with zero expectation and variance $\sigma^2 k_{di}^2$, where k_{di} are known constants that represent the possible heteroscedasticity in the model ($k_{di} = 1$ for all i and d if there is no heteroscedasticity). Estimating $\boldsymbol{\beta}$ using the weighted least squares estimator $\widehat{\boldsymbol{\beta}} = (\sum_{d=1}^{D} \sum_{i \in s_d} a_{di}\, \boldsymbol{x}_{di}\boldsymbol{x}_{di}')^{-1} \sum_{d=1}^{D} \sum_{i \in s_d} a_{di}\, \boldsymbol{x}_{di} Y_{di}$, where $a_{di} = k_{di}^{-2}$, we obtain predictions, through the model, for every individual in the area, $\hat{Y}_{di} = \boldsymbol{x}_{di}'\widehat{\boldsymbol{\beta}}$, $i = 1, \dots, N_d$. The vector of predictions for the area d is then $\widehat{\boldsymbol{y}}_d = (\hat{Y}_{d1}, \dots, \hat{Y}_{dN_d})'$. Using this vector instead of $\boldsymbol{y}_d$ to calculate the indicator, we obtain the synthetic regression estimator of δ_d, i.e

$$\hat{\delta}_d^{REG2-SYN} = \delta_d(\widehat{\boldsymbol{y}}_d).$$

For example, for the mean of the area d, $\delta_d = \bar{Y}_d$, if $\bar{\boldsymbol{X}}_d$ is the vector of means of the p auxiliary variables considered, the synthetic estimator based on the model (16) would be

$$\hat{\bar{Y}}_d^{REG2-SYN} = \bar{\boldsymbol{X}}_d{}'\widehat{\boldsymbol{\beta}}.$$

For an unsampled area, this estimator is obtained in the same way. For a known $\boldsymbol{\beta}$, the bias under the design of the mean estimator is $\bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta} - \bar{Y}_d$, which does not depend on the sample size of the area n_d; therefore, this bias does not decrease when the sample size increases.

Once again, if we wanted to estimate the FGT indicators of poverty, we would have to find variables $\boldsymbol{x}_{di}$ linearly related to $F_{\alpha,di}$; i.e., that verified the model

$$F_{\alpha,di} = \boldsymbol{x}_{di}'\boldsymbol{\beta} + \varepsilon_{di}, \quad i = 1, \dots, N_d, d = 1, \dots, D. \tag{17}$$

However, finding variables linearly related to $F_{\alpha,di}$ is rare in practice. It would be more suitable to assume the model for the variables of interest, that is, the variables used to measure purchasing power, E_{di} or, even better, for a bijective transformation of these variables, $T(E_{di})$since E_{di} are usually highly asymmetric in their distribution and therefore a linear model for these variables would not be very appropriate. In practice, it is very common to use the logarithm transformation; that is, $Y_{di} = \log(E_{di} + c)$ would be taken as a response variable in the model, where $c > 0$ is a positive constant that makes the distribution of Y_{di} approximately normal. This constant can be determined by fitting the model for a sequence of values of c in the range of E_{di}, and taking the value of c for which a measure of skewness of the model residuals (e.g., Pearson's skewness coefficient) is as close as possible to zero.

As in the case of the previous synthetic estimators, if all the auxiliary variables that explain the heterogeneity of Y_{di} in the areas are not available, (i.e., the synthetic model that is assumed is not verified) then these estimators will be biased. However, its variance will be small since the regression coefficient is estimated using the full sample, which is usually large. Therefore, the synthetic regression estimator will have a small sampling error. These estimators require a study of the goodness of fit of the assumed model in order to avoid large biases. Again, at best, if we knew the model exactly, these estimators would use only the auxiliary variable data and not the variable of interest data observed in the area in question, and they do not come close to the direct estimators for areas with sufficient sample size. In addition, there are no known reliable estimators of the MSE under the design that are different for each area.

These estimators can be summarised as follows:

Target indicators: general parameters.

Data requirements:

- Sample observations of the p auxiliary variables related to the variable of interest, obtained from the same survey from which the data of the variable of interest is obtained.
- For means/totals indicators of the response variable considered in the model, population means/totals of the p auxiliary variables considered in the areas, $\bar{\boldsymbol{X}}_d,\ d = 1, \dots, D$. For non-linear indicators in the response variables of the model, the values of the p auxiliary variables are needed for all individuals (m.d.) in that area, $\{\boldsymbol{x}_{di};\ i = 1, \dots, N_d, d = 1, \dots, D\}$.

Advantages:

- It can considerably reduce the variance of the direct estimators and of the estimators obtained from an area-level model.
- It can be estimated in unsampled areas.

Disadvantages:

- The synthetic regression model considered does not represent those cases in which not all the auxiliary variables that explain the heterogeneity between areas are available. Therefore, in these cases, the resulting estimators may have a substantial bias.
- It is necessary to analyse the model thoroughly (e.g., by means of the residuals), as the bias of these estimators depends on the goodness of fit of the model. In particular, it is very important to check if there is an area effect, since this model does not consider it.
- If the model was known exactly, they would not use the variable of interest data for that area.
- It does not extend to the direct estimator when the sample size increases.
- There are no known estimators of the MSE under the design that are stable whilst also different for each area.
- They require readjustment to check the benchmarking property whereby the sum of the estimated totals in the areas of a larger region matches the direct estimator for that area.

D. Composite estimators

As discussed in previous chapters, direct estimators are (at least approximately) unbiased under the sample design, but may have large variance for areas of small sample sizes. On the other hand, synthetic estimators have small variance, but can be considerably biased under the design. Composite estimators are designed to decrease the variance of the direct estimator in exchange for a portion of the bias of a synthetic estimator. The intention is to simultaneously improve the efficiency of the direct estimator and to reduce the bias of the synthetic estimator. Let $\hat{\bar{Y}}_d^{DIR}$ be a generic direct estimator of $\bar{Y}_d$ and $\hat{\bar{Y}}_d^{SYN}$ a synthetic estimator. A composite estimator of $\bar{Y}_d$ would look like

$$\hat{\bar{Y}}_d^{C} = \phi_d \hat{\bar{Y}}_d^{DIR} + (1 - \phi_d)\hat{\bar{Y}}_d^{SYN}, \quad 0 \leq \phi_d \leq 1.$$

The weight ϕ_d given to the direct estimator can be set either semi-optimally by minimising an approximation of the mean squared error (MSE) under the sample design, which can only be done approximately, or by setting it arbitrarily. Drew, Singh, and Choudhry (1982) proposed a weight ϕ_d that depends on the sample size of the area, giving rise to the sample-size dependent (SSD) estimator. Assuming a pre-set value $\delta > 0$ (by default you can assume 1), the proposed weight would look like

$$\phi_d = \begin{cases} 1, & \text{si} \quad \hat{N}_d \geq \delta N_d; \\ \hat{N}_d/(\delta N_d), & \text{si} \quad \hat{N}_d < \delta N_d, \end{cases}$$

where $\hat{N}_d = \sum_{i \in s_d} w_{di}$. To understand the intuitive idea of this estimator, note that, under simple random sampling (SRS) in the population (in that case the size of the area n_d is random), one obtains

$$\hat{N}_d = \sum_{i \in s_d} w_{di} = \sum_{i \in s_d} \frac{N}{n} = N\frac{n_d}{n}$$

and since $\hat{N}_d$ is unbiased, its expectation under the design is equal to $NE_\pi(n_d)/n = N_d$, so $E_\pi(n_d) = nN_d/N$ and therefore the weight proves to be

$$\phi_d = \begin{cases} 1 & \text{si} \quad n_d \geq \delta E_\pi(n_d); \\ n_d/\{\delta E_\pi(n_d)\} & \text{si} \quad n_d < \delta E_\pi(n_d). \end{cases}$$

If we set $\delta = 1$, then the SSD estimator gives a weight of 1 to the direct estimator when the area sample size is greater than or equal to the expected sample size and gives a weight less than 1 otherwise. However, a given area may have a small sample size n_d , but this may exceed the expected size, which would give weight 1 to the direct estimator and therefore there would be no improvement in efficiency with respect to the direct estimator.

The SSD estimator was used in the Canadian Labour Force Survey to obtain estimators for census tracts assuming $\delta = 2/3$ (see Drew, Singh, and Choudhry (1982)). However, for most of the areas considered, the weight of the direct estimator turned out to be $\phi_d = 1$; for a few, the weight was $\phi_d = 0.9$, but in no case was the weight obtained less than 0.8. Therefore, the gain in efficiency over the direct estimator was very limited. As in this application, the problem with this estimator is that it tends to give the direct estimators a weight close to 1 even though the sample size of the area is small, with no improvement in efficiency with respect to the direct estimator. Furthermore, the weight ϕ_d does not consider whether or not the areas are very homogeneous as regards satisfying the model considered by the synthetic estimator. It is therefore independent of the quality of the synthetic estimator (or the goodness of fit of the synthetic model) for each area. Therefore, these estimators can be considered too simple to return a discernible improvement in efficiency over the direct estimators.

As previously stated, it is possible to obtain approximately optimal composite estimators with respect to the sample design assuming the weight ϕ_d that (approximately) minimises the MSE under the composite estimator design, $\text{MSE}_\pi(\hat{\bar{Y}}_d^C)$. Considering that the covariance between the direct estimator and the synthetic estimator is negligible, and by minimising

$$\text{MSE}_\pi(\hat{\bar{Y}}_d^C) \approx \phi_d^2 \text{var}_\pi(\hat{\bar{Y}}_d^{DIR}) + (1 - \phi_d)^2 \text{MSE}_\pi(\hat{\bar{Y}}_d^{SYN}),$$

the optimum weight is obtained

$$\phi_d^* = \text{MSE}_\pi(\hat{\bar{Y}}_d^{SYN})/\{\text{var}_\pi(\hat{\bar{Y}}_d^{DIR}) + \text{MSE}_\pi(\hat{\bar{Y}}_d^{SYN})\}. \tag{18}$$

An estimator of $\text{MSE}_\pi(\hat{\bar{Y}}_d^{SYN})$ is

$$\widehat{\text{MSE}}_\pi(\hat{\bar{Y}}_d^{SYN}) = (\hat{\bar{Y}}_d^{SYN} - \hat{\bar{Y}}_d^{DIR})^2 - \widehat{\text{var}}_\pi(\hat{\bar{Y}}_d^{DIR}),$$

(see Rao and Molina (2015, p.44)). By replacing this estimator in the optimal weight ϕ_d^* given in (18), we obtain an estimator of this optimal weight, expressed as

$$\hat{\phi}_d^* = \widehat{\text{MSE}}_\pi(\hat{\bar{Y}}_d^{SYN})/(\hat{\bar{Y}}_d^{SYN} - \hat{\bar{Y}}_d^{DIR})^2 = 1 - \widehat{\text{var}}_\pi(\hat{\bar{Y}}_d^{DIR})/(\hat{\bar{Y}}_d^{SYN} - \hat{\bar{Y}}_d^{DIR})^2.$$

We can see that this weight depends on the direct estimator $\hat{\bar{Y}}_d^{DIR}$, which is very volatile. This means that the estimated optimal weight $\hat{\phi}_d^*$ is also very volatile. A more stable estimated weight can be obtained by averaging over all the areas, as follows:

$$\begin{aligned}\hat{\phi}^* &= \sum_{\ell=1}^{D} \widehat{\mathrm{MSE}}_\pi\,(\hat{\bar{Y}}_\ell^{SYN}) / \sum_{\ell=1}^{D} (\hat{\bar{Y}}_\ell^{SYN} - \hat{\bar{Y}}_\ell^{DIR})^2 \\ &= 1 - \left\{\sum_{\ell=1}^{D} \widehat{\mathrm{var}}_\pi\,(\hat{\bar{Y}}_\ell^{DIR}) / \sum_{\ell=1}^{D} (\hat{\bar{Y}}_\ell^{SYN} - \hat{\bar{Y}}_\ell^{DIR})^2\right\}\end{aligned}$$

The resulting weight, $\hat{\phi}^*$, is very stable, but it does not depend on the area d; i.e., it is constant for all areas, and doesn't even depend on their sample size. Due to these disadvantages, optimal composite estimators are probably less used in practice than the "model-based" ones that we will see in the next chapter.

Composite estimators are interesting because of the trade-off achieved between bias and variance. However, in the following chapters we will see that composite estimators can be obtained more efficiently than direct estimators from regression models that take account of heterogeneity across areas. These composite estimators will be optimal with respect to the probability distribution triggered by the assumed model, and that is why they are called "model-based" estimators. In these estimators, the weights depend on the sample size of the area and the goodness of fit of the synthetic model, with greater weight given to the direct estimators when the synthetic model is poor (auxiliary variables that are not very informative or very heterogeneous areas) or when the sample size of the area is large, and greater weight is given to the synthetic estimator as the sample size decreases or the model has more predictive capacity. Therefore, model-based estimators exceed these simple composite estimators.

Next, we will summarise the characteristics of the SSD estimator, as the most common representative of the composite estimators:

Target indicators: additive parameters.

Data requirements:

- Sampling weights w_{di} for sample individuals in the area d.
- Population size of the area, N_d, whether the HT estimator of the mean or the Hájek estimator of the total is used.

Advantages:

- They are designed to reduce both the bias of the synthetic estimator and the variance of the direct estimator. They cannot be less efficient than the direct estimator and the bias cannot be greater than the synthetic estimator.

Disadvantages:

- For an area of small sample size, as long as this size is not smaller than the expected sample size, no information is borrowed from the other areas through the synthetic estimator. Therefore, there will be no gain in efficiency with respect to the direct estimator considered.
- The weight given to the synthetic estimator does not depend on how well the variable of interest is explained by the auxiliary variables; i.e., it does not depend on the goodness of fit of the model.

- They cannot be calculated for unsampled areas or domains; i.e., with sample size n_d equal to zero.
- There are no known stable estimators of the MSE under the design that are also different for each area.
- They require readjustment to verify the benchmarking property so that the sum of the estimated totals in the areas of a larger region matches the direct estimator for that region.

Example 4. Composite estimators of poverty incidence, with R. Continuing the previous examples, we now demonstrate how to obtain SSD composite estimators of poverty incidence for the provinces, using the direct HT estimators obtained in Example 1 and the synthetic post-stratified estimators obtained in Example 3. To do so, we call on the ssd() function using the default value of the delta parameter (delta=1) and save the results:

```
povinc.ssd.res<-ssd(dom=prov,sweight=weight,domsize=sizeprov[,c(2,3)],
 direct=povinc.dir.res[,c("Domain","Direct")],synthetic=povinc.psedu.res)
povinc.ssd<-povinc.ssd.res$ssd
```

We analyse the weight given by the SSD estimator to the direct estimator for each province by means of a descriptive summary of these weights:

```
summary(povinc.ssd.res$CompWeight)
```

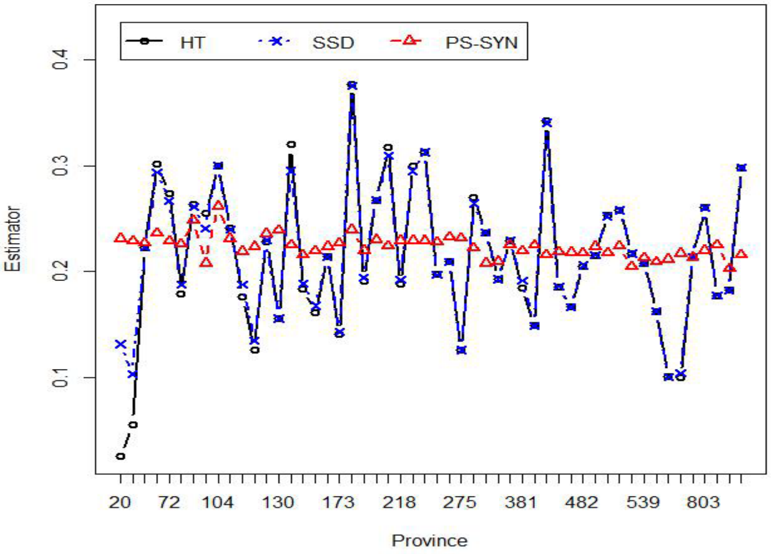
The result is:

```
   Min.  1st Qu.  Median    Mean  3rd Qu.    Max. 
 0.4846  0.8800   0.9779   0.9224   1.0000   1.0000
```

We can see that the direct estimators are given a weight equal to one for at least a quarter of the provinces. In those specific provinces, information is not being borrowed from the others. On the other hand, in this estimator this weight does not depend on the variable of interest. If we estimate, for example, the average income, we get exactly the same weights. Indeed, if we graphically compare the SSD estimates with the direct HT and synthetic post-stratified estimates (Figure 5), we can see that they are very similar to the direct HT estimates. In this graph, the provinces (on the axis) are arranged from smallest to largest sample size, and their sample sizes are indicated on the axis labels. The R code implemented to obtain the previous figures is as follows:

```
o<-order(nd)
k<-2
M<-max(povinc.psedu,povinc.dir,povinc.ssd)
m<-min(povinc.psedu,povinc.dir,povinc.ssd)
plot(1:D,povinc.dir[o],type="n",ylim=c(m,M+(M-m)/k),xlab="Province", ylab="Estimator",xaxt="n")
points(1:D,povinc.dir[o],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:D,povinc.greg[o],type="b",col=3,lty=3,pch=3,lwd=2)
points(1:D,povinc.ssd[o],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:D,povinc.psedu[o],type="b",col=2,lty=2,pch=2,lwd=2)
axis(1, at=1:D, labels=nd[o])
legend(1,M+(M-m)/k,legend=c("HT","GREG","SSD","PS-SYN"),ncol=4,col=c(1,3,4,2),
lwd=rep(2,3),lty=c(1,3,4,2),pch=c(1,3,4,2))
```

Figure 5
HT, PS-SYN and SSD estimates of the poverty incidence for each province
(In proportions)



Source: Prepared by the author.

V. Model-based indirect methods

Model-based small area estimators fall into the group of indirect estimators since they borrow information from other areas. However, they are somewhat more sophisticated than the basic indirect estimators discussed in chapter IV, in that they incorporate heterogeneity between areas that is not explained by the auxiliary variables considered. This is done by incorporating additive random effects on the areas into the regression model considered. We will see that these random effects provide a very good property to linear model-based estimators, which is that they can be written as composite estimators that extend to a direct estimator in areas with sufficient sample size. Having all the variables that fully explain the heterogeneity between areas of our variable of interest will rarely occur. Therefore, these models are significantly more realistic than synthetic models, resulting in estimators with lower bias under the sample design.

A. EBLUP based on the Fay-Herriot model

The Fay-Herriot (FH) model is a popular area-level model that was introduced by Fay and Herriot (1979) to estimate per capita income in small areas of the USA. This model is currently used by the U.S. Census Bureau. Within the Small Area Income and Poverty Estimates (SAIPE) programme to estimate proportions of poor school-age children in counties and school districts (for further details see Bell (1997) or http://www.census.gov/ hhes/www/saipe). This model has also been used in Chile to estimate poverty incidence rates in Chilean comunas (see Casas-Cordero Valencia, Encina and Lahiri (2015)) and in Spain to estimate the poverty incidence and poverty gap in provinces by gender (Molina and Morales, 2009).

This model links the indicators of interest for all the areas δ_d, $d = 1, \dots, D$, assuming that they vary with respect to a vector with values of p auxiliary variables $\boldsymbol{x}_d$ constantly for all the areas, following the linear regression model

$$\delta_d = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d, \quad d = 1, \dots, D, \tag{19}$$

where $\boldsymbol{\beta}$ is the vector of coefficients common to all areas and u_d is the regression error term, different for each area, also known as the random area effect d. These random effects u_d represent the heterogeneity of the indicators δ_d across the areas, which is not due to (or not explained by) the auxiliary variables considered. In the simplest model, such random effects u_d are assumed to be independent and identically distributed (IID), with unknown common variance σ_u^2; this is indicated by $u_d \sim^{iid}(0, \sigma_u^2)$.

Since the true values of the indicators δ_d are not observable, the model (19) cannot be fitted. When using a direct estimator $\hat{\delta}_d^{DIR}$ of δ_d, we must consider that this estimator has a sampling error. The FH model considers this direct estimator $\hat{\delta}_d^{DIR}$ to be unbiased under the design. In this case, we can represent the sampling error of this estimator by using the model:

$$\hat{\delta}_d^{DIR} = \delta_d + e_d, \quad d = 1, \dots, D, \tag{20}$$

where e_d is the sampling error in the area d. It is assumed that the sampling errors e_d are independent of each other and are also independent of the random effects on the areas, u_d, and have zero mean value and known variances ψ_d; i.e., $e_d \sim^{ind}(0, \psi_d)$. In practice, these variances, $\psi_d = \text{var}_\pi(\hat{\delta}_d^{DIR} | \delta_d)$, $d = 1, \dots, D$, are estimated using the microdata from the survey. Combining models (19) and (20), we obtain the linear mixed model expressed as

$$\hat{\delta}_d^{DIR} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \dots, D. \tag{21}$$

Using the Lagrange multiplier method to calculate the linear estimator in the data $\hat{\delta}_d^{DIR}$, $d = 1, \dots, D$, which is unbiased under the model (21), and which minimises the MSE under the model, we obtain the best linear unbiased predictor (BLUP) of $\delta_d = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d$. The resulting estimator is obtained by simply fitting the mixed model (21); i.e., the BLUP under the FH model of δ_d is expressed as

$$\tilde{\delta}_d^{FH} = \boldsymbol{x}_d{}'\widetilde{\boldsymbol{\beta}} + \tilde{u}_d, \tag{22}$$

where $\tilde{u}_d = \gamma_d(\hat{\delta}_d^{DIR} - \boldsymbol{x}_d{}'\widetilde{\boldsymbol{\beta}})$ is the BLUP of u_d, where $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d)$ and where $\widetilde{\boldsymbol{\beta}}$ is the weighted least squares estimator of $\boldsymbol{\beta}$ under the model (21), expressed as

$$\widetilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \gamma_d\, \boldsymbol{x}_d \boldsymbol{x}_d{}'\right)^{-1} \sum_{d=1}^{D} \gamma_d\, \boldsymbol{x}_d \hat{\delta}_d^{DIR}.$$

Note that, substituting $\tilde{u}_d = \gamma_d(\hat{\delta}_d^{DIR} - \boldsymbol{x}_d{}'\widetilde{\boldsymbol{\beta}})$ into the BLUP under the FH model given in (22), we can express the BLUP as a convex linear combination of the direct estimator and the synthetic regression estimator, i.e,

$$\tilde{\delta}_d^{FH} = \gamma_d \hat{\delta}_d^{DIR} + (1 - \gamma_d)\boldsymbol{x}_d{}'\widetilde{\boldsymbol{\beta}}, \tag{23}$$

with a weight for the direct estimator expressed as $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d) \in (0,1)$. This weight depends on the sample size of the area by means of the variance ψ_d of the direct estimator and the goodness of fit of the synthetic model measured by σ_u^2 (in other words, the unexplained heterogeneity between the areas). Therefore, for an area d in which the direct estimator $\hat{\delta}_d^{DIR}$ is efficient due to having sufficient sample size; i.e. with a small sample variance ψ_d compared to the unexplained heterogeneity σ_u^2, $\gamma_d = \sigma_u^2/(\sigma_u^2 + \psi_d)$ is close to one and therefore $\tilde{\delta}_d^{FH}$ gives more weight to the direct estimator. On the other hand, in areas d where the direct estimator lacks quality due to small sample size, where its sample variance ψ_d is larger than the unexplained heterogeneity σ_u^2, then γ_d approaches zero and therefore more weight is given to the synthetic regression estimator $\boldsymbol{x}_d{}'\widetilde{\boldsymbol{\beta}}$, which uses data from all areas to estimate the common parameter $\boldsymbol{\beta}$. In other words, this estimator borrows information from the other areas by means of the synthetic regression estimator $\boldsymbol{x}_d{}'\widetilde{\boldsymbol{\beta}}$ where required, depending on the efficiency of the direct estimator.

Moreover, the fact that the BLUP $\tilde{\delta}_d^{FH}$ comes closer to the direct estimator when the sample size of the area is large (ψ_d small) is a very desirable property, since we do not need to know when an area is "small" enough to use this estimator instead of the direct estimator, since it extends to the direct estimator when the sample size grows, and it also improves the direct estimator in areas with small sample size. Therefore, in principle, this estimator can be used for all areas as long as there is a "small" one (if there were none, it wouldn't be necessary to use it).

The BLUP of δ_d depends on the true value of the variance σ_u^2 of the random effects u_d. In practice, this variance is unknown and must be estimated. Common estimation methods are maximum likelihood (ML) and restricted/residual ML (REML). The REML method corrects the variance estimator σ_u^2 or the degrees of freedom due to estimating the regression coefficients $\boldsymbol{\beta}$ and thus provides a less biased estimator for the finite sample size$n$. A method of adjustment based on moments, which does not need a parametric distribution to obtain the likelihood, is that advanced by Fay and Herriot (1979), which we call the FH method. Let $\hat{\sigma}_u^2$ be a consistent estimator of σ_u^2like those obtained by these methods. By replacing σ_u^2 by $\hat{\sigma}_u^2$ en (22), we obtain the empirical BLUP (EBLUP) of δ_d,

$$\hat{\delta}_d^{FH} = \hat{\gamma}_d \hat{\delta}_d^{DIR} + (1 - \hat{\gamma}_d)\boldsymbol{x}_d{}'\widehat{\boldsymbol{\beta}}, \tag{24}$$

where $\hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \psi_d)$ and $\widehat{\boldsymbol{\beta}} = (\sum_{d=1}^{D} \hat{\gamma}_d\, \boldsymbol{x}_d \boldsymbol{x}_d{}')^{-1} \sum_{d=1}^{D} \hat{\gamma}_d\, \boldsymbol{x}_d \hat{\delta}_d^{DIR}$. In this paper, for purposes of conciseness, we will call the EBLUP based on the FH model given in (24) the FH estimator.

If the parameters of the model $\boldsymbol{\beta}$ and σ_u^2 are known, the MSE of the BLUP, $\tilde{\delta}_d^{FH}$, based on the model (21) is expressed as

$$\mathrm{MSE}(\tilde{\delta}_d^{FH}) = \gamma_d \psi_d \leq \psi_d = \mathrm{var}_\pi(\hat{\delta}_d^{DIR} | \delta_d).$$

Therefore, given the true value of the indicator δ_d, if σ_u^2 and $\boldsymbol{\beta}$ are known, the BLUP under the FH model, $\tilde{\delta}_d^{FH}$, cannot be less efficient than the direct estimator. In practice, σ_u^2 and $\boldsymbol{\beta}$ are estimated and the error due to the estimation of these two parameters is added to the MSE of the FH estimator. However, these two terms that are added to the MSE tend to zero when the number of areas D tends to infinity. Therefore, for a sufficient number of areas D , the FH estimator is still likely to improve on the direct estimator in terms of MSE. That is why these estimators tend to improve in most areas as long as there are a sufficient number of areas. However, improvements in efficiency will be small if the number of areas is not sufficiently large. Unit-level models, based on the total sample size n, can be much more efficient than area-level models, as long as there are auxiliary variables at the individual level that are sufficiently informative about the response variable. However, an advantage of the FH estimator shown in (24) is that it uses the weights of the sample design through the direct estimator and is consistent under the design when the sample size of the area n_d grows, while the weight of the direct estimator is $\gamma_d > 0$. Furthermore, its absolute bias under the design is expressed as

$$(1 - \gamma_d)|\delta_d - \boldsymbol{x}_d{}'\boldsymbol{\beta}| \leq |\delta_d - \boldsymbol{x}_d{}'\boldsymbol{\beta}|,$$

thus, it will be less biased than the synthetic regression estimator based on the same vector of coefficients $\boldsymbol{\beta}$ while $\gamma_d > 0$.

For an unsampled area; i.e., with sample size $n_d = 0$, the variance of the direct estimator ψ_d would tend to infinity and γ_d would tend to zero. Assuming the limit value $\gamma_d = 0$, the synthetic regression estimator is obtained.

$$\hat{\delta}_d^{FH} = \boldsymbol{x}_d{}'\widehat{\boldsymbol{\beta}}.$$

Under normality of u_d and e_d, Prasad and Rao (1990) obtained a second-order approximation (i.e., with error $o(D^{-1})$ when the number of areas D is large) for the MSE of the FH estimator, expressed as

$$\mathrm{MSE}(\hat{\delta}_d^{FH}) = g_{d1}(\sigma_u^2) + g_{d2}(\sigma_u^2) + g_{d3}(\sigma_u^2),$$

where

$$
\begin{aligned}
g_{1d}(\sigma_u^2) &= \gamma_d\psi_d,\\
g_{2d}(\sigma_u^2) &= (1-\gamma_d)^2\boldsymbol{x}_d{}'\left(\sum_{d=1}^{D}\gamma_d\,\boldsymbol{x}_d\boldsymbol{x}_d{}'\right)^{-1}\boldsymbol{x}_d,\\
g_{3d}(\sigma_u^2) &= (1-\gamma_d)^2(\sigma_u^2+\psi_d^2)^{-1}\overline{\text{var}}(\hat{\sigma}_u^2).
\end{aligned}
$$

Here, $\overline{\text{var}}(\hat{\sigma}_u^2)$ is the asymptotic variance of the estimator $\hat{\sigma}_u^2$ of σ_u^2, which depends on the estimation method used, $g_{1d}(\sigma_u^2)$ is the error due to the prediction of the random effect of the area u_d, of the order of $O(1)$ when D grows (i.e. does not tend to zero), $g_{2d}(\sigma_u^2)$ is the error due to the estimation of the vector of regression coefficients $\boldsymbol{\beta}$ and $g_{3d}(\sigma_u^2)$ is the error due to the estimation of the variance σ_u^2, where the last two terms tend to zero when D grows with order $O(D^{-1})$; i.e. at the same rate as D^{-1}. This means that $g_{2d}(\sigma_u^2)$ and $g_{3d}(\sigma_u^2)$ disappear for a large enough D, while $g_{1d}(\sigma_u^2)$ does not disappear, but for moderate D all three terms must be taken into account to avoid underestimation of the MSE.

If $\hat{\sigma}_u^2$ is the REML estimator, the asymptotic variance is obtained as the inverse of Fisher's information $\mathcal{I}(\sigma_u^2)$, and is expressed as

$$
\overline{\text{var}}(\hat{\sigma}_u^2) = \mathcal{I}^{-1}(\sigma_u^2) = 2\left\{\sum_{d=1}^{D}(\,\sigma_u^2+\psi_d)^{-2}\right\}^{-1}. \qquad (25)
$$

In this case, $g_{d2}(\hat{\sigma}_u^2)$ and $g_{d3}(\hat{\sigma}_u^2)$ are respective estimators of $g_{2d}(\sigma_u^2)$ and $g_{3d}(\sigma_u^2)$ unbiased second-order estimators. This means that its bias is $o(D^{-1})$, i.e., it tends to zero faster than D^{-1} when D grows. However, $g_{d1}(\hat{\sigma}_u^2)$ has a non-negligible bias as an estimator of $g_{d1}(\sigma_u^2)$ which turns out to be equal to $-g_{3d}(\sigma_u^2)+o(D^{-1})$. Therefore, to correct for the bias of $g_{1d}(\hat{\sigma}_u^2)$, we must aggregate twice $g_{3d}(\hat{\sigma}_u^2)$. Thus, an unbiased second-order MSE estimator of the FH estimator, known here as the Prasad-Rao estimator, is then expressed as

$$
\text{mse}_{PR}(\hat{\delta}_d^{FH}) = g_{d1}(\hat{\sigma}_u^2) + g_{d2}(\hat{\sigma}_u^2) + 2g_{d3}(\hat{\sigma}_u^2).
$$

If $\hat{\sigma}_u^2$ is the ML estimator, its asymptotic variance is the same as for the REML estimator, given in (25). However, this estimator has a bias that is expressed as

$$
b(\sigma_u^2) = -\{2\mathcal{I}(\sigma_u^2)\}^{-1}\text{traza}\left[\left\{\sum_{d=1}^{D}(\,\sigma_u^2+\psi_d)^{-1}\boldsymbol{x}_d\boldsymbol{x}_d{}'\right\}^{-1}\sum_{d=1}^{D}(\,\sigma_u^2+\psi_d)^{-2}\boldsymbol{x}_d\boldsymbol{x}_d{}'\right].
$$

In this case, the bias of the ML estimator adds a term to the bias of $g_{d1}(\hat{\sigma}_u^2)$ as the estimator of $g_{d1}(\sigma_u^2)$. This bias is equal to $b(\sigma_u^2)\nabla g_{1d}(\sigma_u^2) - g_{3d}(\sigma_u^2)$, where

$$
\nabla g_{1d}(\sigma_u^2) = (1-\gamma_d)^2.
$$

Since $b(\hat{\sigma}_u^2)\nabla g_{1d}(\hat{\sigma}_u^2)$ is an unbiased second-order estimator of $b(\sigma_u^2)\nabla g_{1d}(\sigma_u^2)$, we can correct for the bias of $g_{1d}(\hat{\sigma}_u^2)$ by subtracting this term. In this way, we obtain the following unbiased second-order MSE estimator of the FH estimator,

$$
\text{mse}_{PR}(\hat{\delta}_d^{FH}) = g_{d1}(\hat{\sigma}_u^2) - b(\hat{\sigma}_u^2)\nabla g_{1d}(\hat{\sigma}_u^2) + g_{d2}(\hat{\sigma}_u^2) + 2g_{d3}(\hat{\sigma}_u^2). \qquad (26)
$$

If $\hat{\sigma}_u^2$ is the estimator obtained by the moment-based FH method, the second-order unbiased estimator of the MSE has the same form as (26), but the bias of the FH estimator of σ_u^2 and the asymptotic variance change, and are expressed as

$$\overline{\mathrm{var}}(\hat{\sigma}_u^2) = 2D\left\{\sum_{d=1}^{D}(\sigma_u^2 + \psi_d)^{-1}\right\}^{-2}, \tag{27}$$

$$b(\sigma_u^2) = \frac{2[D\sum_{d=1}^{D}(\sigma_u^2 + \psi_d)^{-2} - \{\sum_{d=1}^{D}(\sigma_u^2 + \psi_d)^{-1}\}^2]}{\{\sum_{d=1}^{D}(\sigma_u^2 + \psi_d)^{-1}\}^3}.$$

When estimating the FGT indicator of order α, $\delta_d = F_{\alpha d}$, using the FH model, auxiliary variables $\boldsymbol{x}_d$ must be found to verify the model

$$F_{\alpha d} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d, \quad d = 1, \dots, D, \tag{28}$$

and it is assumed that the direct estimator $\hat{F}_{\alpha d}^{DIR}$ of $F_{\alpha d}$ satisfies

$$\hat{F}_{\alpha d}^{DIR} = F_{\alpha d} + e_d, \quad d = 1, \dots, D. \tag{29}$$

The linear mixed model obtained by combining (28) and (29) is expressed as

$$\hat{F}_{\alpha d}^{DIR} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d + e_d, \quad d = 1, \dots, D. \tag{30}$$

Fitting this model, the BLUP of $F_{\alpha d} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d$ would be

$$\tilde{F}_{\alpha d}^{FH} = \boldsymbol{x}_d{}'\widetilde{\boldsymbol{\beta}} + \tilde{u}_d, \tag{31}$$

where, in this case, $\tilde{u}_d = \gamma_d(\hat{F}_{\alpha d}^{DIR} - \boldsymbol{x}_d{}'\widetilde{\boldsymbol{\beta}})$ is the BLUP of u_d and $\widetilde{\boldsymbol{\beta}}$ it is calculated as follows:

$$\widetilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D}\gamma_d\,\boldsymbol{x}_d\boldsymbol{x}_d{}'\right)^{-1}\sum_{d=1}^{D}\gamma_d\,\boldsymbol{x}_d\hat{F}_{\alpha d}^{DIR}.$$

The final FH estimator of $F_{\alpha d}$ is obtained by simply replacing the variance σ_u^2 by a consistent estimator $\hat{\sigma}_u^2$ in the BLUP (31).

The characteristics of the FH estimator can be summarised as follows:

Target indicators: general parameters.

Data requirements:

- Aggregate data (e.g., population means) of the p auxiliary variables considered in the areas, $\boldsymbol{x}_d$, $d = 1, \dots, D$.

Advantages:

- It usually improves the efficiency of the direct estimator.
- The considered regression model incorporates unexplained heterogeneity between areas.
- It is a composite estimator that automatically borrows information from the remaining areas (giving greater weight to the synthetic regression estimator) where required (when the direct estimator has greater variance, or smaller sample size). It tends to the direct estimator when the size of the area grows (as ψ_d becomes small).
- If, for an area d, the weight given to the direct estimator is strictly positive ($\gamma_d > 0$), the sampling weights w_{di} are used through the direct estimator $\hat{\delta}_d^{DIR}$; i.e., the sample design is taken into account. Consequently, it is consistent under the design (as is the direct estimator). This means that it will be less affected by informative designs (designs with selection

probabilities for individuals depending on the variable of interest), by considering that the sampling weights are the true ones.

- Because aggregate data is used, the FH estimator is not overly affected by isolated outliers (in this case direct atypical estimators for an area).
- By using only aggregated auxiliary information, it avoids the confidentiality problems of microdata obtained from a census or administrative record.
- For linear direct estimators, the Central Limit Theorem is applied for areas with sufficient sample size. Thus, the model will always have a minimum goodness of fit for areas of sufficient sample size.
- It can be estimated in unsampled areas.
- The Prasad-Rao estimator of the MSE is stable (or efficient) and is unbiased under the design when averaged over many areas.

Disadvantages:

- The estimators are based on a model; thus, it is necessary to analyse the model (e.g., by means of the residuals). For non-linear parameters, we can have linearity problems.
- The sampling variances of the direct estimators ψ_d are assumed to be known, although in practice it is necessary to estimate them, which leads to the same problem of lack of data in an area. Incorporating the estimation error of these variances in the MSE of the FH estimator is not automatic and often the estimated MSE does not incorporate this error.
- The number of observations used to fit the model is the number of areas sampled, which is usually much smaller than the total sample size n used to fit individual-level models. Thus, the model parameters are estimated with lower efficiency and improvements in efficiency compared with the direct estimators will be lower than with individual-level models (this efficiency increases with the number of areas). In our applications we have obtained very small gains over the direct estimator.
- When estimating several indicators that depend on a common variable (e.g., $F_{\alpha d}$ for different values of α), as opposed to methods based on unit-level models, modelling, and searching of useful auxiliary variables is required for each of the indicators separately.
- The MSE estimator under the Prasad-Rao model is correct under the model with normality of u_d and e_d, and is not unbiased under the design for the MSE under the design for a particular area.
- Once the model has been fitted at the area level, the estimators $\hat{\delta}_d^{FH}$ cannot be disaggregated to subdomains or subareas within areas unless a new model is found that is suitable for that new level or, alternatively, a multilevel random effects model is fitted.
- They require refitting to verify the benchmarking property: that the sum of the estimated totals in the areas of a larger region matches the direct estimator for that area.

Example 5. FH estimators of poverty incidence, with R. Continuing with the previous examples, we demonstrate how to obtain FH estimators of poverty incidence in R for the provinces. Firstly, to check whether the hypothesis of normality of the model is verified, we can analyse graphically the distribution of the direct estimators of poverty incidence by means of the histogram:

```
hist(povinc.dir,prob=TRUE,main="",xlab="HT estimators pov. incidence")
```

The shape of this histogram (not included for purposes of conciseness) is somewhat asymmetric but is not too distant from a normal density, which is to be expected since the Central Limit Theorem is applied to the direct estimators of the areas.

Next, we load the datasets with the population sizes of the provinces and by nationality, age, and employment status groups (some were already loaded in the previous examples):

```
data(sizeprov)
data(sizeprovnat)
data(sizeprovage)
data(sizeprovedu)
data(sizeprovlab)
```

We use these population sizes to calculate the proportions of individuals in each category within each province. These will be our explanatory variables in a Fay-Herriot model:

```
Nd<-sizeprov[,3]
Ndnat<-as.matrix(sizeprovnat[,-c(1,2)])
Ndage<-as.matrix(sizeprovage[,-c(1,2)])
Ndedu<-as.matrix(sizeprovedu[,-c(1,2)])
Ndlab<-as.matrix(sizeprovlab[,-c(1,2)])

Pdnat<-Ndnat/Nd
Pdage<-Ndage/Nd
Pdedu<-Ndedu/Nd
Pdlab<-Ndlab/Nd

# Design matrix for FH model
X<-cbind(const=rep(1,D),nat1=Pdnat[,1],Pdage[,3:5],Pdedu[,c(1,3)],Pdlab[,c(2,3)])
```

We call on the function that calculates the FH estimators of poverty incidence for the provinces, using the direct HT estimators obtained in Example 1 and their corresponding sampling variances:

```
povinc.FH.res<-eblupFH(povinc.dir~X-1,vardir=povinc.dir.res$SD^2)
povinc.FH<-povinc.FH.res$eblup
```

Using the estimated regression coefficients obtained from fitting the Fay-Herriot model, we can also calculate synthetic regression estimators based on the model at the area level:

```
povinc.rsyn1<-X%*%povinc.FH.res$fit$estcoef[,1]
```

Although these estimators are based on the estimator of the regression coefficients obtained from the fitting of the Fay-Herriot model and not from the synthetic model, they are also synthetic estimators because they do not consider heterogeneity between areas that is not explained by the considered auxiliary variables. Moreover, the estimators of the regression coefficients obtained under both models, using the same auxiliary variables, are asymptotically equivalent. Thus, for a large number of areas, they will both be very similar.

As the FH estimators are composite estimators between direct and synthetic regression estimators, we calculate the weights given to the direct estimators in the composite and show a descriptive summary of them:

```
gammad<-povinc.FH.res$fit$refvar/(povinc.FH.res$fit$refvar+povinc.dir.res$SD^2)
summary(gammad)
```

Result:

```
  Min.   1st Qu.  Median   Mean     3rd Qu.  Max.
0.4537   0.7182   0.8108   0.7906   0.8977   0.9477
```

We see that, unlike the SSC estimators, in this case the weight given to the direct estimator does not equal one for any province, although it does assume values close to one for some provinces.

We now graphically compare the FH estimates with the direct HT and synthetic RSYN1 estimates for each province. The provinces (on the axis) are arranged from smallest to largest sample size, and we indicate their sample sizes on the axis:

```
o<-order(nd)
k<-6
M<-max(povinc.dir,povinc.FH,povinc.rsyn1)
m<-min(povinc.dir,povinc.FH,povinc.rsyn1)
plot(1:D,povinc.dir[o],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="Estimator",

 xaxt="n")
points(1:D,povinc.dir[o],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:D,povinc.FH[o],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:D,povinc.rsyn1[o],type="b",col=3,lty=3,pch=3,lwd=2)
axis(1, at=1:D, labels=nd[o])
legend(1,M+(M-m)/k,legend=c("DIR","FH","RSYN1"),ncol=3,col=c(1,4,3),lwd=rep(2,3),
 lty=c(1,4,3),pch=c(1,4,3))
```

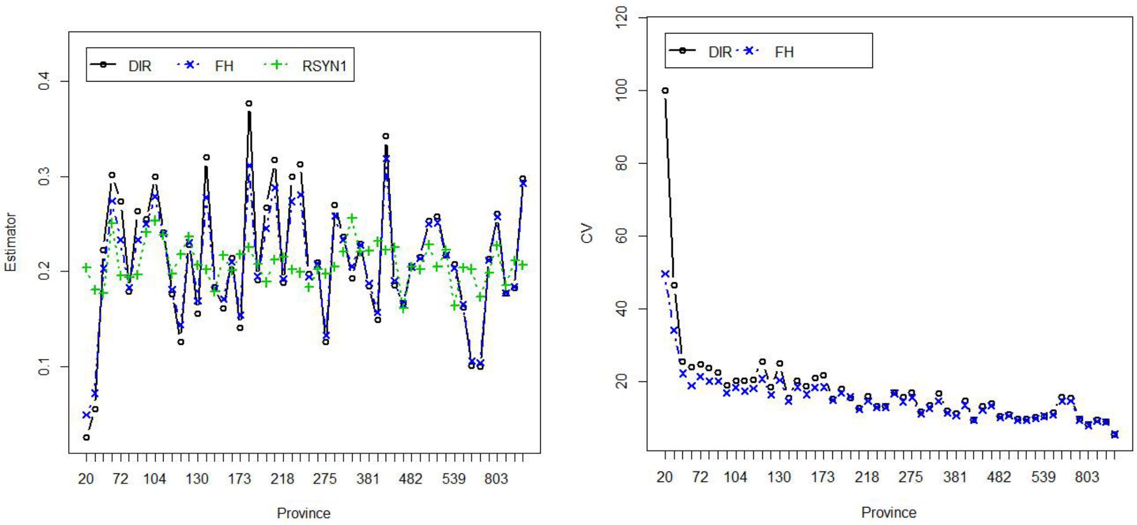
Finally, we estimate the MSE of the FH estimators by calling on the `mseFH()` function, we calculate the estimated CVs and plot the MSEs together with the variances of the direct estimators:

```
povinc.FH.mse.res<-mseFH(povinc.dir~X-1,vardir=povinc.dir.res$SD^2)

povinc.FH.mse<-povinc.FH.mse.res$mse
povinc.FH.cv<-100*sqrt(povinc.FH.mse)/povinc.FH

M<-max(povinc.dir.var,povinc.FH.mse)
m<-min(povinc.dir.var,povinc.FH.mse)
plot(1:D,povinc.dir.cv[o],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="CV",xaxt="n")
points(1:D,povinc.dir.var[o],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:D,povinc.FH.mse[o],type="b",col=4,lty=4,pch=4,lwd=2)
axis(1, at=1:D, labels=nd[o])
legend(1,M+(M-m)/k,legend=c("DIR", "FH"),ncol=3,col=c(1,4),lwd=rep(2,2),lty=c(1,4),pch=c(1,4))
```

Once again, we can see in Figure 6 (left) that the synthetic regression estimators assume very similar values for all provinces, unlike the direct estimators, which vary more across provinces. The FH estimators are close to the direct estimators, but they also borrow information from the other provinces by means of synthetic estimators, especially for the provinces with smaller sample size (left-hand graph). Although in this example the considered auxiliary variables are not very powerful, figure 6 (right) indicates that the FH estimators are more efficient than the direct estimators.

Finally, we will compare the estimated CVs for the HT, GREG, and FH estimators for the 5 provinces with smaller sample sizes:

Figure 6
FH, direct HT and RSYN1 estimates of poverty incidence for the provinces (left), and MSEs estimated from the FH and direct HT estimators (right)
(In proportions)



Source: Prepared by the author.

```
compardirFH<-data.frame(povinc.dir.cv,povinc.greg.cv,povinc.FH.cv)

selprov<-o[1:5]
compardirFH[selprov,]
```

Results:

```
    povinc.dir.CV povinc.greg.cv povinc.FH.cv
42       99.97815       94.72703     49.34572
5        46.35946       42.04802     33.74811
40       25.33449       21.77035     21.64444
34       23.80085       19.02477     18.27171
44       24.57017       16.86049     20.47468
```

We can see the reduction in CVs achieved by the FH estimators compared to the HT direct estimators. They are also more efficient than the GREG estimators for the four provinces with smaller sample sizes, and these improvements are significant for the two provinces with smaller sample sizes.

B. Model-based EBLUP with nested errors

The model with nested errors was proposed by Battese, Harter and Fuller (1977) to estimate corn and soybean production at a county level in the U.S. This model linearly relates the values of a variable of interest Y_{di} for the individual i within the area d, with the values of p auxiliary variables for that same individual, as follows:

$$Y_{di} = \boldsymbol{x}_{di}'\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \dots, N_d,\ d = 1, \dots, D, \tag{32}$$

where $\boldsymbol{\beta}$ is the vector of coefficients of the auxiliary variables, common to all areas, u_d is the random effect of the area and e_{di} is the error at individual level. Random effects represent the unexplained heterogeneity of the values Y_{di} across the areas. Random effects are considered independent of errors, with $u_d \sim^{iid} (0, \sigma_u^2)$ and $e_{di} \sim^{ind} (0, \sigma_e^2 k_{di}^2)$, being k_{di} known constants representing possible heteroscedasticity.

Note that the mean of the area d can be decomposed into the sum of the values observed in the sample and those not sampled, as follows:

$$\bar{Y}_d = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} Y_{di}\right).$$

It is not necessary to predict the values observed in the sample as they are given to us. The BLUP of $\bar{Y}_d$ under the model with nested errors (32) is obtained by simply fitting the model to the sampling data and predicting the values of the out-of-sample variables Y_{di} from the area d, i.e.

$$\tilde{\bar{Y}}_d^{BLUP} = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} \tilde{Y}_{di}^{BLUP}\right), \tag{33}$$

where, assuming the weighted least squares estimator $\widetilde{\boldsymbol{\beta}}$ from $\boldsymbol{\beta}$ under the model (32), the predicted values are

$$\begin{aligned}\tilde{Y}_{di}^{BLUP} &= \boldsymbol{x}_{di}{}'\widetilde{\boldsymbol{\beta}} + \tilde{u}_d,\\ \tilde{u}_d &= \gamma_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}{}'\widetilde{\boldsymbol{\beta}}), \gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/a_{d\cdot}),\end{aligned}$$

where $\bar{y}_{da} = a_{d\cdot}^{-1}\sum_{i\in s_d} a_{di}\, Y_{di}$ and $\bar{\boldsymbol{x}}_{da} = a_{d\cdot}^{-1}\sum_{i\in s_d} a_{di}\,\boldsymbol{x}_{di}$ are the weighted sample means of the response variable and the auxiliary variables, respectively, with weights $a_{di} = k_{di}^{-2}$, and where $a_{d\cdot} = \sum_{i\in s_d} a_{di}$. Once again $\tilde{u}_d$ is the BLUP of u_d and the predicted values $\tilde{Y}_{di}^{BLUP}$ are the BLUPs of the variables Y_{di}, $i \in r_d$, under the model (32).

We construct the vector of response variables for the area d, $\boldsymbol{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$ and the corresponding matrix of auxiliary variables, $\boldsymbol{X}_d = (\boldsymbol{x}_{d1}, \dots, \boldsymbol{x}_{dN_d})'$. Under the nested error model (32), $\boldsymbol{y}_d \sim^{ind} N\,(\boldsymbol{X}_d\boldsymbol{\beta}, \boldsymbol{V}_d)$, $d = 1, \dots, D$, where

$$\boldsymbol{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}_{N_d}' + \sigma_e^2 \boldsymbol{A}_d,$$

where $\boldsymbol{A}_d = \mathrm{diag}(k_{di}^2; i = 1, \dots, N_d)$. We now decompose the vector $\boldsymbol{y}_d$ of the area d into the subvectors for the in-sample units and for the out-of-sample units as follows: $\boldsymbol{y}_d = (\boldsymbol{y}_{ds}{}', \boldsymbol{y}_{dr}{}')'$, and, similarly, the matrices $\boldsymbol{X}_d$ and $\boldsymbol{V}_d$,

$$\boldsymbol{X}_d = \begin{pmatrix}\boldsymbol{X}_{ds}\\ \boldsymbol{X}_{dr}\end{pmatrix}, \quad \boldsymbol{V}_d = \begin{pmatrix}\boldsymbol{V}_{ds} & \boldsymbol{V}_{dsr}\\ \boldsymbol{V}_{drs} & \boldsymbol{V}_{dr}\end{pmatrix}.$$

With this notation, the weighted least squares estimator of $\boldsymbol{\beta}$ is expressed as

$$\widetilde{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \boldsymbol{X}_{ds}\, \boldsymbol{V}_{ds}^{-1}\boldsymbol{X}_{ds}{}'\right)^{-1} \sum_{d=1}^{D} \boldsymbol{X}_{ds}\, \boldsymbol{V}_{ds}^{-1}\boldsymbol{y}_{ds}. \tag{34}$$

For areas with negligible sampling fraction, i.e., where $n_d/N_d \approx 0$, the BLUP of the mean $\bar{Y}_d$ can be written as follows:

$$\tilde{\bar{Y}}_d^{BLUP} \approx \gamma_d\{\bar{y}_{da} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_{da})'\widetilde{\boldsymbol{\beta}}\} + (1 - \gamma_d)\bar{\boldsymbol{X}}_d{}'\widetilde{\boldsymbol{\beta}}.$$

As $\gamma_d \in (0,1)$, he BLUP is a weighted mean between the estimator $\bar{y}_{da} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_{da})'\widetilde{\boldsymbol{\beta}}$, known as the survey regression estimator and the synthetic regression estimator, $\bar{\boldsymbol{X}}_d{}'\widetilde{\boldsymbol{\beta}}$. The survey regression estimator is obtained by adapting the same model (32) but taking the area effects u_d as fixed rather than random. Note, also, that this weighted mean is similar to that obtained using the FH estimator given in (24), but where the survey-regression estimator $\bar{y}_{da} + (\bar{\boldsymbol{X}}_d - \bar{\boldsymbol{x}}_{da})'\widetilde{\boldsymbol{\beta}}$ assumes the role of direct estimator. Indeed, this estimator can be considered as direct, since its variance is $O(n_d^{-1})$; i.e., its variance increases when the sample size of the area n_d becomes small.

In order to interpret this estimator, let us consider, for simplicity, a homoscedastic model; i.e., with $k_{di} = 1$ for all i and d. In this case, you have $\gamma_d = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/n_d)$. For an area with a small sample size n_d , γ_d is close to zero and the BLUP is close to the synthetic regression estimator, which borrows information from the other areas. However, for an area with a large sample size n_d , γ_d is close to one and BLUP is close to the survey regression estimator. Moreover, γ_d also depends on heterogeneity between areas as measured by σ_u^2. If the areas are very heterogeneous (σ_u^2 is large compared to σ_e^2/n_d), or equivalently, if the considered auxiliary variables do not explain much of the variability, then γ_d is close to one and more weight is given to the survey regression estimator, which is similar to a direct estimator. Otherwise, if the areas are homogeneous or, in other words, if the auxiliary variables are strong predictors, then more weight is given to the synthetic estimator obtained by means of regression with these auxiliary variables.

Again, the BLUP given in (33) depends on the true values of the variance components of the model (32), $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)'$. Substituting the true $\boldsymbol{\theta}$ for a consistent estimator $\widehat{\boldsymbol{\theta}} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ in the BLUP (33), we obtain the EBLUP, expressed as

$$\hat{\hat{Y}}_d^{EBLUP} = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} \hat{Y}_{di}^{EBLUP}\right), \tag{35}$$

where, if $\widehat{\boldsymbol{\beta}}$ is the result of substituting $\boldsymbol{\theta}$ by the estimator $\widehat{\boldsymbol{\theta}}$ in $\widetilde{\boldsymbol{\beta}}$ given in (34), the predicted values are now

$$\begin{aligned}\hat{Y}_{di}^{EBLUP} &= \boldsymbol{x}_{di}'\widehat{\boldsymbol{\beta}} + \hat{u}_d,\\ \hat{u}_d &= \hat{\gamma}_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}'\widehat{\boldsymbol{\beta}}), \hat{\gamma}_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2/a_{d\cdot}),\end{aligned}$$

The BLUP is unbiased under the model (32) and is optimal, in that it minimises the MSE, between the linear in-sample and unbiased estimators. By substituting $\boldsymbol{\theta}$ for the estimator $\widehat{\boldsymbol{\theta}}$, the EBLUP remains unbiased under the model (32), under certain conditions for the estimator $\widehat{\boldsymbol{\theta}}$. The usual estimation methods, namely ML, REML and the Henderson III method, satisfy these conditions. However, neither the BLUP nor the EBLUP are unbiased under the sample design. In fact, they do not take account of the sample design and are therefore normally designed for simple random sampling (SRS). In any case, EBLUPs give a marked improvement in efficiency over direct estimators and even over FH estimators, since they use much more detailed information and in a more efficient way (without reducing the data by half). Under sample designs with unequal probabilities, they may have a non-negligible bias under the design. You and Rao (2002) proposed a variation called pseudo EBLUP that includes the sampling weights and is consistent under the design when the area size n_d grows.

For an unsampled area, i.e., with sample size $n_d = 0$, assuming $\gamma_d = 0$, we obtain the synthetic regression estimator $\bar{\boldsymbol{X}}_d{}'\widehat{\boldsymbol{\beta}}$.

Under SRS and assuming $k_{di} = 1$, for all i and d, given that the survey regression estimator is approximately unbiased under the design, the bias under the BLUP design when. $n_d/N_d \approx 0$ is $-(1-\gamma_d)(\bar{Y}_d - \bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta})$. Therefore, the relative absolute bias (RAB). under the design is equal to

$$(1-\gamma_d)\left|\frac{\bar{Y}_d - \bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta}}{\bar{Y}_d}\right| \le \left|\frac{\bar{Y}_d - \bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta}}{\bar{Y}_d}\right|,$$

i.e., it is smaller than the relative absolute bias under the design of the synthetic regression estimator $\bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta}$ for the same vector of coefficients $\boldsymbol{\beta}$, $|(\bar{Y}_d - \bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta})/\bar{Y}_d|$, while $\gamma_d > 0$. If we set an upper limit B

for the relative absolute bias (e.g., $B = 0.20$ or $B = 0.10$), if this limit B is exceeded for any of the areas, we can replace the relative absolute bias of the synthetic estimator by a constant quantity for each area, such as the maximum, i.e., we consider that

$$M = \max_{1 \le d \le D} \left| \frac{\bar{Y}_d - \bar{\boldsymbol{X}}_d{}'\boldsymbol{\beta}}{\bar{Y}_d} \right|.$$

The quantity $(1-\gamma_d)M$ decreases monotonically with the sample size of the area n_d, by means of γ_d. We can find the sample size n_d^* starting from which $(1-\gamma_d)M$ exceeds B. If $M > B$ (otherwise the SAR does not exceed B for any province), the resulting sample size is

$$n_d^* = \frac{\sigma_e^2}{\sigma_u^2}\left(\frac{M}{B} - 1\right).$$

Thus, for areas with sample size $n_d < n_d^*$, the relative absolute bias could exceed the upper limit B and we may decide not to generate estimates for those areas. However, n_d^* depends on certain unknown quantities. Therefore, in practice, we estimate these unknown quantities and obtain an estimated value of n_d^*. An estimator would be

$$\hat{n}_d^* = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_u^2}\left(\frac{\widehat{M}}{B} - 1\right),$$

where

$$\widehat{M} = \max_{1 \le d \le D} \left| \frac{\hat{\bar{Y}}_d^{EBLUP} - \bar{\boldsymbol{X}}_d{}'\widehat{\boldsymbol{\beta}}}{\hat{\bar{Y}}_d^{EBLUP}} \right|,$$

assuming that $\widehat{M} > B$.

The MSE of the EBLUP $\hat{\bar{Y}}_d^{EBLUP}$ of $\bar{Y}_d$, as well as a second-order estimator of this MSE, can be approximated by using a suitable large analytical second-order formula for D in much the same way as the Prasad-Rao formula described in the introduction for the FH estimator. Another option that does not require a large number of areas D, although computationally more expensive, is to turn to bootstrapping procedures. Here we give an overview of a parametric bootstrapping procedure for finite populations proposed by González-Manteiga et al. (2008), particularised here for the estimation of area means, $\bar{Y}_d$. The bootstrapping procedure is as follows:

1. Fit the model (32) to the sampling data $\boldsymbol{y}_s = (\boldsymbol{y}_{1s}{}', \dots, \boldsymbol{y}_{Ds}{}')'$ and obtain the estimators of the parameters of the model $\widehat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$.
2. Generate the effects on the areas as follows: $u_d^{*(b)} \sim^{iid} N\,(0, \hat{\sigma}_u^2)$, $d = 1, \dots, D$.
3. Independently of the effects on the areas $u_d^{*(b)}$, generate bootstrap errors for the sample units in the area, $e_{di}^{*(b)} \sim^{iid} N\,(0, \hat{\sigma}_e^2)$, $i \in s_d$. Also generate the population mean values of the errors in the areas, $\bar{E}_d^{*(b)} \sim^{iid} N\,(0, \hat{\sigma}_e^2/N_d)$, $d = 1, \dots, D$.
4. Calculate the true bootstrap mean values of the areas,
$$\bar{Y}_d^{*(b)} = \bar{\boldsymbol{X}}_d{}'\widehat{\boldsymbol{\beta}} + u_d^{*(b)} + \bar{E}_d^{*(b)}, \quad d = 1, \dots, D.$$
Note that the computation of the mean $\bar{Y}_d^{*(b)}$ does not require the individual values $\boldsymbol{x}_{di}$, for each out-of-sample unit in the area $i \in r_d$.
5. Using the vectors of values of the auxiliary variables for the sampling units $\boldsymbol{x}_{di}$, $i \in s_d$, generate the response variables for the sampling units based on the model
$$Y_{di}^{*(b)} = \boldsymbol{x}_{di}{}'\widehat{\boldsymbol{\beta}} + u_d^{*(b)} + e_{di}^{*(b)}, \quad i \in s_d, \quad d = 1, \dots, D. \qquad (36)$$

6. For the original sample $s = s_1 \cup \cdots \cup s_D$, let $\boldsymbol{y}_s^{*(b)} = ((\boldsymbol{y}_{1s}^{*(b)})', \dots, (\boldsymbol{y}_{Ds}^{*(b)})')'$ be the bootstrap vector of values in the sample. Fit the model (32) to the bootstrap data $\boldsymbol{y}_s^{*(b)}$ and compute the bootstrap EBLUPs $\hat{\bar{Y}}_d^{EBLUP*(b)}$, $d = 1, \dots, D$.
7. Repeat steps 2) - 6) for $b = 1, \dots, B$ and we obtain the true mean values $\bar{Y}_d^{*(b)}$ and the corresponding EBLUPs $\hat{\bar{Y}}_d^{EBLUP*(b)}$ for the bootstrap replication b. The naive bootstrap estimators of the MSE of the EBLUPs $\hat{\bar{Y}}_d^{EBLUP}$, obtained by means of the parametric bootstrapping are

$$mse_B\left(\hat{\bar{Y}}_d^{EBLUP}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{Y}}_d^{EBLUP*(b)} - \bar{Y}_d^{*(b)}\right)^2, \quad d = 1, \dots, D. \qquad (37)$$

The bootstrap estimator (37) is first-order unbiased rather than second-order, i.e., its bias does not decrease faster than D^{-1} when the number of areas D grows. There are various bias corrections in the literature but they either produce estimators that can assume unwanted negative values or else they are strictly positive but not second-order unbiased. Besides, these corrections increase the variance of the MSE estimator. Thus, the naive bootstrap estimator that does not perform bias correction is an acceptable choice from among the non-analytical estimators.

Summary of characteristics of the model-based EBLUP with nested errors:

Target indicators: mean values/totals of the variable of interest.

Data requirements:

- Microdata from the p considered auxiliary variables, from the same survey where the variable of interest is observed.
- Area of interest obtained from the same survey where the variable of interest is observed.
- Population mean values of the p auxiliary variables considered in the areas, $\bar{\boldsymbol{X}}_d$, $d = 1, \dots, D$.

Advantages:

- The number of observations used to fit the model is the total sample size n, much larger than the number of observations (equal to the number of areas) in the FH models. Thus, the model parameters are estimated more efficiently and the improvement in efficiency over the direct estimators will be greater than with FH models.
- The considered regression model incorporates unexplained heterogeneity between areas.
- It is a composite estimator, which automatically borrows information from the remaining areas (giving greater weight to the synthetic regression estimator) where required (when the sample size is small). It tends to the survey regression estimator when the area grows in size.
- Unlike the FH model, no variance needs to be known.
- The MSE estimator under the model is a stable estimator of the MSE under the design and is unbiased under the design when averaged across several areas.
- Estimates can be disaggregated for any required subdomain or subarea within the areas, even at the individual level.
- It can be estimated in unsampled areas.

Disadvantages:

- The estimators are based on a model; thus, it is necessary to analyse the model (e.g., by means of the residuals).
- It does not take the sample design into account. Therefore, it is not unbiased under the design and is more suitable for simple random sampling. It will be affected by informative sample designs.
- It is affected by isolated outlying observations or by a lack of normality.
- Microdata is usually obtained from a census or administrative record, and there are often confidentiality issues that limit the use of this type of data.
- The MSE estimator under the Prasad-Rao model is suitable under the model with normality and is not unbiased under the design for the MSE under the design for a specific area.
- They require readjustment to verify the benchmarking property: that the sum of the estimated totals in the areas of a larger region matches the direct estimator for that area.

Example 6. EBLUPs based on the model with nested errors of poverty incidence, with R. Continuing with the previous examples, we demonstrate how to obtain in R the EBLUPs of poverty incidence based on a model with nested errors. In a predefined dataset in R, the values of the out-of-sample auxiliary variables are available for the five provinces with the smallest sample sizes. Using this data and the sample, we can calculate the population means of these variables for these provinces, but we do not have the true means for the other provinces. Therefore, we can only demonstrate the collection of the EBLUPs for those provinces, even though the model is fitted to the sample with the provinces.

First of all, we load the dataset containing the values of the out-of-sample auxiliary variables for the selected provinces and we calculate the population means of these variables in the provinces. To do so, we use the in-sample values (incomedata dataset) and the out-of-sample values (Xoutsamp). Moreover, we include the provincial codes in the first column of the matrix of the mean values:

```
data(Xoutsamp)
<-length(selprov)                # Number of provinces selected
p<-dim(Xoutsamp)[2]-1            # Number of auxiliary variables

auxvar<-names(Xoutsamp)[-1]      # Names of  aux. var. in Xoutsamp
meanXpop<-matrix(0,nr=I,nc=p)    # Matrix with means of aux. var.
Ni<-numeric(I)                   # Population size of the provinces

for (i in 1:I){                  # Loop for selected provinces
  d<-selprov[i]
  Xsd<-incomedata[prov==d,auxvar]               # Aux. var. sampling values
  Xrd<-Xoutsamp[Xoutsamp$domain==d,-1]          # Non-sample values
  Ni[i]<-dim(Xrd)[1]+dim(Xsd)[1]# Population size of the prov.
  for (k in 1:p){
    meanXpop[i,k]<-(sum(Xrd[,k])+sum(Xsd[,k]))/Ni[i]
  }
}
Xmean<-data.frame(selprov,meanXpop)
```

We now call on the function that calculates the EBLUPs of the poverty incidence for the selected provinces, based on the model with nested errors adapted to the sample data (for all provinces). We save the estimates obtained in a vector:

```
povinc.BHF.res<-eblupBHF(poor ~ age2+age3+age4+age5+nat1+educ1+educ3+labor1+labor2,
 dom=prov,selectdom=selprov,meanxpop=Xmean,popnsize=sizeprov[,-1])

povinc.BHF<-numeric(D)
povinc.BHF[selprov]<-povinc.BHF.res$eblup$eblup
```

We check the results of the model fit with nested errors and compute the synthetic regression estimator based on the individual-level model:

```
betaest<-povinc.BHF.res$fit$fixed     # Regression coefficients
upred<-povinc.BHF.res$fit$random   # Predicted effects on prov.
sigmae2est<-povinc.BHF.res$fit$errorvar  # Estimated var. of the error
sigmau2est<-povinc.BHF.res$fit$refvar   # Estimated variance of the effects of the provinces

povinc.rsyn2<-numeric(D)
povinc.rsyn2[selprov]<-cbind(1,meanXpop)%*%betaest
```

We analyse how much weight the EBLUP gives to the survey regression estimator:

```
gammad.BHF<-sigmau2est/(sigmau2est+sigmae2est/nd)
summary(gammad.BHF)
```

Result:

```
   Min.  1st Qu.   Median   Mean      3rd Qu.  Max.
 0.3458  0.7743   0.8606   0.8352    0.9276   0.9741
```

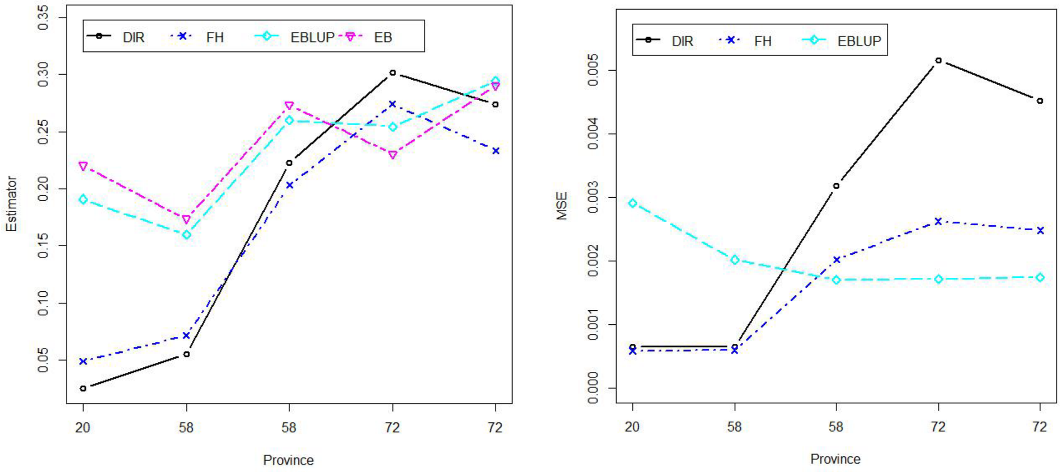
The closer the gammad.BHF result is to zero for an area, the more information is being borrowed from the synthetic regression estimator at individual level. In this case, there is one province for which a lot of information is being borrowed, given that the minimum value of gammad.BHF is relatively small.

We now calculate the MSE estimators of the EBLUPs using the parametric bootstrap described above. To do so, we call on the pbmseBHF() function using B=200 bootstrap replications. This function also returns the EBLUPs and results of the fit exactly the same as the eblupBHF() function.

```
povinc.mse.res<-pbmseBHF(poor~age3+age4+age5+nat1+educ1+educ3+labor1+labor2,
dom=prov,selectdom=selprov,meanxpop=Xmean,popnsize=sizeprov[,-1],B=200)
```

Finally, we compare the EBLUPs based on the model with nested errors with the direct HT and FH estimators, plotting the point estimates obtained and their estimated MSEs for the five provinces selected:

Figure 7
EBLUPs based on the model with nested errors of poverty incidence for the provinces together with direct HT and FH estimates (left), and estimated MSEs from the three estimators (right)
(In proportions)



Source: Prepared by the author.

The R code used to generate the previous figure is as follows:

```
M<-max(povinc.dir[selprov],povinc.FH[selprov],povinc.BHF[selprov])
m<-min(povinc.dir[selprov],povinc.FH[selprov],povinc.BHF[selprov])
plot(1:5,povinc.dir[selprov],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="Estimator",
 xaxt="n")
points(1:5,povinc.dir[selprov],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:5,povinc.FH[selprov],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:5,povinc.BHF[selprov],type="b",col=5,lty=5,pch=5,lwd=2)
axis(1, at=1:5, labels=nd[selprov])
legend(1,M+(M-m)/k,legend=c("DIR","FH","EBLUP"),ncol=3,col=c(1,4,5),lwd=rep(2,3),
 lty=c(1,4,5),pch=c(1,4,5))

M<-max(povinc.dir.var[selprov],povinc.FH.mse[selprov],povinc.BHF.mse[selprov])
m<-min(povinc.dir.var[selprov],povinc.FH.mse[selprov],povinc.BHF.mse[selprov])
plot(1:5,povinc.dir.cv[selprov],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="CV",
 xaxt="n")
points(1:5,povinc.dir.var[selprov],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:5,povinc.FH.mse[selprov],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:5,povinc.BHF.mse[selprov],type="b",col=5,lty=5,pch=5,lwd=2)
axis(1, at=1:5, labels=nd[selprov])
legend(1,M+(M-m)/k,legend=c("DIR", "FH", "EBLUP"),ncol=3,col=c(1,4,5),lwd=rep(2,3),
 lty=c(1,4,5),pch=c(1,4,5))
```

According to Figure 7 (left), we can see how, for the five provinces with the smallest sample size, the FH estimators assume similar values to the direct estimators but are slightly more stable for the 5 selected provinces than the direct and FH estimators. EBLUPs are clearly more stable for the 5 selected provinces than the direct and FH estimators. Moreover, as we can observe in figure 7 (right), the estimated MSEs of the FH estimators are smaller for the provinces on the left, because they borrow more information from the other provinces, since the nested error model is fitted with all the individuals in the sample (from the D=52 provinces). On the other hand, these MSEs grow gradually as the sample size decreases, which makes sense. In contrast, the estimated MSEs of the direct and FH estimators assume extremely small values for the provinces with smaller sample size (which is not very credible). In the case of the direct estimators, their variances are estimated with the limited data from each province and, therefore, these estimated variances (like the MSEs) are not reliable. BLUPs based on the FH model with known parameters have an MSE that cannot exceed the variance of the direct estimators; if these variances are incorrectly estimated, then the MSE of the FH estimator is also incorrect in that case.

C. ELL Method

The Elbers, Lanjouw and Lanjouw (2003) method, which we will call the ELL method, is the method traditionally used by the World Bank to build poverty or inequality maps. This method was the first to appear in the literature that can estimate more complex indicators than means or totals, as long as they are a function of a variable that measures individual purchasing power (usually net disposable income or expenditure). This method assumes the model with nested errors (32) for the log transformation of this variable, where the random effects are for the first-stage units of the sample design (clusters) rather than for the areas of interest. However, to facilitate comparability with the other methods presented in this paper, whilst also simplifying the notation, we will consider the first stage units to be equal to the areas. In this case, if E_{di} is the variable that measures the individual's purchasing power i in the area d, assuming $Y_{di} = \log\left(E_{di} + c\right)$, where $c > 0$ is a constant, the ELL model is

$$Y_{di} = \boldsymbol{x}_{di}'\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \dots, N_d,\ d = 1, \dots, D, \tag{38}$$

where $u_d \sim^{iid}(0, \sigma_u^2)$ and $e_{di} \sim^{ind}(0, \sigma_e^2 k_{di}^2)$, with u_d and e_{di} being independent, and k_{di} known constants representing possible heteroscedasticity.

The ELL estimator of a general parameter $\delta_d = \delta_d(\boldsymbol{y}_d)$ under the model (38) is obtained by means of a bootstrap procedure. This bootstrap procedure provides a numerical approximation of the theoretical ELL estimator, which is expressed as the marginal expectation $\hat{\delta}_d^{ELL} = E[\delta_d]$, unlike the EB predictor considered in Chapter V.B, which conditions the sample $\boldsymbol{y}_s$. The same bootstrap procedure is used to obtain an estimate of the MSE of the ELL estimator.

The bootstrap procedure works as follows. First of all, residuals of the model (38) fitted to the data are used to generate random effects u_d^* for each area $d = 1, \dots, D$, and errors e_{di}^*, for each individual $i = 1, \dots, N_d$, $d = 1, \dots, D$. From these, from the estimator $\widehat{\boldsymbol{\beta}}$ of the regression parameter $\boldsymbol{\beta}$, and using the values of the auxiliary variables for the in-sample and out-of-sample individuals, bootstrap values of the response variable are generated for all the individuals in the population, as follows:

$$Y_{di}^* = \boldsymbol{x}_{di}'\widehat{\boldsymbol{\beta}} + u_d^* + e_{di}^*, \quad i = 1, \dots, N_d, d = 1, \dots, D.$$

This provides us with a census of the response variable, with which any type of indicator can be calculated. This generation process is repeated for $a = 1, \dots, A$, obtaining A complete censuses. For each census a, we compute the indicator of interest $\delta_d^{*(a)} = \delta_d(\boldsymbol{y}_d^{*(a)})$, where $\boldsymbol{y}_d^{*(a)} = (Y_{d1}^{*(a)}, \dots, Y_{dN_d}^{*(a)})'$ are the values of the response variable in the area d in the bootstrap census a. Finally, the ELL estimator is obtained by averaging over the A censuses,

$$\hat{\delta}_d^{ELL} = \frac{1}{A}\sum_{a=1}^{A} \delta_d^{*(a)}.$$

Also, in this method, the MSE is estimated as follows:

$$\mathrm{mse}_{ELL}(\hat{\delta}_d^{ELL}) = \frac{1}{A}\sum_{a=1}^{A}(\,\delta_d^{*(a)} - \hat{\delta}_d^{ELL})^2.$$

To estimate the FGT indicator of order α using this method, we first write this indicator as a function of the response variables of the model $Y_{di} = \log\,(E_{di} + c)$. By substituting $E_{di} = \exp\,(Y_{di}) - c$ in the formula of the FGT indicator given in (1), we obtain:

$$F_{\alpha d} = \frac{1}{N_d}\sum_{i=1}^{N_d}\left(\frac{z + c - exp\,(Y_{di})}{z}\right)^{\alpha} I(\exp\,(Y_{di}) < z + c). \qquad (39)$$

Thus, we calculate this indicator with the values Y_{di}^{*} generated for each census a, as follows:

$$F_{\alpha d}^{*(a)} = \frac{1}{N_d}\sum_{i=1}^{N_d}\left(\frac{z + c - exp\,(Y_{di}^{*(a)})}{z}\right)^{\alpha} I\left(\exp\,(Y_{di}^{*(a)}\right) < z + c),$$

and the ELL estimator of $F_{\alpha d}$ is then approximated by averaging these indicators over the A generated censuses, i.e,

$$\hat{F}_{\alpha d}^{ELL} = \frac{1}{A}\sum_{a=1}^{A} F_{\alpha d}^{*(a)}.$$

Finally, the MSE of the estimator $\hat{F}_{\alpha d}^{ELL}$ is estimated as follows:

$$\mathrm{mse}_{ELL}(\hat{F}_{\alpha d}^{ELL}) = \frac{1}{A}\sum_{a=1}^{A}(\,F_{\alpha d}^{*(a)} - \hat{F}_{\alpha d}^{ELL})^2.$$

It is easy to verify that, for areas of large population size N_d (usually the case in real applications), if we use this method to estimate the mean of the area d, $\bar{Y}_d$, by averaging $\bar{Y}_d^{*(a)} \approx \bar{\boldsymbol{X}}_d{}'\hat{\boldsymbol{\beta}} + u_d^{*(a)}$ over the A censuses, the average of the bootstrap random effects $u_d^{*(a)}$, over the bootstrap repetitions, is $A^{-1}\sum_{a=1}^{A} u_d^{*(a)} \approx E(u_d) = 0$. Therefore, the ELL estimator, $\hat{\bar{Y}}_d^{ELL} = E[\bar{Y}_d]$, turns out to be the synthetic regression estimator,

$$\hat{\bar{Y}}_d^{ELL} = \bar{\boldsymbol{X}}_d{}'\hat{\boldsymbol{\beta}}.$$

This is due to the fact that the marginal mean $E[\delta_d]$, without conditioning the data available from Y_{di} in the sample, does not use these sample observations and therefore sticks to the prediction obtained through the model, without considering the random effects on the areas, as these disappear. Thus, the ELL estimator has the same problems as the synthetic regression estimator; namely, it can be highly biased if the regression model without the random effects is not verified; i.e., if the considered auxiliary variables do not explain all the heterogeneity of the response variable across the areas.

Moreover, in the bootstrap method used, unlike in the usual bootstrap methods, the model is not refitted and estimated with bootstrap samples (which should be drawn from bootstrap censuses). Therefore, you are not replicating the real-world process in the bootstrap world. As a result, the MSE estimated by this method does not correctly reproduce the error incurred in real-world estimation. Finally, in the original ELL method, the random effects included in the model are for the clusters or first-stage sampling units and not for the areas of interest. If this model is considered, but the available auxiliary variables do not account for all the heterogeneity between areas, the error of the ELL estimator may be seriously underestimated.

Summary of the characteristics of the ELL estimator:

Target indicators: general parameters.

Data requirements:

- Microdata from the p considered auxiliary variables, from the same survey where the variable of interest is observed.
- Area of interest obtained from the same survey where the variable of interest is observed.
- Microdata from the considered p auxiliary variables in the areas drawn from a census or administrative record (measured in the same way as in the survey).

Advantages:

- Based on individual-level data, which provides more detailed information than area-level data. Moreover, the sample size is usually much larger (n compared to D).
- Any indicators can be estimated, as long as they are defined as a function of the response variables Y_{di}.
- They are unbiased under the model if the model parameters are known.
- Once the model is fitted, it can be estimated for any subarea or subdomain. It can even be estimated at the individual level.
- Once the model is fitted, all the indicators required (which are a function of Y_{di}) can be estimated at the same time, without the need to fit a different model for each indicator.

Disadvantages:

- ELL estimators may have a high MSE under the model and may even perform worse than direct estimators if the unexplained heterogeneity between areas is significant, see Molina and Rao (2010). For the estimation of means, the ELL estimators are synthetic regression estimators, which assume a model without random effects on the areas.
- They are model-based. It is, therefore, necessary to check that the model fits the data correctly.
- They are not unbiased under the design and may have considerable bias under information design.
- They can be seriously affected by isolated outliers.
- If the model includes cluster effects rather than area of interest effects, but there is heterogeneity across areas, the ELL estimators underestimate the true MSE. Even if area effects are included in the model, the ELL estimators of the MSE do not estimate correctly the true MSE of the ELL estimators for each area.

D. Best empirical predictor under the model with nested errors

The best/Bayes predictor (BP) based on the model with nested errors was proposed by Molina and Rao (2010) to estimate general non-linear indicators. These authors have used it to estimate the poverty incidence and poverty gap in the Spanish provinces by gender. It has also been used by the National Council for the Evaluation of Social Development Policy (*Consejo Nacional para la Evaluación de la Política de Desarrollo Social* (CONEVAL)) in Mexico in comparative studies with other methods, such as the ELL, for the estimation of poverty and inequality indicators in Mexican municipalities. This method assumes that the variables $Y_{di} = \log\,(E_{di} + c)$ follow the model (32) under normality for the random effects on the areas u_d and for the errors e_{di}. Under this model, the vectors of variables for each area, $\boldsymbol{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$, $d = 1, \dots, D$, are independent and verify $\boldsymbol{y}_d \sim^{ind} N\,(\boldsymbol{\mu}_d, \boldsymbol{V}_d)$, with vector of means $\boldsymbol{\mu}_d = \boldsymbol{X}_d\boldsymbol{\beta}$, being $\boldsymbol{X}_d = (\boldsymbol{x}_{d1}, \dots, \boldsymbol{x}_{dN_d})'$ and covariance matrix $\boldsymbol{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}_{N_d}{}' + \sigma_e^2 \boldsymbol{A}_d$, where. $\boldsymbol{A}_d = \text{diag}(k_{di}^2; i = 1, \dots, N_d)$. For a general indicator defined as a function of $\boldsymbol{y}_d$, i.e., $\delta_d = \delta_d(\boldsymbol{y}_d)$ the best predictor is one that minimises the MSE and is expressed as

$$\tilde{\delta}_d^B(\boldsymbol{\theta}) = E_{\boldsymbol{y}_{dr}}[\delta_d(\boldsymbol{y}_d)|\boldsymbol{y}_{ds}; \boldsymbol{\theta}], \tag{40}$$

where the expectation is assumed with respect to the distribution of the out-of-sample vector of values $\boldsymbol{y}_{dr}$ from the domain d given the values in the sample $\boldsymbol{y}_{ds}$. This conditioned distribution depends on the true value of the model parameters for $\boldsymbol{\theta}$. By replacing $\boldsymbol{\theta}$ with a consistent estimator $\widehat{\boldsymbol{\theta}}$ in the best predictor (40), we obtain the so-called empirical best/Bayes(EB) predictor, $\hat{\delta}_d^{EB} = \tilde{\delta}_d^B(\widehat{\boldsymbol{\theta}})$. Once again, the usual estimation methods, which provide consistent estimators even in the absence of normality, are ML and REML, both under normal likelihood, and the Henderson III method.

Under the nested error model (32), the distribution of $\boldsymbol{y}_{dr}|\boldsymbol{y}_{ds}$, needed to calculate the best predictor (40), is obtained as follows. First of all, we decompose the matrices $\boldsymbol{X}_d$ and $\boldsymbol{V}_d$ into the in-sample and out-of-sample parts in a similar way to how we decomposed $\boldsymbol{y}_d$, i.e,

$$\boldsymbol{y}_d = \begin{pmatrix} \boldsymbol{y}_{ds} \\ \boldsymbol{y}_{dr} \end{pmatrix}, \quad \boldsymbol{X}_d = \begin{pmatrix} \boldsymbol{X}_{ds} \\ \boldsymbol{X}_{dr} \end{pmatrix}, \quad \boldsymbol{V}_d = \begin{pmatrix} \boldsymbol{V}_{ds} & \boldsymbol{V}_{dsr} \\ \boldsymbol{V}_{drs} & \boldsymbol{V}_{dr} \end{pmatrix}.$$

Since $\boldsymbol{y}_d$ follows a normal distribution, then the conditioned ones also have normal distribution, i.e,

$$\boldsymbol{y}_{dr}|\boldsymbol{y}_{ds} \overset{ind}{\sim} N\,(\boldsymbol{\mu}_{dr|s}, \boldsymbol{V}_{dr|s}), \quad d = 1, \dots, D, \tag{41}$$

where the vector of conditioned means and the corresponding covariance matrix take the form

$$\boldsymbol{\mu}_{dr|s} = \boldsymbol{X}_{dr}\boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T\boldsymbol{\beta})\mathbf{1}_{N_d - n_d}, \tag{42}$$

$$\boldsymbol{V}_{dr|s} = \sigma_u^2(1 - \gamma_d)\mathbf{1}_{N_d - n_d}\mathbf{1}_{N_d - n_d}^T + \sigma_e^2 \text{diag}_{i \in r_d}(k_{di}^2), \tag{43}$$

where $\mathbf{1}_k$ is a vector of those of size k. Specifically, for the individual $i \in r_d$, we have

$$Y_{di}|\boldsymbol{y}_{ds} \sim N(\mu_{di|s}, \sigma_{di|s}^2), \tag{44}$$

where the conditioned mean and variance are expressed as

$$\mu_{di|s} = \boldsymbol{x}_{di}{}'\boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T\boldsymbol{\beta}), \tag{45}$$

$$\sigma_{di|s}^2 = \sigma_u^2(1 - \gamma_d) + \sigma_e^2 k_{di}^2. \tag{46}$$

If we now wish to estimate the FGT poverty indicator of order α, $\delta_d = F_{\alpha d}$, we first assume that $Y_{di} = \log\,(E_{di} + c)$, for $c > 0$, verifies the model with nested errors. We rewrite the FGT indicator in question as a function of the response variables in the model Y_{di}, i.e., as in (39), and we calculate the expectation that defines the best predictor $\tilde{F}_{\alpha d}^B = E_{\boldsymbol{y}_{dr}}[F_{\alpha d}|\boldsymbol{y}_{ds}; \boldsymbol{\theta}]$. To do so, we separate the sum that defines the FGT indicator given in (1) into the in-sample and the out-of-sample parts and, by inserting the expectation into the sum, we obtain

$$\tilde{F}_{\alpha d}^B(\boldsymbol{\theta}) = \frac{1}{N_d}\left(\sum_{i \in s_d} F_{\alpha,di} + \sum_{i \in r_d} \tilde{F}_{\alpha,di}^B\,(\boldsymbol{\theta})\right), \tag{47}$$

where $\tilde{F}^B_{\alpha,di}(\boldsymbol{\theta}) = E[F_{\alpha,di}|\boldsymbol{y}_{ds};\boldsymbol{\theta}]$ and the expectation is assumed with respect to the distribution of $Y_{di}|\boldsymbol{y}_{ds}$, $i \in r_d$, given in (44)-(46). For $\alpha = 0{,}1$, the expectations are easy to calculate, and are respectively expressed as

$$\tilde{F}^B_{0,di}(\boldsymbol{\theta}) = \Phi(\alpha_{di}), \tag{48}$$

$$\tilde{F}^B_{1,di}(\boldsymbol{\theta}) = \Phi(\alpha_{di})\left\{1 - \frac{1}{z}\left[\exp\left(\mu_{di|s} + \frac{\sigma^2_{di|s}}{2}\right)\frac{\mathit{\Phi}(\alpha_{di} - \sigma_{di|s})}{\mathit{\Phi}(\alpha_{di})} - c\right]\right\}, \tag{49}$$

where $\Phi(\cdot)$ is the distribution function of a standard Normal random variable and $\alpha_{di} = [\log{(z + c)} - \mu_{di|s}]/\sigma_{di|s}$, with $\mu_{di|s}$ and $\sigma^2_{di|s}$ given in (45)-(46).

For more complex $\delta_d = \delta_d(\boldsymbol{y}_d)$ indicators, e.g., FGT indicators for $\alpha \neq 0{,}1$, the expectation defining the best predictor can often not be calculated analytically. In such cases, the best predictor can be approximated empirically using Monte Carlo simulation. The process would be as follows:

1. Obtain an estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\sigma}^2_u, \hat{\sigma}^2_e)'$ from the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2_u, \sigma^2_e)'$ by fitting the model (32) to the data $(\boldsymbol{y}_s, \boldsymbol{X}_s)$.
2. Generate, for $a = 1, \ldots, A$, vectors of response variables for the out-of-sample individuals in the area d, $\boldsymbol{y}^{(a)}_{dr}$, based on the distribution of $\boldsymbol{y}_{dr}|\boldsymbol{y}_{ds}$ given in (41)-(43), with $\boldsymbol{\theta}$ replaced by its estimator $\hat{\boldsymbol{\theta}}$ obtained in (a).
3. Augment the generated vector $\boldsymbol{y}^{(a)}_{dr}$ with the sample data $\boldsymbol{y}_{ds}$ to form a census vector for the area d, $\boldsymbol{y}^{(a)}_{d} = (\boldsymbol{y}'_{ds}, (\boldsymbol{y}^{(a)}_{dr})')'$. Using $\boldsymbol{y}^{(a)}_{d}$, calculate the interest indicator $\delta^{(a)}_d = \delta_d(\boldsymbol{y}^{(a)}_d)$ and repeat for $a = 1, \ldots, A$. The Monte Carlo approximation of the EB predictor of the δ_d indicator is obtained by averaging the indicators over the A simulated censuses, i.e.

$$\hat{\delta}^{EB}_d = \frac{1}{A}\sum_{a=1}^{A}\delta^{(a)}_d. \tag{50}$$

In step (b), we have to simulate A times a vector $\boldsymbol{y}^{(a)}_{dr}$ with multivariate Normal distribution of size $N_d - n_d$, which can be really large (e.g., the size of a province), which can be computationally very difficult or even impossible due to the large size of the multivariate vector to be generated. This can be avoided by noting that the covariance matrix of this vector, $\boldsymbol{V}_{dr|s}$, given in (43), corresponds to the covariance matrix of a random vector $\boldsymbol{y}^{(a)}_{dr}$ generated from the model

$$\boldsymbol{y}^{(a)}_{dr} = \boldsymbol{\mu}_{dr|s} + v^{(a)}_d \boldsymbol{1}_{N_d - n_d} + \boldsymbol{\epsilon}^{(a)}_{dr}, \tag{51}$$

where $v^{(a)}_d$ and $\boldsymbol{\epsilon}^{(a)}_{dr}$ are independent, and verify, respectively

$$v^{(a)}_d \sim N(0, \sigma^2_u(1 - \gamma_d)), \quad \boldsymbol{\epsilon}^{(a)}_{dr} \sim N(\boldsymbol{0}_{N_d - n_d}, \sigma^2_e \mathrm{diag}_{i \in r_d}(k^2_{di})); \tag{52}$$

(see Molina and Rao (2010)). Using the model (51)-(52), instead of generating a multivariate Normal vector $\boldsymbol{y}^{(a)}_{dr}$ of size $N_d - n_d$, it is only necessary to generate the $1 + N_d - n_d$ independent Normal variables $v^{(a)}_d \sim^{ind} N\,(0, \sigma^2_u(1 - \gamma_d))$ and $\epsilon^{(a)}_{di} \sim^{ind} N\,(0, \sigma^2_e k^2_{di})$, for $i \in r_d$. Using the vector $\boldsymbol{y}^{(a)}_{dr}$ generated from the model (51), in step (c) we construct the census vector $\boldsymbol{y}^{(a)}_d = (\boldsymbol{y}_{ds}{}', (\boldsymbol{y}^{(a)}_{dr})')'$ and calculate the indicator of interest $\delta^{(a)}_d = \delta_d(\boldsymbol{y}^{(a)}_d)$.

For an unsampled area d (i.e. with $n_d = 0$), we generate $\boldsymbol{y}^{(a)}_{dr}$ from the model (51) by taking $\gamma_d = 0$ and, as there is no sampling part in this case, the census vector of the area d is equal to the vector generated $\boldsymbol{y}^{(a)}_d = \boldsymbol{y}^{(a)}_{dr}$.

In the case of complex indicators, calculating analytical approximations for the MSE of the corresponding EB predictors is complicated. Molina and Rao (2010) describe a parametric bootstrap method for estimating the MSE based on the bootstrap method for finite populations by. González-Manteiga et al. (2008). This method consists of the following steps:

1. Fit the model (32) to the sample data $\boldsymbol{y}_s = (\boldsymbol{y}_{1s}', \dots, \boldsymbol{y}_{Ds}')'$, obtaining estimates of the model parameters, $\widehat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$.
2. Generate bootstrap effects on the areas as follows:
$$u_d^{*(b)} \overset{iid}{\sim} N\,(0, \hat{\sigma}_u^2), \quad d = 1, \dots, D.$$
3. Generate, independent of $u_1^{*(b)}, \dots, u_D^{*(b)}$, bootstrap errors
$$e_{di}^{*(b)} \overset{iid}{\sim} N\,(0, \hat{\sigma}_e^2), \quad i = 1, \dots, N_d, d = 1, \dots, D$$
4. Generate a bootstrap population (or census) of response variable values by means of the model,
$$Y_{di}^{*(b)} = \boldsymbol{x}_{di}'\widehat{\boldsymbol{\beta}} + u_d^{*(b)} + e_{di}^{*(b)}, \quad i = 1, \dots, N_d, d = 1, \dots, D.$$
5. We define the census vector of response variables of the area d, given by $\boldsymbol{y}_d^{*(b)} = (Y_{d1}^{*(b)}, \dots, Y_{dN_d}^{*(b)})'$. Calculate the indicators of interest from the bootstrap census. $\delta_d^{*(b)} = \delta_d(\boldsymbol{y}_d^{*(b)})$, $d = 1, \dots, D$.
6. For the original sample $s = s_1 \cup \cdots \cup s_D$, let $\boldsymbol{y}_s^{*(b)} = ((\boldsymbol{y}_{1s}^{*(b)})', \dots, (\boldsymbol{y}_{Ds}^{*(b)})')'$ be the vector containing the bootstrap observations whose indices are in the sample, i.e. containing the variables $Y_{di}^{*(b)}$, $i \in s_d$, $d = 1, \dots, D$. Once more, fit the model (32) to the bootstrap data $\boldsymbol{y}_s^{*(b)}$ and obtain the bootstrap EB predictors of the indicators of interest, $\hat{\delta}_d^{EB*(b)}$, $d = 1, \dots, D$.
7. Repeat steps 2) - 6) for $b = 1, \dots, B$, and we obtain the true values, $\delta_d^{*(b)}$, and the corresponding EB predictors, $\hat{\delta}_d^{EB*(b)}$, for each area $d = 1, \dots, D$, and for each bootstrap replication, $b = 1, \dots, B$.
8. The naive bootstrap estimators of the MSE of the EB predictors, $\hat{\delta}_d^{EB}$, are expressed as
$$\mathrm{mse}_B(\hat{\delta}_d^{EB}) = B^{-1}\sum_{b=1}^{B}\left(\hat{\delta}_d^{EB*(b)} - \delta_d^{*(b)}\right)^2, \quad d = 1, \dots, D.$$

Note that, in order to estimate complex indicators, both the ELL method described in the previous chapter and the EB method presented in this chapter require data from a survey with observations of the variable of interest and auxiliary variables for all areas, $\{(y_{di}, \boldsymbol{x}_{di}); i \in s_d, d = 1, \dots, D\}$, as well as a census with the values of the same auxiliary variables for all population units, $\{\boldsymbol{x}_{di}; i = 1, \dots, N_d, d = 1, \dots, D\}$. In principle, the EB method must also identify in the census those units that are also in the sample within each area s_d. Linking survey and census data is not always possible in practice. However, the sample size of the area, n_d, is typically very small compared to the population size of the area, N_d. Next, we can use the Census best predictor proposed by Correa, Molina, and Rao (2012), which is obtained by calculating the conditioned expectations $\tilde{F}_{\alpha,di}^{B}(\boldsymbol{\theta})$, also for the individuals in the sample as if they were not observed, i.e., the Census best predictor of $F_{\alpha d}$ is expressed as

$$\tilde{F}_{\alpha d}^{CB}(\boldsymbol{\theta}) = \frac{1}{N_d}\sum_{i=1}^{N_d}\tilde{F}_{\alpha,di}^{B}\,(\boldsymbol{\theta}). \quad (53)$$

In the same way as the EB predictor, we define the Census EB predictor of $F_{\alpha d}$, replacing in (53) a consistent estimator of $\boldsymbol{\theta}$. If the expectation defining $\tilde{F}_{\alpha,di}^{B}(\boldsymbol{\theta})$ cannot be calculated analytically, as happens when the indicator has a complicated form, in each replication of the Monte Carlo procedure described in (1)-(3), we generate the complete census vector $\boldsymbol{y}_d$ instead of just the vector of out-of-sample observations $\boldsymbol{y}_{dr}$; i.e. we apply the Monte Carlo approximation (50) by generating $\boldsymbol{y}_d^{(a)} = \boldsymbol{\mu}_{d|s} + v_d^{(a)}\mathbf{1}_{N_d-n_d} + \boldsymbol{\epsilon}_d^{(a)}$, where $\boldsymbol{\mu}_{d|s} = \boldsymbol{X}_d\boldsymbol{\beta} + \gamma_d(\bar{y}_{da} - \bar{\boldsymbol{x}}_{da}^T\boldsymbol{\beta})\mathbf{1}_{N_d}$ and $\boldsymbol{\epsilon}_d^{(a)} \sim N(\mathbf{0}_{N_d}, \sigma_e^2 \mathrm{diag}_{i=1,\dots,N_d}(k_{di}^2))$. If the sampling fraction n_d/N_d is negligible, as it usually is in most actual cases, the *Census EB* estimator from $\delta_d = F_{\alpha d}$ will be practically equal to the original EB estimator.

For indicators whose calculation comes with a high computational cost, such as those that require ordering the individuals of the population according to their purchasing power, like the Fuzzy monetary and Fuzzy supplementary indicators, the computational time for the total procedure, including the bootstrap method for the calculation of the MSE, escalates. In this case, Ferretti, and Molina (2012) proposed a variation of the EB predictor, known as fast EB, which is much faster computationally. In the Monte Carlo procedure (1)-(3) for the EB predictor approximation, this procedure replaces the generation of the census in step (2) by the generation of a sample (different in each Monte Carlo replication) and the calculation of the true values of the indicators in step (3) by the calculation of design-based estimators, which only need a sample instead of the full census.

EB predictor properties (approximate for Census EB if n_d/N_d is negligible):

Target indicators: general parameters.

Data requirements:

- Microdata from the p considered auxiliary variables, from the same survey where the variable of interest is observed.
- Area of interest obtained from the same survey where the variable of interest is observed.
- Microdata from the considered p auxiliary variables from a census or administrative record (measured in the same way as in the survey).

Advantages:

- Based on individual-level data, which provides more detailed information than area-level data (it is also possible to incorporate area-level variables). Moreover, the sample size is usually much larger (n compared to D).
- Any indicators can be estimated, as long as they are defined as a function of the response variables Y_{di}.
- They are unbiased under the model if the model parameters are known.
- They are optimal in the sense of minimising the MSE under the model, for known values of the parameters.
- They perform substantially better than the ELL estimators in terms of MSE under the model (32) when the unexplained heterogeneity between areas is significant. For unsampled areas (with $n_d = 0$), the EB and ELL estimators are practically the same. They will also be practically the same, in this case for all the areas, if all the heterogeneity between areas is explained by the auxiliary variables ($\sigma_u^2 = 0$).
- Once the model is fitted, it can be estimated for any subarea or subdomain. It can even be estimated at the individual level.
- Once the model is fitted, all the indicators required (which are a function of Y_{di}) can be estimated at the same time, without the need to fit a different model for each indicator.

Disadvantages:

- They are model-based. Therefore, it is necessary to check that the model fits correctly (e.g., by means of the residuals).
- They do not take the sample design into account. They are not unbiased under the design and may have considerable bias under information design.
- They can be seriously affected by isolated outliers or lack of normality.
- The MSE estimators obtained using the parametric bootstrap method are computationally intensive.

Example 7. EB estimators of poverty incidence, with R. Continuing with the previous examples, we demonstrate how to obtain the EB estimators of poverty incidence in R, based on a model with nested errors for the logarithm of income (transferred with a constant). The poverty line has been calculated in advance as 60% of median income, and it proves to be $z = 6557.143$. Using this line, we need to define the function that gives us the poverty incidence:

```
 povertyincidence <- function(y) {
 result <- mean(y < 6557.143)
 return (result)
}
```

We now call upon the function that calculates the EB estimators by selecting the poverty incidence function as the indicator, taking the logarithm transformation (default), and adding the constant=3500 constant to the income before this transformation, and using replications for the Monte Carlo approximation of the EB estimators. The above-mentioned constant is selected so that the residuals of the fit show an approximately symmetrical distribution, since the EB method described is based on normal distribution. Before calling on the function, we set the random number generator seeds so that the function will give us the same estimates in the event of repeating the call to this function, and we initialize the vector that will contain the EB estimators. :

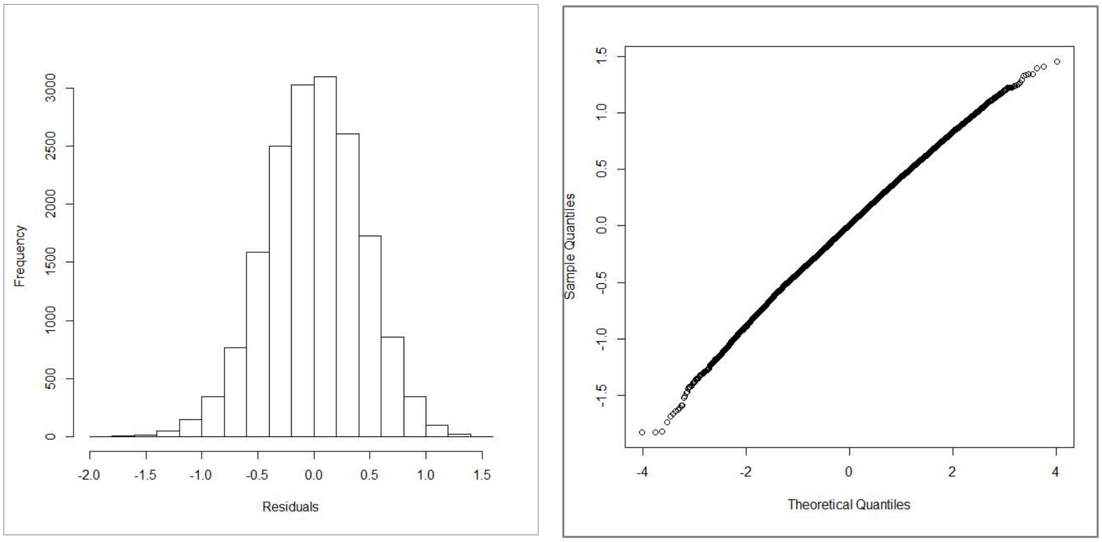
```
povinc.EB<-numeric(D)

set.seed(123)      # We set the seed for random numbers
res.EB<-ebBHF(income~age2+age3+age4+age5+nat1+educ1+educ3+labor1+labor2,dom=prov,
selectdom=selprov,Xnonsample=Xoutsamp,MC=50,constant=3500,indicator=povertyincidence)
povinc.EB[selprov]<-res.EB$eb$eb$eb
```

For any model, the residuals should be analysed to check that the data doesn't present clear evidence counter to the assumed model. Since the EB method requires normality, we plot a histogram and a q-q plot of normality of the residuals:

```
resid.EB<-res.EB$fit$residuals
hist(resid.EB,main="",xlab="Residuals")
qqnorm(resid.EB,main="")
```

Both charts (figure 8) show that the distribution of the residuals is approximately normal. In contrast, if we fit the model to income without the log transformation, both the histogram and the q-q normality plot (not included for purposes of conciseness) show a markedly skewed distribution to the right. This transformation is, therefore, necessary in order not to move away from the normality hypothesis.

Figure 8
Histogram (left) and q-q plot of normality (right) of the residuals from the model fit with errors nested to the logarithm of income
(In units)



Source: Prepared by the author.

Finally, we calculate the bootstrap MSE estimators of the EB estimators with B=200 bootstrap replications and MC=50 replications for the Monte Carlo approximation of the EB estimators.

```
set.seed(123)
povinc.mse.res<-
pbmseebBHF(income~age2+age3+age4+age5+nat1+educ1+educ3+labor1+labor2,dom=prov,selectdom
=selprov,Xnonsample=Xoutsamp,B=200,MC=50,constant=3500,
 indicator=povertyincidence)

povinc.eb.mse<-numeric(D)
povinc.eb.mse[selprov]<-povinc.mse.res$mse$mse
```

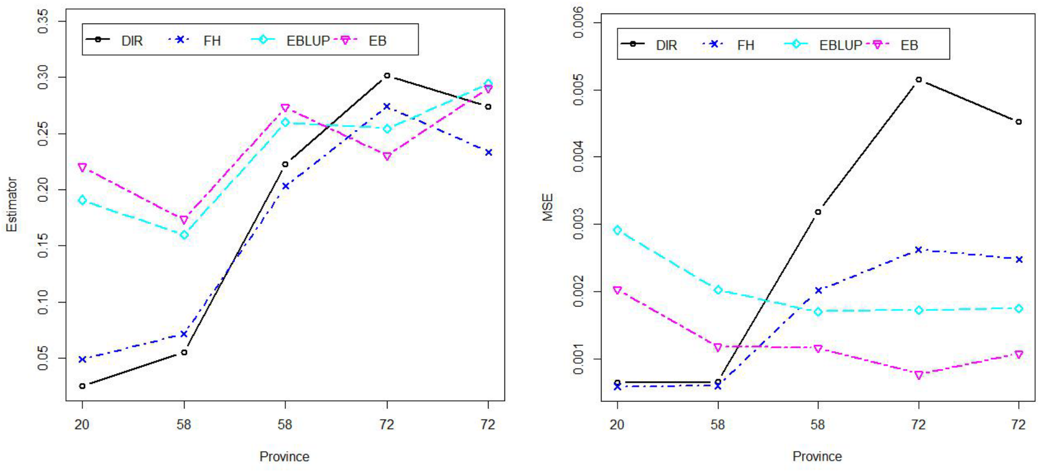
Finally, we graphically compare the EB estimators with the HT, FH and EBLUP direct estimators based on the model with nested errors of the poverty incidence for the selected provinces:

```
k<-6
M<-max(povinc.dir[selprov],povinc.FH [selprov],povinc.BHF [selprov],povinc.EB [selprov])
m<-min(povinc.dir[selprov],povinc.FH [selprov],povinc.BHF [selprov],povinc.EB [selprov])
plot(1:5,povinc.dir[selprov],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="Estimator",
 xaxt="n")
points(1:5,povinc.dir[selprov],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:5,povinc.FH[selprov],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:5,povinc.BHF[selprov],type="b",col=5,lty=5,pch=5,lwd=2)
points(1:5,povinc.EB[selprov],type="b",col=6,lty=6,pch=6,lwd=2)
axis(1, at=1:5, labels=nd[selprov])
legend(1,M+(M-m)/k,legend=c("DIR", "FH", "EBLUP", "EB"),ncol=4,col=c(1,4,5,6),lwd=rep(2,4),
lty=c(1,4,5.6),pch=c(1,4,5.6))
M<-max(povinc.dir.var[selprov],povinc.FH.mse[selprov],povinc.BHF.mse[selprov],
 povinc.eb.mse[selprov])
m<-min(povinc.dir.var[selprov],povinc.FH.mse[selprov],povinc.BHF.mse[selprov],
 povinc.eb.mse[selprov])
plot(1:5,povinc.dir.var[selprov],type="n",ylim=c(m,M+(M-m)/k),xlab="Province",ylab="CV",
 xaxt="n")
points(1:5,povinc.dir.var[selprov],type="b",col=1,lty=1,pch=1,lwd=2)
points(1:5,povinc.FH.mse[selprov],type="b",col=4,lty=4,pch=4,lwd=2)
points(1:5,povinc.BHF.mse[selprov],type="b",col=5,lty=5,pch=5,lwd=2)
points(1:5,povinc.eb.mse[selprov],type="b",col=6,lty=6,pch=6,lwd=2)
axis(1, at=1:5, labels=nd[selprov])
legend(1,M+(M-m)/k,legend=c("DIR", "FH", "EBLUP", "EB"),ncol=4,col=c(1,4,5,6),lwd=rep(2,4),
 lty=c(1,4,5.6),pch=c(1,4,5.6))
```

According to figure 9 (left), the EB estimators are very similar to the EBLUPs. This is reasonable since both are based on an individual-level model, although the EB estimators fit the model for the logarithm of income, while the EBLUPs fit the model to the binary indicator of either having or not having income below the (variable poor) threshold.

Theoretically, the model assumed by the EBLUPs is not true, since the response variable is binary, and the predictors can provide values outside the interval. Moreover, despite the similarity between the EB and EBLUP estimates, Figure 9 (right) indicates that the EB estimators are more efficient than the EBLUPs.

Figure 9
EB and EBLUP estimates based on the nested error model, FH, and HT direct (left), and MSEs of estimators (right) for the selected provinces
(In proportions)



Source: Prepared by the author.

E. Hierarchical Bayes method under the nested error model

The calculation of EB (or Census-EB) estimators together with their estimated MSEs is computationally intensive and may not be practical for very large populations or for very complex indicators (e.g., those that need to be ordered). Note that to obtain the Monte Carlo approximation of the EB estimator, it is necessary to construct A censuses $\boldsymbol{y}^{(a)}$, $a = 1, \dots, A$, which can be very large.

Moreover, when using bootstrap to estimate the MSE, the Monte Carlo approximation must be repeated for each bootstrap replication. In order to develop a computationally more efficient method, Molina, Nandram and Rao (2014) proposed the hierarchical Bayes (HB) method for estimating general indicators. This procedure does not require the use of bootstrap methods for estimating the MSE as it provides samples from the posterior distribution, from which posterior variances that assume the role of MSE, or any other summary measure, can be easily obtained.

The HB method is based on reparameterising the model with nested errors (32) in terms of the intraclass correlation coefficient $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$ and considering prior distributions for the model parameters $(\boldsymbol{\beta}, \rho, \sigma_e^2)$ that reflect the lack of prior information about them. Specifically, we consider the following HB model:

$$\begin{aligned}
&\text{(i) } Y_{di}|u_d, \boldsymbol{\beta}, \sigma_e^2 \overset{ind}{\sim} N\,(\boldsymbol{x}_{di}{}'\boldsymbol{\beta} + u_d, \sigma_e^2 k_{di}^2), \quad i = 1, \dots, N_d,\\
&\text{(ii) } u_d|\rho, \sigma_e^2 \overset{iid}{\sim} N\left(0, \frac{\rho}{1-\rho}\sigma_e^2\right), \quad d = 1, \dots, D,\\
&\text{(iii) } \pi(\boldsymbol{\beta}, \rho, \sigma_e^2) \propto \frac{1}{\sigma_e^2}, \quad \epsilon \le \rho \le 1-\epsilon,\ \sigma_e^2 > 0, \boldsymbol{\beta} \in R^p,
\end{aligned}$$

where $\epsilon > 0$ is selected very small to reflect lack of prior information. (See the application produced by Molina, Nandram, and Rao (2014), where inference is not sensitive to small changes of ϵ.)

The posterior distribution of the model parameters can be calculated based on the conditioned distributions using the chain rules as follows. Firstly, note that, with the HB method, the random effects $\boldsymbol{u} = (u_1, \dots, u_D)'$ are considered as additional parameters. Then, the joint density of the vector of the parameters $\boldsymbol{\theta} = (\boldsymbol{u}', \boldsymbol{\beta}', \sigma_e^2, \rho)'$ given the observations of the sample $\boldsymbol{y}_s$ is expressed as

$$\pi(\boldsymbol{u}, \boldsymbol{\beta}, \sigma_e^2, \rho|\boldsymbol{y}_s) = \pi_1(\boldsymbol{u}|\boldsymbol{\beta}, \sigma_e^2, \rho, \boldsymbol{y}_s)\pi_2(\boldsymbol{\beta}|\sigma_e^2, \rho, \boldsymbol{y}_s)\pi_3(\sigma_e^2|\rho, \boldsymbol{y}_s)\pi_4(\rho|\boldsymbol{y}_s), \tag{54}$$

where all conditioned densities except π_4 have known forms. Since ρ is defined in a closed interval within $(0{,}1)$, we can generate values of π_4 using a grid method. For more details see Molina, Nandram and Rao (2014)). Thus, samples of $\boldsymbol{\theta} = (\boldsymbol{u}', \boldsymbol{\beta}', \sigma_e^2, \rho)'$ can be generated directly from the posterior distribution given in (54), without the need to use Markov Chain Monte Carlo (MCMC) methods. Under general conditions, an independent posterior distribution can be ensured.

Given $\boldsymbol{\theta}$, under the HB model (i)-(iii), the variables Y_{di} for all the individuals in the population are independent and verify

$$Y_{di}|\boldsymbol{\theta} \overset{ind}{\sim} N\,(\boldsymbol{x}_{di}{}'\boldsymbol{\beta} + u_d, \sigma_e^2 k_{di}^2), \quad i = 1, \dots, N_d,\ d = 1, \dots, D. \tag{55}$$

The predictive density of $\boldsymbol{y}_{dr}$ is expressed as

$$f(\boldsymbol{y}_{dr}|\boldsymbol{y}_s) = \int \prod_{i \in r_d} f\,(Y_{di}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)d\boldsymbol{\theta},$$

where $\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)$ is given in (54). Finally, the HB estimator of the parameter $\delta_d = \delta_d(\boldsymbol{y}_d)$ is

$$\hat{\delta}_d^{HB} = E_{\boldsymbol{y}_{dr}}(\delta_d|\boldsymbol{y}_s) = \int \delta_d\,(\boldsymbol{y}_d)f(\boldsymbol{y}_{dr}|\boldsymbol{y}_s)d\boldsymbol{y}_{dr}. \tag{56}$$

This estimator can be approximated using Monte Carlo simulation. To do so, we generate samples of the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)$ as follows. First, we generate a value $\rho^{(a)}$ from $\pi_4(\rho|\boldsymbol{y}_s)$ using a grid method (see Molina, Nandram, & Rao, 2014); then, we generate $\sigma_e^{2(a)}$ from $\pi_3(\sigma_e^2|\rho^{(a)}, \boldsymbol{y}_s)$; next, $\boldsymbol{\beta}^{(a)}$ is generated from $\pi_2(\boldsymbol{\beta}|\sigma_e^{2(a)}, \rho^{(a)}, \boldsymbol{y}_s)$ and, finally, $\boldsymbol{u}^{(a)}$ is generated from $\pi_1(\boldsymbol{u}|\boldsymbol{\beta}^{(a)}, \sigma_e^{2(a)}, \rho^{(a)}, \boldsymbol{y}_s)$. This process is repeated several A times, in order to obtain a random sample $\boldsymbol{\theta}^{(a)}$, $a = 1, \dots, A$, from $\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)$. For each generated value $\boldsymbol{\theta}^{(a)}$ of $\pi(\boldsymbol{\theta}|\boldsymbol{y}_s)$, we generate the out-of-sample values $\{Y_{di}^{(a)}, i \in r_d\}$ from the distribution given in (55) obtaining, for each area d, the vector of out-of-sample variables $\boldsymbol{y}_{dr}^{(a)}$. By joining it to the data vector in the sample $\boldsymbol{y}_{ds}$, we construct the census vector $\boldsymbol{y}_d^{(a)} = (\boldsymbol{y}_{ds}', (\boldsymbol{y}_{dr}^{(a)})')'$. Now, using $\boldsymbol{y}_d^{(a)}$, we calculate the indicator in question $\delta_d^{(a)} = \delta_d(\boldsymbol{y}_d^{(a)})$, and repeat for $a = 1, \dots, A$. Finally, the HB estimator of δ_d is the posterior mean, which is approximated as follows:

$$\hat{\delta}_d^{HB} = E_{\boldsymbol{y}_{dr}}(\delta_d | \boldsymbol{y}_s) \approx \frac{1}{A} \sum_{a=1}^{A} \delta_d^{(a)}. \tag{57}$$

Since there are no sample observations for unsampled areas ($n_d = 0$), we have $\boldsymbol{y}_{dr}^{(a)} = \boldsymbol{y}_d^{(a)}$, and we therefore generate the complete census vector $\boldsymbol{y}_d^{(a)} = (Y_{d1}^{(a)}, \dots, Y_{dN_d}^{(a)})'$ from the distribution (55).

As a measure of estimation error of the HB estimator, $\hat{\delta}_d^{HB}$, the approximate posterior variance is provided in a similar way,

$$V(\delta_d | \boldsymbol{y}_s) \approx \frac{1}{A} \sum_{a=1}^{A} \left(\delta_d^{(a)} - \hat{\delta}_d^{HB}\right)^2. \tag{58}$$

In the specific case of the FGT indicator of order α, $\delta_d = F_{\alpha d}$, in the Monte Carlo run a, we calculate $F_{\alpha d}^{(a)}$ using $\boldsymbol{y}_d^{(a)}$ applying (39) and the HB estimator is

$$\hat{F}_{\alpha d}^{HB} \approx \frac{1}{A} \sum_{a=1}^{A} F_{\alpha d}^{(a)}. \tag{59}$$

As with the ELL and EB methods, if one wishes to estimate a non-linear indicator, this method requires the availability, in addition to the survey data from which the values of the variable of interest are extracted, of a census or administrative record from which to obtain the microdata of the auxiliary variables. If it is not possible to identify survey individuals in the census or record, a Census HB estimator can be calculated in a similar way to Census EB. In this estimator, even if there were values in the sample $\boldsymbol{y}_{ds}$, these would be ignored and the complete census vector $\boldsymbol{y}_d^{(a)}$would be generated, by generating each value $Y_{di}^{(a)}$ of (55) and the procedure would be the same as if the area were not sampled.

Summary of the HB estimator based on the model with nested errors:

Target indicators: general parameters.

Data requirements:

- Microdata from the p considered auxiliary variables, from the same survey where the variable of interest is observed.
- Area of interest obtained from the same survey where the variable of interest is observed.
- Microdata from the considered p auxiliary variables from a census or administrative record (measured in the same way as in the survey).

Advantages:

- Based on individual-level data, which provides more detailed information than area-level data (it is also possible to incorporate area-level variables). Moreover, the sample size is usually much larger (n compared to D).
- Any indicators can be estimated, as long as they are defined as a function of the response variables Y_{di}.

- They are unbiased under the model if the model parameters are known.
- They are optimal in that they minimise the posterior variance.
- In our simulation studies, they prove to be practically equal to the EB estimators.
- Once the model is fitted, it can be estimated for any subarea or subdomain. It can even be estimated at the individual level.
- Once the model is fitted, all the indicators required (which are a function of Y_{di}) can be estimated at the same time, without the need to fit a different model for each indicator.
- Unlike most Bayesian procedures, the proposed HB method does not require the use of MCMC methods and therefore does not require the convergence of the Markov chains to be monitored.
- Bootstrap methods are not required for MSE estimation. Therefore, the total computational time can be significantly less than in the EB + bootstrap method.
- The calculation of credible intervals or any other summary of the posterior distribution is automatic.

Disadvantages:

- They are model-based. It is, therefore, necessary to check that the model fits correctly (e.g., through predictive or cross-validation residuals (see Molina, Nandram, & Rao, 2014)).
- They do not take the sample design into account. They are not unbiased under the design and may have considerable bias under information design.
- They can be seriously affected by isolated outliers or non-normality.
- The HB method cannot be directly extended to more complex models without losing some of the advantages mentioned above, such as avoiding the application of MCMC methods.

F. Methods based on generalised linear mixed models

Access to certain educational or health services, or the availability of certain housing amenities, are usually measured in a particular area in terms of the proportion of people in that area who may or may not have access to the service or amenity in question. The linear mixed models considered so far do not provide predictions in the natural space $[0,1]$ where these proportions are. Generalised linear mixed models (GLMM) are generally used to obtain predictions in this space. If $Y_{di} \in \{0,1\}$ is the binary variable that measures the lack or otherwise of the service or amenity in question, the most usual estimation model in small areas is the GLMM with random effects in the areas, given by

$$Y_{di}|v_d \sim \mathrm{Bern}(p_{di}), g(p_{di}) = \boldsymbol{x}_{di}'\boldsymbol{\alpha} + v_d, v_d \overset{iid}{\sim} N\,(0, \sigma_v^2), i = 1, \dots, N_d, d = 1, \dots, D, \quad (60)$$

where v_d is the effect of the area d, $\boldsymbol{\alpha}$ is the vector of regression coefficients and g: $(0,1) \rightarrow R$ is the link function (bijective, with continuous derivative). In particular, the logistical link given by. $g(p) = \log\,(p/(1-p))$ is probably the most widely used in practice.

As discussed above, the best predictor under the model (which minimises the MSE under the model) of the ratio $P_d = \bar{Y}_d$, is expressed as

$$\tilde{P}_d^B(\boldsymbol{\theta}) = E(P_d|\boldsymbol{y}_{ds}; \boldsymbol{\theta}) = \frac{1}{N_d}\left\{\sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} E\,(Y_{di}|\boldsymbol{y}_{ds}; \boldsymbol{\theta})\right\}. \quad (61)$$

The distribution of $Y_{di}|\boldsymbol{y}_{ds}$ depends on the vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \sigma_v^2)'$ of parameters of the model. In practice, we obtain the EB predictor by replacing $\boldsymbol{\theta}$ with a consistent estimator $\widehat{\boldsymbol{\theta}}$ in the best predictor, i.e., $\hat{P}_d^{EB} = \tilde{P}_d^{B}(\widehat{\boldsymbol{\theta}})$.

The estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\alpha}}', \hat{\sigma}_v^2)$ of $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \sigma_v^2)$ is obtained by fitting the GLMM model given in (60) to the sample data $\boldsymbol{y}_s = (\boldsymbol{y}_{1s}', \ldots, \boldsymbol{y}_{Ds}')'$. If you want to fit the model using the maximum likelihood method, you need to maximise the likelihood given by $f(\boldsymbol{y}_s) = \int_{R^D} f\,(\boldsymbol{y}_s|\boldsymbol{v}) f(\boldsymbol{v}) d\boldsymbol{v}$, where $\boldsymbol{v} = (v_1, \ldots, v_D)'$. Under the GLMM mentioned above, such a likelihood has no explicit form. For this fitting method it is therefore necessary to use approximations of the integral (e.g., numerical) together with numerical maximisation techniques. Once the model has been fitted, we need to calculate the expectations $E(Y_{di}|\boldsymbol{y}_{ds}; \widehat{\boldsymbol{\theta}})$ that define the EB predictor. One way to approximate this expectation would be to use Bayes' Theorem and the fact that, given v_d, the variables $\{Y_{di}; i = 1, \ldots, N_d\}$ are all independent. In this case, such an expectation can be expressed as follows:

$$E(Y_{di}|\boldsymbol{y}_{ds}; \widehat{\boldsymbol{\theta}}) = \frac{E\{h(\boldsymbol{x}_{di}'\boldsymbol{\alpha} + v_d) f(\boldsymbol{y}_{ds}|v_d); \widehat{\boldsymbol{\theta}}\}}{E\{f(\boldsymbol{y}_{ds}|v_d); \widehat{\boldsymbol{\theta}}\}}, \quad i \in r_d, \tag{62}$$

where $h = g^{-1}$ is the inverse link and

$$\begin{aligned} f(\boldsymbol{y}_{ds}|v_d) &= \prod_{i \in s_d} p_{di}^{Y_{di}} (1 - p_{di})^{(1-Y_{di})} \\ &= \prod_{i \in s_d} h\,(\boldsymbol{x}_{di}'\boldsymbol{\alpha} + v_d)^{Y_{di}} \{1 - h(\boldsymbol{x}_{di}'\boldsymbol{\alpha} + v_d)\}^{(1-Y_{di})}. \end{aligned} \tag{63}$$

For the logistic link, the inverse link is:

$h(\boldsymbol{x}_{di}'\boldsymbol{\alpha} + v_d) = \exp\,(\boldsymbol{x}_{di}'\boldsymbol{\alpha} + v_d)/\{1 + \exp\,(\boldsymbol{x}_{di}'\boldsymbol{\alpha} + v_d)\}$. Using (63), we can approximate the two expectations which appear in (62) by means of Monte Carlo simulation, generating $v_d^{(r)} \sim N(0, \hat{\sigma}_v^2)$, $r = 1, \ldots, R$, and then calculating

$$E(Y_{di}|\boldsymbol{y}_{ds}; \widehat{\boldsymbol{\theta}}) \approx \frac{R^{-1}\sum_{r=1}^{R} h\,(\boldsymbol{x}_{di}'\widehat{\boldsymbol{\alpha}} + v_d^{(r)}) \hat{f}(\boldsymbol{y}_{ds}|v_d^{(r)})}{R^{-1}\sum_{r=1}^{R} \hat{f}\,(\boldsymbol{y}_{ds}|v_d^{(r)})}, \quad i \in r_d, \tag{64}$$

where $\hat{f}$ is the conditioned density $f(\boldsymbol{y}_{ds}|v_d)$, with $\boldsymbol{\alpha}$ replaced by $\widehat{\boldsymbol{\alpha}}$.

The best predictor (61) has minimal MSE and is unbiased under the model (60). However, fitting the GLMM and calculating the Monte Carlo approximation of $\hat{P}_d^{EB}$ as described above, requires significant computational time. Estimating the MSE of the EB predictors using a resampling procedure increases the computational time, making it impractical for very large populations. Moreover, when estimating the parameters of the model $\boldsymbol{\theta}$ and replacing the estimators in order to obtain the empirical version of the best (EB) predictor, we lose the unbiasedness.

There are simple estimators that, although not optimal, are very similar to the optimal estimators under certain conditions and can be obtained directly from the output of the usual software for GLMM fitting. When estimating a ratio, if $\widehat{\boldsymbol{\alpha}}$ and $\hat{v}_d$ are the estimators of $\boldsymbol{\alpha}$ and v_d returned by the software, a plug-in estimator can be calculated by simply predicting the out-of-sample values by means of the model, i.e., assuming

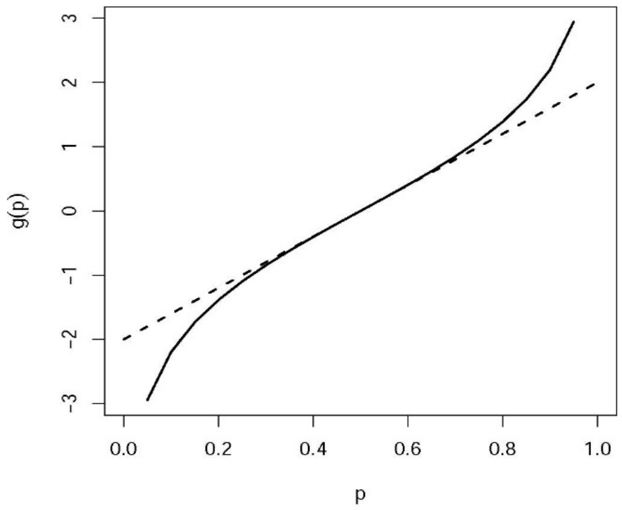
$$\hat{P}_d^{PI} = \frac{1}{N_d}\left(\sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} \hat{p}_{di}\right), \tag{65}$$

where $\hat{p}_{di} = h(\boldsymbol{x}_{di}'\widehat{\boldsymbol{\alpha}} + \hat{v}_d)$ is the predicted value of the out-of-sample observation Y_{di}, $i \in r_d$. Where $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \sigma_v^2)$ is known, the plug-in estimator, $\hat{P}_d^{PI}$, cannot have lower MSE than the best predictor $\hat{P}_d^{B}$.

In fact, unlike the best predictor, the plug-in estimator is not unbiased unless the link function is linear. However, the plug-in estimator is much easier to calculate. The two estimators match when the link function $g(\cdot)$ is linear. In the case of the logistic link $g(p) = \log\left(p/(1-p)\right)$, this is approximately linear for $p \in (0.2,0.8)$ as shown in Figure 10. This approximate linearity of $g(p)$ for central values of p leads us to think that the plug-in estimator (65) based on the model with logistic link should be very similar to the EB predictor, $\hat{P}_d^{EB}$, in terms of MSE, at least for not very extreme values of p. Moreover, this approximate linearity for central values of p also makes both the EB and plug-in estimators of the ratio $P_d = \bar{Y}_d$, resemble the EBLUP, $\hat{P}_d^{EBLUP} = \hat{\bar{Y}}_d^{EBLUP}$, based on the model with nested errors described in Chapter I. This means that, for estimating proportions of individuals with neither too few nor extremely frequent characteristics, it also makes sense to use the EBLUP.

Both the EB and plug-in methods, based on non-linear models such as the GLMM given in (60), even estimating mean values $\bar{Y}_d$, need to have the values of the auxiliary variables for all individuals (microdata), obtained from a census or an administrative record. This is required for calculating the expectation $E(Y_{di}|\boldsymbol{y}_{ds};\boldsymbol{\theta})$ in the case of the EB predictor, or to predict the probability $\hat{p}_{di}$ in the case of the plug-in estimator. However, in addition to the survey data, the EBLUP of $\bar{Y}_d$ only requires the population means of these variables in the areas. Such aggregated data is generally available without confidentiality restrictions.

Figure 10
Logistic link



Source: Prepared by the author.

In principle, the GLMM given in (60) could be used to estimate poverty incidence (FGT indicator of order $\alpha = 0$), F_{0d}. For the poverty gap (FGT indicator with $\alpha = 1$), F_{1d}, it would not make sense to use it because they are not ratios, since the individual values $F_{1,di}$ are not binary variables. In the case of poverty incidence, taking $Y_{di} = I(E_{di} < z)$as a binary response variable, we obtain $P_d = F_{\alpha d}$. The resulting best predictor assumes the expression (47) from Section V.D, but the expectation appearing in the second term would be with respect to the conditioned distribution under the model(60) and would have to be approximated numerically; e.g., as in (64) since, in this case, the conditioned distributions $Y_{di}|\boldsymbol{y}_{ds}$ do not have a known form. As mentioned, the plug-in estimator (65) would have a lower computational cost. Again, if the survey units in the census or register cannot be identified, a Census EB estimator can be used to substitute the observations of the sample in the predictor (61) Y_{di}, $i \in s_d$, with predictions obtained as in (62), or using $\hat{p}_{di}$ as a prediction in the case of the Census plug-in estimator.

The MSE of the corresponding predictor (whether EB or plug-in) can be estimated using a bootstrap procedure as follows (see González-Manteiga et al., 2007):

1. Fit the GLMM given in (60) to the sample data s, obtaining estimators $\hat{\sigma}_v^2$ and $\hat{\boldsymbol{\alpha}}$ of the parameters of the model.
2. Generate bootstrap random effects
$$v_d^{*(b)} \overset{iid}{\sim} N\,(0, \hat{\sigma}_v^2), \quad d = 1, \dots, D.$$
3. Generate a bootstrap census $\boldsymbol{y}_d^{*(b)} = (Y_{d1}, \dots, Y_{dN_d})'$, as follows:
$$Y_{di}^{*(b)} \overset{ind}{\sim} \mathrm{B\,ern}(p_{di}^{*(b)}), p_{di}^{*(b)} = h(\boldsymbol{x}_{di}'\hat{\boldsymbol{\alpha}} + v_d^{*(b)}), i = 1, \dots, N_d, d = 1, \dots, D, \qquad (66)$$
and calculate the true values of the indicators $P_d^{*(b)} = \bar{Y}_d^{*(b)}$, $d = 1, \dots, D$.
4. For each area $d = 1, \dots, D$, extract the sample elements of that area from the bootstrap census $\boldsymbol{y}_d^{*(b)}$, Y_{di}, $i \in s_d^{*(b)}$, constructing the vector $\boldsymbol{y}_{ds}^{*(b)}$. Let $\boldsymbol{y}_s^{*(b)} = ((\boldsymbol{y}_{1s}^{*(b)})', \dots, (\boldsymbol{y}_{Ds}^{*(b)})')'$ be the vector with the sample values of all the areas, with $s = s_1 \cup \cdots \cup s_D$ being the original sample.
5. Fit the model (60) to the bootstrap data $\boldsymbol{y}_s^{*(b)}$ and calculate the bootstrap predictors $\hat{P}_d^{EB*(b)}$, $d = 1, \dots, D$.
6. Repeat steps 2) - 5), for $b = 1, \dots, B$. The bootstrap estimator of the MSE of the predictor $\hat{P}_d^{EB}$ is expressed as
$$mse_B(\hat{P}_d^{EB}) = B^{-1} \sum_{b=1}^{B} (\,\hat{P}_d^{EB*(b)} - P_d^{*(b)})^2.$$

Summary of characteristics of the GLMM-based EB/plug-in predictor compared to methods applicable to mean estimation:

Target indicators: Proportions or totals of a binary variable (e.g., lack or otherwise of a certain commodity or service).

Data requirements:

- Microdata from the p considered auxiliary variables, from the same survey where the variable of interest is observed.
- Area of interest obtained from the same survey where the variable of interest is observed.
- Microdata from the considered p auxiliary variables from a census or administrative record (measured in the same way as in the survey).

Advantages:

- The number of observations used to fit the model is the total sample size n, much larger than the number of areas in the FH models. The model parameters are, therefore, estimated very efficiently and the improvements in efficiency over direct estimators will be greater than with FH models.
- The considered regression model incorporates unexplained heterogeneity between areas.
- Unlike the FH model, no variance needs to be known.

- The MSE estimator under the model obtained (e.g., by means of bootstrap procedures) is a stable estimator of the MSE under the design and is unbiased under the design when averaged over many areas.
- Estimates can be disaggregated for any required subdomain or subarea within the areas, even at the individual level.
- It can be estimated in unsampled areas.

Disadvantages:

- They are based on a model, and it is therefore necessary to analyse this model (e.g., through the residuals).
- It does not take the sample design into account. Therefore, it is not unbiased under the design and is more suitable for simple random sampling. It will be affected by informative sample designs.
- Microdata is usually obtained from a census or administrative record, and there are often confidentiality issues that limit the use of this type of data.
- The estimator of the MSE under the model obtained (e.g., by means of bootstrap procedures) is correct under the considered model and is not unbiased under the design for the MSE under the design for a given area.
- The EB predictor (unlike the plug-in estimator) has a high computational cost.
- The MSE of the EB predictor obtained (e.g., by means of a bootstrap procedure) has an excessively high computational cost and may not be practical for very large populations. This cost is significantly lower for the plug-in predictor.
- They require refitting to verify the benchmarking property: that the sum of the estimated totals in the areas of a larger region matches the direct estimator for that area.

VI. Application: estimating average income and poverty rates in Montevideo

In this chapter we are going to use some of the techniques described above to estimate average incomes and incidence of non-extreme poverty for the census tracts and for both genders in Montevideo, Uruguay. To this end, we will use data from the Continuous Household Survey (*Encuesta Continua de Hogares* or ECH) and the Population Census, both from 2011. This application is for illustrative purposes only and can probably be improved by carrying out a more intensive search for auxiliary information. Therefore, the results obtained in this application should not be considered as definitive estimates.

Since the parameters of the models to be considered may depend on gender, for each type of estimator we will fit separate models for each gender. Specifically, we will compute direct estimates using the ECH microdata for each tract and gender, FH estimates based on the basic area-level model (21), using certain population totals extracted from the census as auxiliary information for each tract and gender and, finally, Census EB estimates based on the basic individual-level model (38) for the logarithm of income, using census microdata from some variables also measured in the ECH. Note that even if only the mean income, which is a linear parameter in the income of individuals in the area, were estimated, when performing a non-linear (logarithmic) transformation of the response variable in the model with nested errors (38), the target parameter, written as a function of the values of the model's response variable, is a non-linear parameter in the values of the model's response variable. Thus, in this case, the EBLUP does not make sense since it is a linear estimator in the values of the model response variable in the sample and we need to resort to the EB methodology. Additionally, since individuals from the ECH are not identified in the census, we consider the Census EB estimator. In addition to the point estimators, estimates of the MSEs of each estimator will be obtained. Calculations have been performed using the R sae packages (Molina and Marhuenda, 2015) and lme4 (Bates et al. 2015).

The population sizes according to the complete census questionnaire (for residents in private dwellings) of the 2011 Population Census in Montevideo are $\mathrm{N} = 656{,}162$ for females and $\mathrm{N} = 566{,}698$

for males. The ECH sample sizes, after discounting missing data, are $n = 26{,}233$ for females and $n = 22{,}464$ for males. For the $D = 25$ tracts that appear in the census, the sample sizes vary between 56 and 3482 for females and between 65 and 2820 for males. Although these are not excessively small sample sizes, we will see that small area estimation techniques can still provide more accurate estimates, by measuring such accuracy in terms of mean squared error. It should also be borne in mind that, according to the available data, poverty incidence in Montevideo is relatively low and, in order to estimate these numbers accurately using direct estimators, the sample sizes needed by tract and gender must be larger than for estimating proportions close to 0.5 or means of continuous variables, such as mean income. In fact, even if the sample size is not excessively small, the direct estimator may be equal to zero due to the fact that no individuals are obtained with incomes below the poverty line. The non-extreme poverty threshold for urban areas in 2011 is 3,182 Uruguayan pesos.

For both mean income $\bar{E}_d = N_d^{-1} \sum_{i=1}^{N_d} E_{di}$ and poverty incidence $F_{0d} = N_d^{-1} \sum_{i=1}^{N_d} I\,(E_{di} < z)$ for each census tract and gender, the corresponding direct estimators, $\hat{\bar{E}}_d^{DIR}$ and $\hat{F}_{0d}^{DIR}$, and their estimated sample variances $\widehat{\mathrm{var}}_\pi(\hat{\bar{E}}_d^{DIR})$ and $\widehat{\mathrm{var}}_\pi(\hat{F}_{0d}^{DIR})$ are obtained using the ECH microdata in formulae 4) - 6). This is provided by the direct() function of the sae package, by introducing the sampling weights of the ECH. For population sizes of the census tracts, N_d, we use the sizes obtained from the Census.

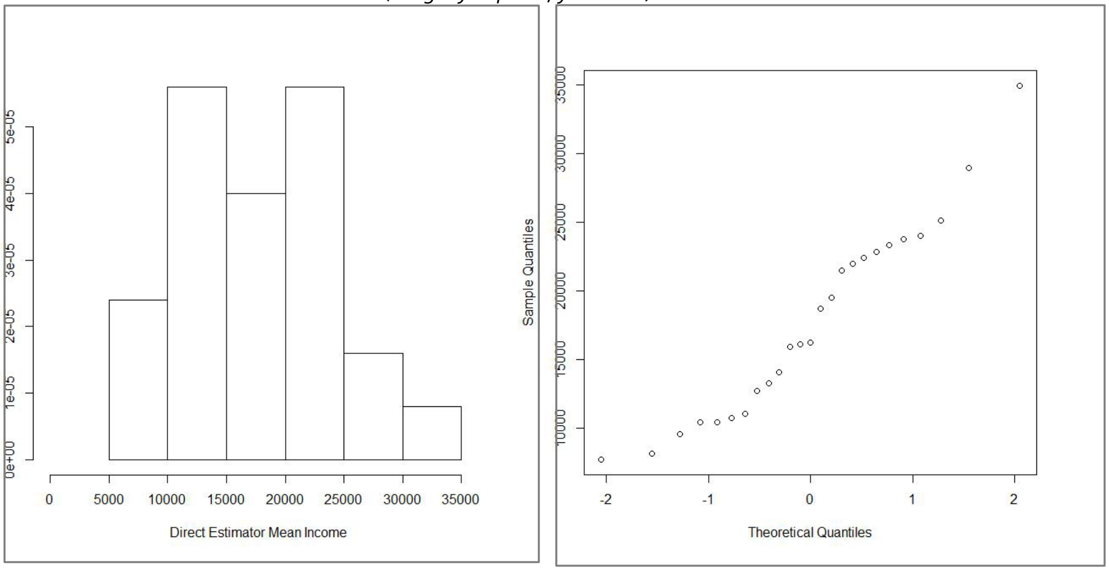
The FH estimators and their estimated mean square errors are obtained from the model (21) for $\delta_d = \bar{E}_d$ or $\delta_d = F_{0d}$. In the case of average income, $\delta_d = \bar{E}_d$, for both genders, we consider as auxiliary variables aggregated at the census tract level (components of $\boldsymbol{x}_d$ in the model), the census ratios of literate individuals, unemployed (but not retired) individuals, average age, and average years in education. For the incidence of poverty, $\delta_d = F_{0d}$, only the ratios of literate individuals and unemployed individuals are significant. FH estimators can be obtained using the eblupFH() function of the sae package which implements the formula given in (24). As vector of response variables of the model, the vector is established of previously obtained direct estimations $\hat{\bar{E}}_d^{DIR}$ or $\hat{F}_{0d}^{DIR}$ as the case may be and, as variances ψ_d, the estimations of the sample variances $\widehat{\mathrm{var}}_\pi(\hat{\bar{E}}_d^{DIR})$ or $\widehat{\mathrm{var}}_\pi(\hat{F}_{0d}^{DIR})$. The estimated MSEs, $\mathrm{mse}_{PR}(\hat{\delta}_d^{FH})$, are obtained by using the analytical formulae in Section V.A, for the REML method fit, and in R they are obtained using the mseFH() function from the previous package.

The FH estimators of poverty incidence can assume the value zero (as can the direct estimators) in those domains where there are no individuals with incomes below the poverty line. Moreover, the MSEs estimated by means of the above analytical formula also assume the value zero. In these cases, we consider such FH estimates to be unreliable and, instead, we calculate synthetic estimators $\hat{\delta}_d^{FH} = \boldsymbol{x}_d{}'\hat{\boldsymbol{\beta}}$. Their MSEs are obtained using the formula (6.2.14) of Rao and Molina (2015), replacing the REML estimator of the variance of the domain effects.

Although the EBLUP based on the Fay-Herriot model does not require normality, the analytical approximation of the MSE obtained in this way does require normality. As we can see in the histogram and the q-q plot of normality for females (figure 11), the distribution of the direct estimators of mean income for the D=25 census tracts does not conform to a normal distribution, but it is not too far away either, bearing in mind that the number of observations used to construct the histogram (D=25) is small. For males, the graphs are similar. This is not the case for the direct estimators of non-extreme incidence of poverty (see figure 12). It should, therefore, be borne in mind that the estimated MSEs of these poverty incidences may not correspond to the facts.

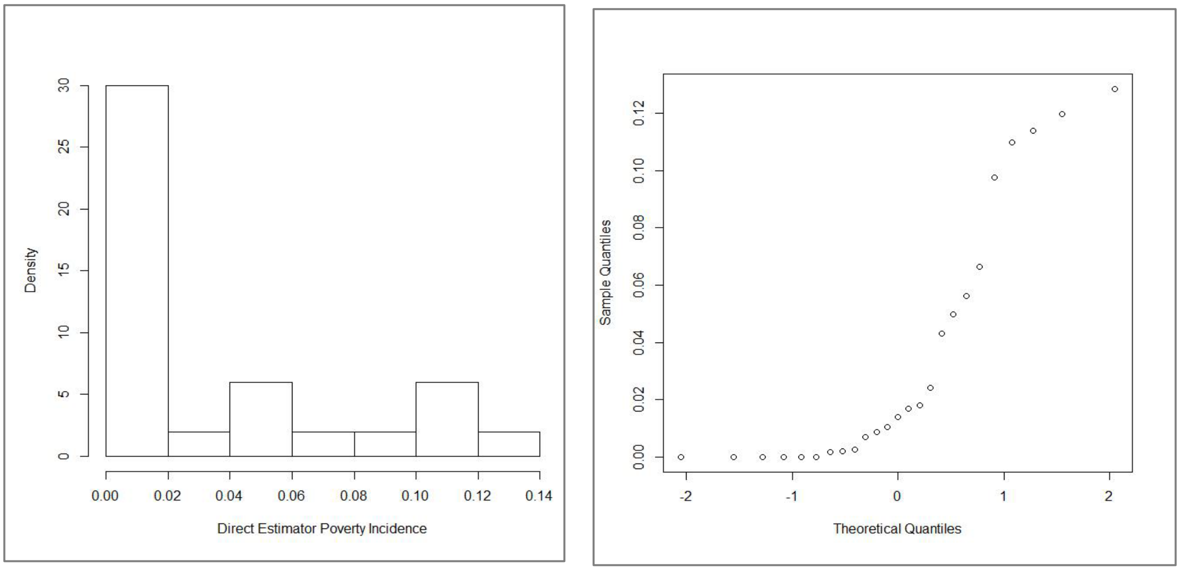
Finally, we obtain the Census EB estimators based on the individual-level model (38), using as response variable $\log(\text{income} + 1000)$, where the addition of constant 1000 to income has been determined so that the histogram of the residuals of the fitted model is approximately symmetrical (see the histogram of the original income and of the income transformed in this way in Figure 13). As auxiliary variables at the individual level in $\boldsymbol{x}_{di}$, we consider the indicators of activity status categories, age, and years in education. Figure 14 for females shows an approximately linear increasing relationship between transformed earnings and age or years in education. The graph for males is similar. Since the transformation of earnings is monotonic, this relationship indicates that, as age or years in education increase, so do earnings.

Figure 11
Histogram (left) and q-q normality plot (right) of the direct estimators of mean income for the D=25 Montevideo census tracts, for females
(Uruguayan pesos, year 2011)



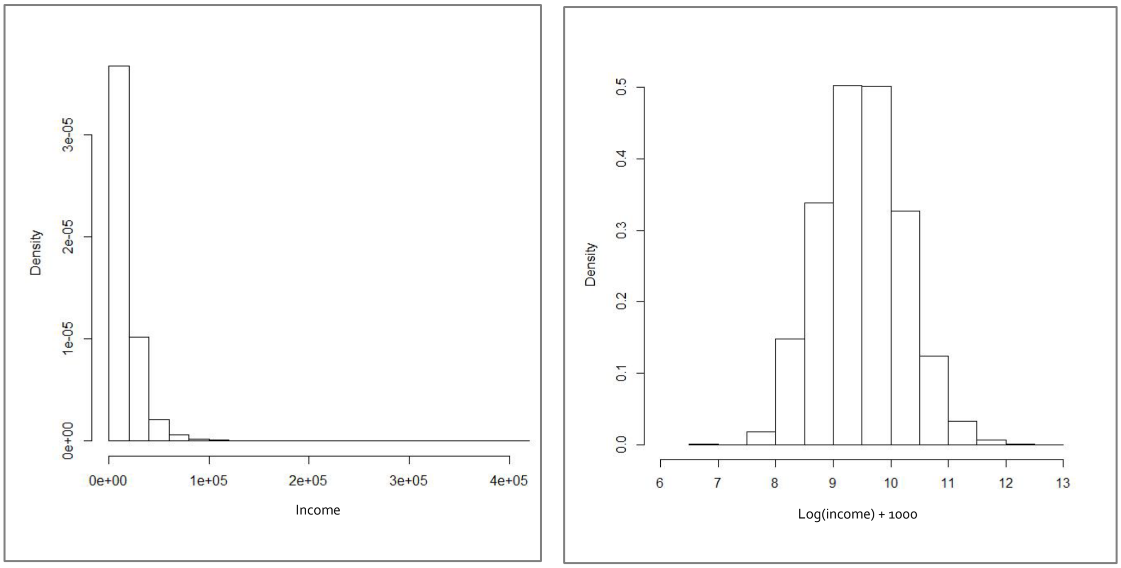
Source: Prepared by the author.

Figure 12
Histogram (left) and q-q normality plot (right) of the direct estimators of the non-extreme poverty incidence for the D=25 Montevideo census tracts, for females
(In proportions)



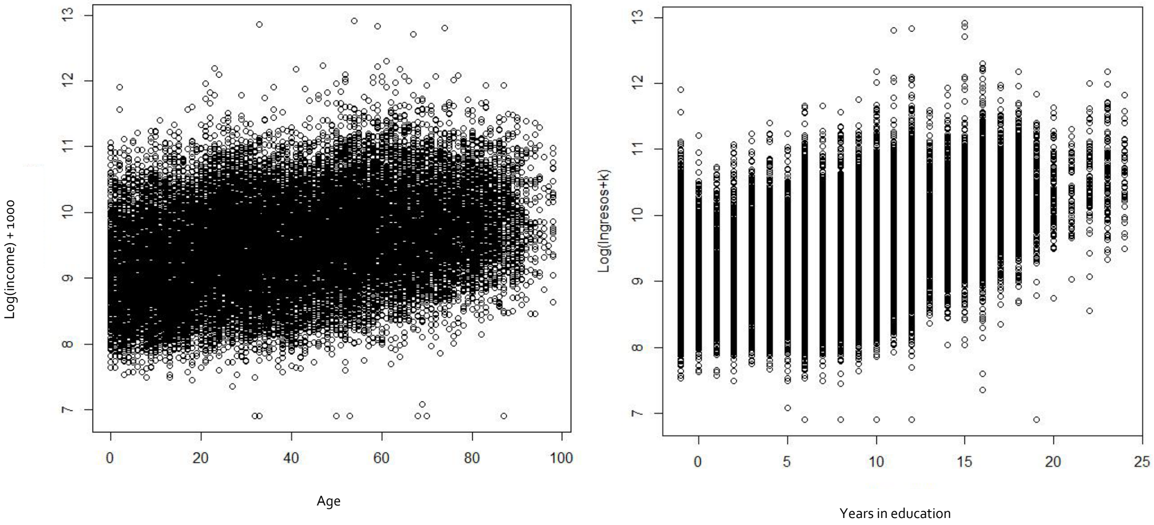
Source: Prepared by the author.

Figure 13
Histogram of untransformed (left) and log (income + 1000) transformed (right) income for females
(Uruguayan pesos, year 2011)



Source: Prepared by the author.

Figure 14
Transformed income in comparison with age (left) and in comparison with years in education (right), for females
(Uruguayan pesos – logarithmic transformation -, year 2011)



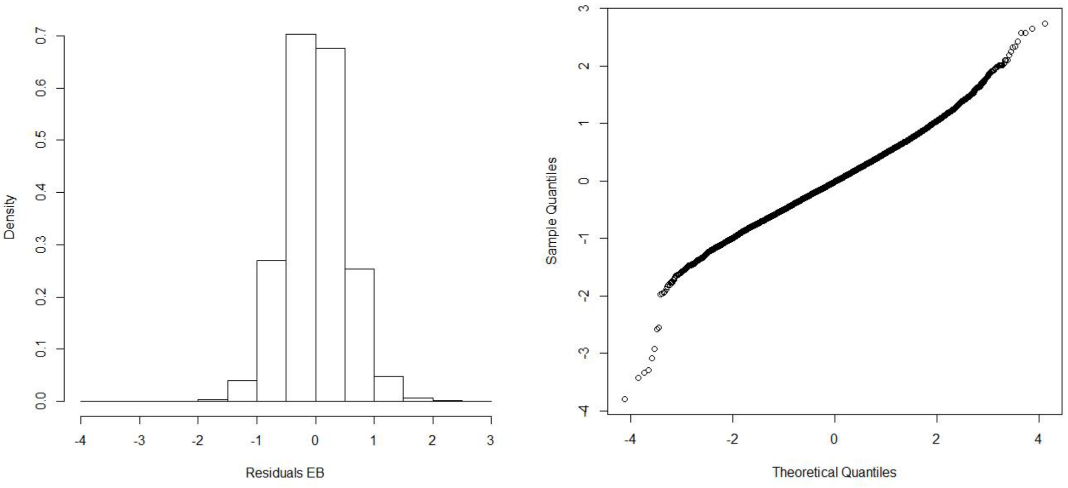
Source: Prepared by the author.

The Census EB estimators of poverty incidence $F_{\alpha d}$ based on the model with nested errors for transformed income are calculated using formulae (53) and (48), replacing $\boldsymbol{\theta}$ by the estimator $\widehat{\boldsymbol{\theta}}$; in this case, we have used the REML estimator. Although, as seen in Example 7, the ebBHF() function of the sae package provides the EB estimators but no t the Census EB estimators, if the sample fractions of the areas are small, we can use the same function to obtain approximate values of the Census EB estimators, setting the Xnonsample attribute of that function (matrix of values of the auxiliary variables for the out-of-sample part of the population) equal to the matrix with the census microdata of those variables for all individuals in the tracts considered. In this case, it can be seen that the Census EB estimates and those obtained in this way show really small differences.

The same fitted model is used to obtain the Census EB estimators of average earnings. The Census EB estimators of average earnings $\delta_d = \bar{E}_d$ based on this model are obtained in a similar way. Specifically, they are obtained as follows: $\hat{\bar{E}}_d^{CEB} = N_d^{-1}\sum_{i=1}^{N_d}\hat{E}_{di}^{CEB}$ where, taking into account that income E_{di} is obtained as a function of the response variables in the model Y_{di} as follows: $E_{di} = \exp\,(Y_{di}) + 1000$, then $\hat{E}_{di}^{CEB} = E[\exp\,(Y_{di})|\boldsymbol{y}_s; \hat{\boldsymbol{\theta}}] + 1000$. This expectation can be obtained using the Monte Carlo approximation (50) and implemented in the ebBHF() function, or by means of the analytical formula given in Molina and Martín (2018). In this case, this analytical formula has been used because it has practically no computational cost.

The estimated MSEs of the Census EB estimators are obtained by means of a slight modification of the bootstrap procedure described in Chapter V (originally designed for the EB estimators), using. $B = 500$ bootstrap replications. The difference between the EB and Census EB estimators lies in the fact that ECH units cannot be identified in the census. Therefore, in each bootstrap replication, we cannot generate census vectors $\boldsymbol{y}_d^{*(b)}$, $d = 1, \dots, D$, and take the part of the sample from them $\boldsymbol{y}_s^{*(b)}$. In the case of the Census EB estimators, we generate the bootstrap censuses $\boldsymbol{y}_d^{*(b)}$, $d = 1, \dots, D$, using the values of the auxiliary variables of the census and, on the other hand, we generate the bootstrap sample vector $\boldsymbol{y}_s^{*(b)}$ using the values of the same auxiliary variables, but taken from the ECH. The true bootstrap values of the parameters are obtained from the generated bootstrap censuses, $\delta_d^{(b)} = \delta_d(\boldsymbol{y}_d^{*(b)})$, $d = 1, \dots, D$.

Figure 15
Histogram (left) and q-q normality plot (right) of the residuals of the model with nested errors for transformed income, for females
(In proportions)

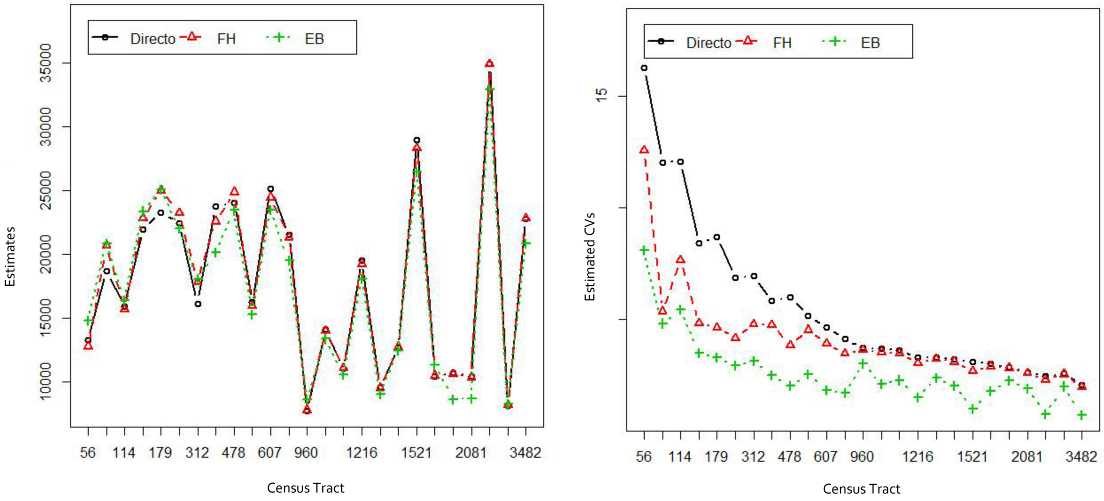


Source: Prepared by the author.

The EB (or Census EB) method and the bootstrap procedure used are based on the normality hypothesis; it is, therefore, in this case, crucial to check if this hypothesis is verified, at least approximately. Figure 15 shows the histogram and q-q normality plot of the residuals from the model fit for transformed earnings for females. Although real data hardly fits a model exactly, and any test will reject the null hypothesis of normality if the sample size is as large as it is in this case, we can see from these figures that the distribution is not too far from normal. If one wishes to use a distribution that fits income somewhat better, one can use the EB method based on a multivariate GB2 model such as the one proposed by Graf, Marín and Molina (2018).

The detailed numerical results for each census tract are shown in tables 1-4 in the annex. Next, we analyse these results graphically and comment on the results obtained for the different estimators. Figure 16 shows the values obtained from the direct, FH and Census EB estimators, of mean income (left), and the estimated CVs of these estimators (right) for each census tract, for Females. Census tracts (axis x) are ordered from smallest to largest sample size and their sample sizes have been indicated on the axis labels x. We can see how the three estimators assume similar values, although the direct and FH estimators obtain practically the same values in this case. This is due to the fact that, when estimating average income, the sample sizes of the tracts are not excessively small, and the weight given by the FH estimators to the corresponding direct estimators is close to one. This is an advantage of estimators based on models with random effects. However, and although the sample sizes are moderate, as we can see in the chart on the right, the Census EB estimators are clearly more efficient than the direct and FH estimators for all census tracts. This is because they use a greater amount of information: the microdata from the census. For males (figure 17), we can draw similar conclusions.

Figure 16
Direct, FH and Census EB estimates (left) of average income, and CVs of the estimators (right) for the D=25 census tracts of Montevideo, for females
(Uruguayan pesos, year 2011)



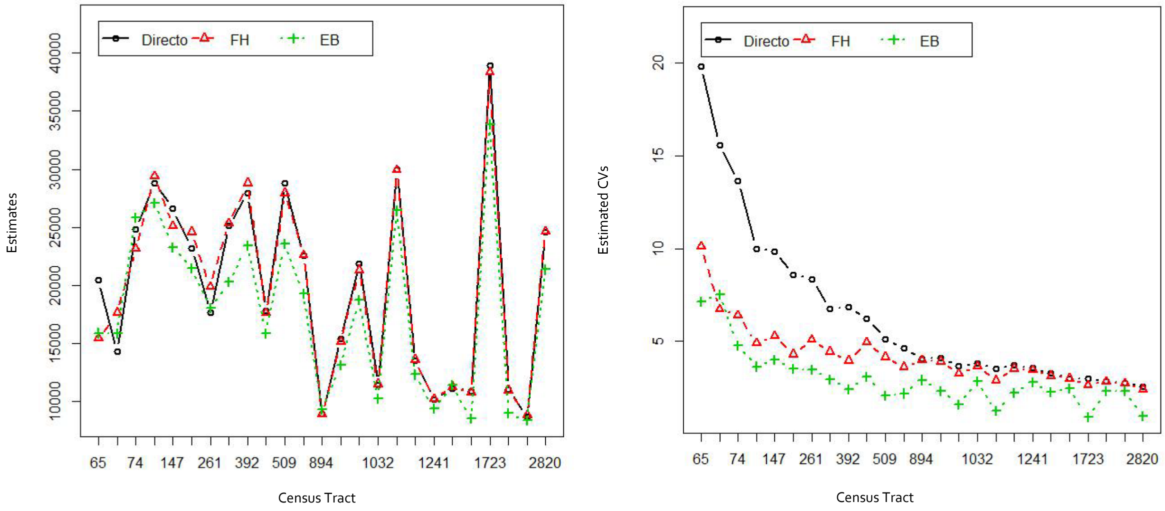
Source: Prepared by the author.
Note: Census tracts (axis x) ordered from smallest to largest sample size, with sample sizes shown on the axis.

For the poverty incidence, the estimates and mean squared errors for females and males are shown in figures 18 and 19 respectively. In this case we show MSEs instead of CVs because, in the case of ratios, for a fixed sample size, CVs increase as the ratio decreases; therefore, CVs are less meaningful as measures of estimation error, especially when the estimated proportions assume very small values, as is the case here. Once again, the values of the three estimators are similar for all census tracts except for those with the smallest sample size. In fact, in these tracts, the direct estimators for Females assume the (implausible) value of zero because there are no sampled individuals with incomes below the threshold. In fact, the estimated variances of the direct estimators also assume the value zero for these tracts. Note that the estimated variances of the direct estimators are also based on the few observations sampled for each tract and gender. If we consider the direct estimators to be unreliable, their estimated variances are also unreliable. For domains with direct estimators equal to zero, the FH estimators and their MSEs are also theoretically zero. In such cases, as mentioned above, the synthetic estimators obtained from the same model have been used. We can observe in the figures on the right how the MSEs of the direct and FH estimators show large fluctuations. Note that the MSEs of the FH estimators are especially large for the domains where synthetic estimators have been used. In contrast, the MSEs of the EB estimators increase gently in relation to the sample size of the census tract. In addition to taking more reasonable values, the estimated MSEs of the EB estimators remain below the MSEs of the other two estimators for most census tracts.

Figure 17

Direct, FH and Census EB estimates (left) of average income, and CVs of the estimators (right) for the D=25 census tracts of Mondevideo, for males

(Uruguayan pesos, year 2011)



Source: Prepared by the author.

Note: Census tracts (axis x) ordered from smallest to largest sample size, with sample sizes indicated on the axis.

It should be stressed that model-based estimators can even provide estimates for unsampled areas, although this is not recommended since it is not possible to analyse the goodness-of-fit of the model for these areas. And, on the subject of goodness of fit, as stated above, for the incidence of non-extreme poverty, the normality hypothesis in the Fay-Herriot model is not verified. This is because the sample sizes are small for some of the tracts and the true poverty incidences appear to be quite small, with the result that the direct estimators have a markedly skewed distribution and the Central Limit Theorem is not verified. Although normality is not a requirement for obtaining the FH estimator, it is assumed for the estimation of the MSE using the analytical formulae considered in Section V.A and provided by the mseFH() function of the sae package. In fact, an additional drawback of the estimators obtained from this Fay-Herriot model is that they can result in negative values or values greater than one which, when it comes to ratios, is not suitable. A simple solution is to truncate the estimates to zero when they are negative and one when they exceed this value. Another possibility is to consider the regression model (19) for a bijective transformation of poverty incidence, $g(F_{0d})$, which translates values from the space $[0,1]$ to real values. However, the same transformation of the direct estimator, $g(\hat{F}_{0d}^{DIR})$, need not be an unbiased estimator for $g(F_{0d})$ and, therefore, the model (20) is not verified for $g(\hat{F}_{0d}^{DIR})$. In this case, the FH model for $g(\hat{F}_{0d}^{DIR})$ would have an additional bias, unless one considers the model (20) for $\hat{F}_{0d}^{DIR}$ together with the regression model above for $g(F_{0d})$. In this case, the two models considered cannot be summarised in a linear mixed model such as that given in (21), i.e., they are unmatched models. Estimators based on mismatched models of this type have been obtained by You and Rao (2002b) based on Bayesian inference.

As we have seen, when we have auxiliary information at the individual level, the improved efficiency of the estimators that use this information is usually greater. However, both data sources from the same year have been used in this application. In those years where an updated census is not available, estimators based on individual-level models may provide somewhat biased estimates. In these cases, therefore, it is advisable to look for other sources of current data, such as administrative records. When there are no updated sources of data at the individual level, it is recommended to keep to area-level models. In some cases, aggregate data sources can be found at a lower level than the area. In that case, models for aggregate data could be used at that level, including two-fold subarea level models (see Torabi and Rao (2014)).

VII. Conclusions

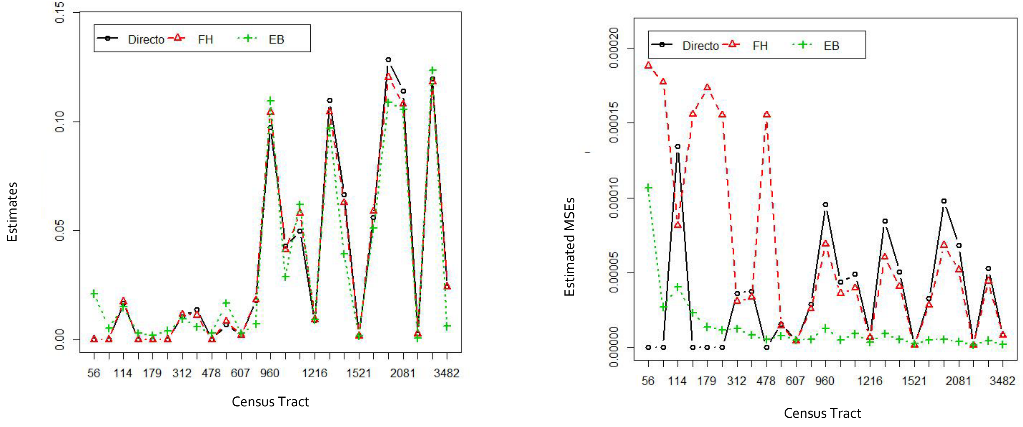
This paper has dealt with the problem of disaggregating statistical estimates in population areas or subgroups. Recommendations are given on the limitations of disaggregation of direct estimates and there is a description of basic indirect methods, as well as some more sophisticated ones, that can overcome these limitations. As has been seen throughout this paper, the methods to be used in each specific application depend mainly on the form of the indicator in question and on the type of auxiliary information available, as there are no universal methods that can be used for any available type of indicator or information. Thus, in each case, a study must be made of the potentially applicable methods, depending on the data requirements and assumptions that each method assumes. In applications that allow the use of various methods, the accuracy of the final estimators will depend on the extent to which the available auxiliary variables are good predictors of the variable being modelled in each case and the extent to which the corresponding assumptions are verified.

It should not be forgotten that, while the most accurate estimates possible are demanded, their error measures (usually the mean square errors) must also be estimated as accurately as possible, or, at the very least, there must be no underestimation of these, so as not to provide an erroneously optimistic picture of the estimates obtained. As mentioned above, when producing estimates at the local level, the communities living in each area often have information (albeit subjective) about the plausible values of the indicators in question, and the estimates provided may contradict such local knowledge. Thus, it is always necessary to remind those who use statistical data that such data has a certain degree of error, and the error measurements accompanying these data should reflect the actual errors made for each area.

Well-extended methods for the estimation of the mean squared errors of the corresponding indirect estimators have also been included in this paper. However, no reference is made in this paper to error measures that might incorporate non-sampling errors, such as coverage errors, non-response errors, errors in the data, substitution of missing data, etc. These issues require further study within small area estimation.

Figure 18

Direct, FH and Census EB estimates (left) of poverty rates, and MSEs of the estimators (right) for the D=25 Montevideo census tracts, for females

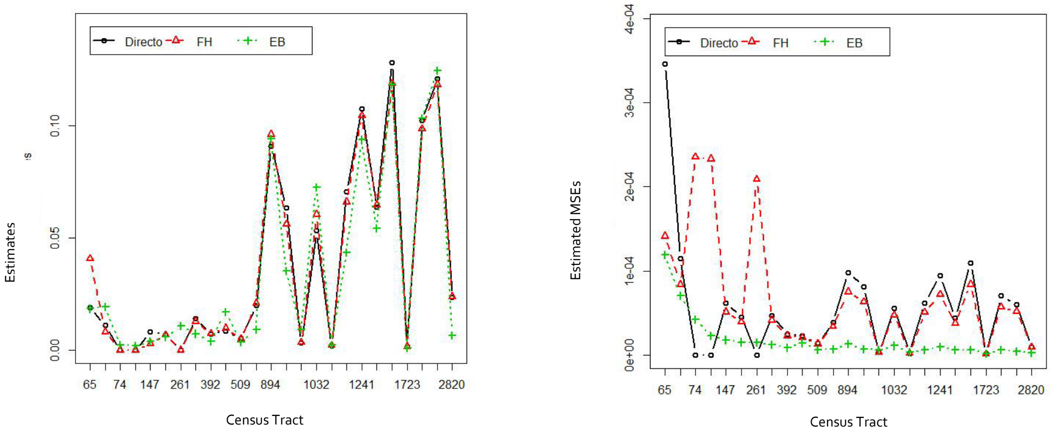


Source: Prepared by the author.
Note: Census tracts (axis x) ordered from smallest to largest sample size, with sample sizes indicated on the axis.

Figure 19

Direct, FH and Census EB estimates (left) of poverty rates, and MSEs of the estimators (right) for the D=25 Montevideo census tracts, for males

(In proportions)



Source: Prepared by the author.
Note: Census tracts (axis x) ordered from smallest to largest sample size, with sample sizes indicated on the axis.

Nor should this paper be considered an exhaustive compendium of methods for disaggregation (or for error estimation), as there are a large number of methods not described due to limited space, (see Rao and Molina (2015) for a more complete description of most previously published methods). This paper has sought to provide an introduction to the subject in question, including the basic methods, in that they form the basis for the study of more advanced methods, including only a few of the more advanced methods that are designed for the estimation of indicators on living conditions.

Bibliography

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67, 1-48.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data, Journal of the American Statistical Association, 83, 28-36.

Bell, W. (1997). Models for county and state poverty estimates. Preprint, Statistical Research Division, U. S. Census Bureau.

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

Casas-Cordero Valencia, C., Encina, J. and Lahiri, P. (2015). Poverty Mapping for the Chilean Comunas, In M. Pratesi (Ed.), Analysis of Poverty Data by Small Area Estimation: Methods for poverty mapping, New York: Wiley.

Correa, L., Molina, I., and Rao, J.N.K., (2012). Comparison of methods for estimation of poverty indicators in small areas. Unpublished report.

Datta, G.S., Fay, R.E. and Ghosh, M. (1991). Hierarchical and Empirical Bayes Multivariate Analysis in Small Area Estimation, in Proceedings of Bureau of the Census 1991 Annual Research Conference, U.S. Bureau of the Census, Washington, DC, 63-79.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimation in Survey Sampling, Journal of the American Statistical Association, 87, 376-382.

Drew, D., Singh, M.P. and Choudhry, G.H. (1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey, Survey Methodology, 8, 17-47.

Elbers, C., Lanjouw, J.O. and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. Econometrica, 71, 355-364.

Estevao, V., Hidiroglou, M. A. and Särndal, C. E. (1995), Methodological Principles for a Generalized Estimation Systems at Statistics Canada, Journal of Official Statistics, 11, 181-204.

Fay, R.E. (1987). Application of Multivariate Regression to Small Domain Estimation, in R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh (Eds.), Small Area Statistics, New York: Wiley, 91-102.

Fay, R.E. and Herriot, R.A. (1979). Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data, Journal of the American Statistical Association, 74, 269-277.

Ferretti, C. and Molina, I. (2012). Fast EB Method for Estimating Complex Poverty Indicators in Large Populations. Journal of the Indian Society of Agricultural Statistics, 66, 105-120.

Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures, Econometrica, 52, 761-766.

Fuller, W.A. (1975), Regression Analysis for Sample Surveys, Sankhyā, Series C, 37, 117-132 (1999). Environmental Surveys Over Time. Journal of Agricultural, Biological and Environmental Statistics, 4, 331-345.

_____ (1999). Environmental Surveys Over Time. Journal of Agricultural, Biological and Environmental Statistics, 4, 331-345.

Graf, M., Marín, J.M. and Molina, I. (2018). A generalized mixed model for skewed distributions applied to small area estimation. Unpublished manuscript.

Ghosh, M. and Steorts, R.C. (2013). Two-stage benchmarking as applied to small area estimation, Test, 22, 670-687.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. and Santamaría, L. (2008). Bootstrap Mean Squared Error of a Small-Area EBLUP, Journal of Statistical Computation and Simulation, 75, 443-462.

_____ (2007), Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model, Computational Statistics and Data Analysis, 51, 2720-2733.

González-Manteiga, W. (2010), Small area estimation under Fay-Herriot Models with nonparametric estimation of heteroscedasticity, Statistical Modelling, 10, 215-239.

Lumley, T. (2017). Survey: analysis of complex survey samples. R package version 3.32.

Prasad, N.G.N. and Rao, J.N.K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators, Journal of the American Statistical Association, 85, 163-171.

Molina, I. and Marhuenda, Y. (2015), sae: An R Package for Small Area Estimation, The R Journal, 7, 81-98.

Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. Computational Statistics and Data Analysis, 58, 308-325.

Marhuenda, Y., Molina, I., Morales, D., and Rao, J.N.K. (2018). Poverty mapping in small areas under a two-fold nested error regression model. Journal of the Royal Statistical Society, Series A, to be published shortly.

Molina, I. and Martín, N. (2018). Empirical best prediction under a nested error model with log transformation, Annals of Statistics, to be published shortly.

Molina, I. and Morales, D. (2009). Small area estimation of poverty indicators. Boletín de Estadística e Investigación Operativa (Bulletin of Statistics and Operations Research), 25, 218-225.

Molina, I., Nandram, B. and Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach. Annals of Applied Statistics, 8, 852-885.

Molina, I., and Rao, J.N.K. (2010). Small Area Estimation of Poverty Indicators. Canadian Journal of Statistics, 38, 369-385.

Molina, I., Salvati, N. and Pratesi, M. (2009). Bootstrap for estimating the MSE of the Spatial EBLUP. Computational Statistics, 24, 441-458.

Neri, L., Ballini, F. and Betti, G. (2005). Poverty and inequality in transition countries. Statistics in Transition, 7, 135-157.

Observatorio Social, Chilean Ministry of Social Development (2017). Metodología de estimación de pobreza a nivel comunal, con datos de Casen 2015. Aplicación de metodologías de estimación directa, de estimación para áreas pequeñas (SAE) e imputación de medias por conglomerados (IMC). (Poverty estimation methodology at the comuna level, with data from Casen 2015. Application of direct estimation methodologies, small area estimation (SAE) and cluster mean imputation (CMI)). Serie Documentos Metodológicos Casen (Casen Methodological Documents Series) 3428.

Pfeffermann, D. and Burk, L. (1990). Robust small area estimation combining time series and cross-sectional data. Survey Methodology, 16, 217-237.

Rao, J.N.K. and Molina (2015). Small area estimation, Second Ed., Hoboken, NJ: Wiley.

Rao, J.N.K. and Yu, M. (1992). Small area estimation by combining time series and cross-sectional data, Proceedings of the Section on Survey Research Method, American Statistical Association, 1-9.

Sen A. (1976), Poverty: An Ordinal Approach to Measurement. Econometrica, 44, 219-231.

Stukel, D. and Rao, J.N.K. (1999). On small-area estimation under two-fold nested error regression models. Journal of Statistical Planning and Inference, 78, 131-147.

Tillé, Y. and Matei, A. (2016). sampling: Survey Sampling. R package version 2.8.

U.S. Bureau of Labor Statistics and U.S. Census Bureau. (2006). Design and Methodology: Current Population Survey, Technical Paper 66. Available at https://www.cen-sus.gov/prod/2006pubs/tp-66.pdf

Torabi, M., and Rao, J.N.K. (2014). On small area estimation under a sub-area level model. Journal of Multivariate Analysis, 127, 36-55.

You, Y., and Rao, J.N.K. (2002a). A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights. Canadian Journal of Statistics, 30, 431-439.

You, Y., and Rao, J.N.K. (2002b). Small area estimation using unmatched sampling and linking models, Canadian Journal of Statistics, 30, 3-15.

Annex

Results of the estimation of average incomes and poverty rates in Montevideo

Table A1

Direct, FH and Census EB estimates of average income, mean squared errors and estimated coefficients of variation of each estimator, for each census tract in Montevideo, for females

(In Uruguayan pesos)

Tract	n_d	Direct			FH			Census EB		
		Est	var	cv	Est	mse	cv	Est	mse	cv
1	93	18 693.71	5 057 851.88	12.03	20 714.21	1 240 278.11	5.38	21 095.71	1 042 681.38	4.84
2	56	13 277.12	4 664 014.87	16.27	12 804.17	2 589 382.73	12.57	14 721.23	1 423 595.65	8.10
3	114	15 950.53	3 705 060.97	12.07	15 709.41	1 449 903.59	7.66	16 405.42	804 002.00	5.47
4	172	21 964.73	3 420 289.14	8.42	22 818.88	1 222 004.22	4.84	23 513.94	685 803.09	3.52
5	277	22 414.35	2 388 487.30	6.90	23 267.98	946 571.10	4.18	22 041.95	423 246.74	2.95
6	179	23 313.57	4 128 976.00	8.72	25 010.83	1 352 338.30	4.65	25 195.60	703 110.23	3.33
7	421	23 755.31	1 924 432.84	5.84	22 592.03	1 163 899.22	4.78	20 071.35	256 888.38	2.53
8	312	16 154.17	1 263 151.79	6.96	17 851.16	736 152.51	4.81	18 028.97	321 051.60	3.14
9	1 113	11 063.69	161 639.01	3.63	11 127.05	151 476.54	3.50	10 598.71	59 822.24	2.31
10	3 482	22 823.33	230 630.51	2.10	22 817.42	209 158.99	2.00	20 764.66	24 042.91	0.75
11	2 081	10 473.08	76 660.70	2.64	10 390.16	74 127.46	2.62	8 758.45	29 006.70	1.94
12	1 216	19 519.30	419 289.74	3.32	19 251.93	347 714.08	3.06	18 123.74	75 945.76	1.52
13	1 844	10 741.54	95 042.46	2.87	10 634.85	91 443.15	2.84	8 652.12	38 841.43	2.28
14	792	21 514.74	790 097.23	4.13	21 340.52	560 681.72	3.51	19 518.73	113 098.81	1.72
15	607	25 157.71	1 369 375.39	4.65	24 472.89	931 112.13	3.94	23 471.75	187 914.34	1.85
16	960	7 748.40	84 359.99	3.75	7 817.03	81 443.96	3.65	8 592.14	68 306.77	3.04
17	2 278	8 167.08	44 950.84	2.60	8 217.67	44 341.90	2.56	8 286.72	28 268.28	2.03
18	2 227	34 942.88	746 573.51	2.47	34 893.09	656 698.10	2.32	33 015.11	64 575.27	0.77
19	504	16 244.32	709 726.52	5.19	15 953.47	526 515.95	4.55	15 340.75	156 341.38	2.58
20	1 402	12 724.70	168 612.47	3.23	12 758.27	158 968.95	3.13	12 436.10	65 960.74	2.07
21	1 667	10 435.48	99 478.80	3.02	10 526.60	95 005.58	2.93	11 354.96	43 098.30	1.83
22	1 073	14 104.97	272 183.12	3.70	14 011.56	248 767.70	3.56	13 370.10	82 132.18	2.14
23	478	24 032.62	2 084 022.21	6.01	24 886.94	920 424.29	3.85	23 547.61	230 604.23	2.04
24	1 521	28 948.32	822 681.41	3.13	28 302.65	588 417.88	2.71	26 395.05	74 187.48	1.03
99	1 364	9 614.82	101 119.90	3.31	9 548.70	96 674.58	3.26	9 081.88	47 799.87	2.41

Source: Prepared by the author.

Table A2
Direct, FH and Census EB estimates of non-extreme poverty (in %), mean squared errors and estimated coefficients of variation of each estimator, for each census tract in Montevideo, for females
(In Uruguayan pesos)

Tract	n_d	Direct		FH		Census EB	
		Est	var	Est	mse	Est	mse
1	93	0.00	0.0000	0.94	1.7750	0.50	0.2707
2	56	0.00	0.0000	6.48	1.8817	2.24	1.0674
3	114	1.68	1.3444	1.74	0.8148	1.51	0.4029
4	172	0.00	0.0000	0.00	1.5588	0.30	0.2303
5	277	0.00	0.0000	0.00	1.5516	0.42	0.1158
6	179	0.00	0.0000	0.00	1.7350	0.20	0.1362
7	421	1.39	0.3732	1.12	0.3327	0.58	0.0810
8	312	1.06	0.3617	1.16	0.3075	0.98	0.1230
9	1 113	4.99	0.4874	5.80	0.3948	6.21	0.0888
10	3 482	2.41	0.0813	2.43	0.0786	0.55	0.0184
11	2 081	11.40	0.6827	10.79	0.5181	10.55	0.0393
12	1 216	0.88	0.0681	0.92	0.0662	0.90	0.0341
13	1 844	12.85	0.9809	12.03	0.6817	10.97	0.0521
14	792	1.80	0.2872	1.82	0.2580	0.70	0.0532
15	607	0.20	0.0406	0.20	0.0403	0.29	0.0480
16	960	9.75	0.9567	10.43	0.6890	11.09	0.1267
17	2 278	11.97	0.5287	11.84	0.4425	12.28	0.0449
18	2 227	0.27	0.0128	0.26	0.0127	0.06	0.0148
19	504	0.69	0.1549	0.83	0.1450	1.72	0.0783
20	1 402	6.64	0.5047	6.27	0.4047	3.83	0.0518
21	1 667	5.61	0.3258	5.89	0.2824	5.08	0.0487
22	1 073	4.30	0.4345	4.11	0.3600	2.89	0.0497
23	478	0.00	0.0000	0.00	1.5526	0.29	0.0510
24	1 521	0.17	0.0136	0.17	0.0135	0.17	0.0233
99	1 364	10.98	0.8465	10.43	0.6014	9.65	0.0898

Source: Prepared by the author.

Table A3
Direct, FH and Census EB estimates of mean income, mean squared errors and estimated coefficients of variation of each estimator, for each census tract in Montevideo, for males
(In Uruguayan pesos)

Tract	n_d	Direct			FH			Census EB		
		Est	var	cv	Est	mse	cv	Est	mse	cv
1	74	24 836.71	11 478 551.50	13.64	23 153.44	2 192 630.31	6.40	25 762.01	1 516 931.82	4.78
2	65	20 460.87	16 472 932.99	19.84	15 443.37	2 435 603.57	10.11	15 765.25	1 269 312.32	7.15
3	72	14 298.89	4 973 475.40	15.60	17 664.10	1 405 480.47	6.71	16 069.72	1 460 721.97	7.52
4	147	26 634.78	6 878 201.92	9.85	25 124.25	1 749 062.81	5.26	23 452.91	883 029.38	4.01
5	218	23 222.64	3 977 409.63	8.59	24 597.89	1 109 370.86	4.28	21 535.17	568 700.05	3.50
6	141	28 783.87	8 278 641.51	10.00	29 394.58	2 072 316.24	4.90	26 974.58	956 465.19	3.63
7	343	25 127.32	2 855 763.61	6.73	25 322.16	1 257 216.13	4.43	20 211.77	353 216.26	2.94
8	261	17 701.92	2 187 547.36	8.36	19 904.65	1 018 640.94	5.07	17 975.03	384 896.60	3.45
9	1 032	11 475.41	190 257.67	3.80	11 517.36	174 746.99	3.63	10 253.44	84 239.24	2.83
10	2 820	24 575.73	396 689.08	2.56	24 658.69	348 365.27	2.39	21 279.59	39 334.03	0.93
11	1 882	11 079.86	99 418.16	2.85	10 975.09	95 447.04	2.81	9 008.44	42 747.93	2.30
12	1 009	21 868.24	638 076.52	3.65	21 342.01	481 650.83	3.25	18 731.09	88 898.30	1.59
13	1 712	10 897.03	106 703.48	3.00	10 766.13	103 066.30	2.98	8 566.47	44 008.29	2.45
14	641	22 605.70	1 090 803.69	4.62	22 636.34	664 159.72	3.60	19 332.12	177 272.29	2.18
15	509	28 797.75	2 175 732.47	5.12	27 945.27	1 339 582.93	4.14	23 540.89	232 540.09	2.05
16	894	8 920.20	130 401.00	4.05	8 893.51	125 439.87	3.98	9 342.98	73 446.28	2.90
17	2 095	8 749.60	58 161.78	2.76	8 830.39	57 265.48	2.71	8 402.07	37 470.56	2.30
18	1 723	38 931.89	1 332 962.25	2.97	38 347.54	1 019 441.37	2.63	33 874.56	93 681.13	0.90
19	417	17 855.77	1 230 524.64	6.21	17 640.61	755 965.09	4.93	15 893.35	237 964.08	3.07
20	1 179	13 531.48	248 104.13	3.68	13 611.71	229 000.88	3.52	12 318.07	72 710.30	2.19
21	1 498	11 147.80	132 574.18	3.27	11 290.17	125 020.79	3.13	11 380.39	65 707.49	2.25
22	929	15 394.11	397 876.11	4.10	15 137.23	349 742.04	3.91	13 156.00	92 731.14	2.31
23	392	27 941.71	3 640 071.50	6.83	28 804.90	1 294 705.53	3.95	23 483.98	321 582.84	2.41
24	1 170	29 940.63	1 097 458.14	3.50	29 943.72	745 397.44	2.88	26 410.63	108 953.03	1.25
99	1 241	10 230.59	130 610.91	3.53	10 227.46	124 563.28	3.45	9 379.01	67 878.14	2.78

Source: Prepared by the author.

Table A4
Direct, FH and Census EB estimates of non-extreme poverty (in %), mean squared errors and estimated coefficients of variation of each estimator, for each census tract in Montevideo, for males

(In Uruguayan pesos)

Tract	n_d	Direct		FH		Census EB	
		Est	var	Est	mse	Est	mse
1	74	0.00	0.0000	0.00	2.3526	0.24	0.4195
2	65	1.89	3.4599	4.06	1.4112	1.94	1.1930
3	72	1.10	1.1480	0.82	0.8380	1.90	0.7071
4	147	0.80	0.6193	0.30	0.5133	0.37	0.1798
5	218	0.68	0.4522	0.66	0.3936	0.57	0.1535
6	141	0.00	0.0000	0.00	2.3300	0.19	0.2299
7	343	1.39	0.4707	1.27	0.4145	0.73	0.1259
8	261	0.00	0.0000	1.89	2.0897	1.15	0.1482
9	1 032	5.32	0.5585	6.04	0.4740	7.40	0.1099
10	2 820	2.34	0.0959	2.38	0.0928	0.61	0.0233
11	1 882	10.23	0.7013	9.86	0.5710	10.38	0.0589
12	1 009	0.31	0.0312	0.34	0.0309	0.93	0.0605
13	1 712	12.81	1.0956	11.88	0.8397	11.89	0.0636
14	641	1.99	0.3835	2.09	0.3381	0.88	0.0724
15	509	0.53	0.1388	0.50	0.1337	0.36	0.0644
16	894	9.07	0.9783	9.60	0.7497	9.33	0.1292
17	2 095	12.09	0.5981	11.83	0.5141	12.47	0.0419
18	1 723	0.17	0.0134	0.16	0.0134	0.08	0.0199
19	417	0.83	0.2247	0.98	0.2099	1.72	0.1449
20	1 179	7.04	0.6212	6.60	0.5067	4.28	0.0596
21	1 498	6.36	0.4374	6.47	0.3789	5.45	0.0606
22	929	6.35	0.8152	5.60	0.6296	3.39	0.0734
23	392	0.71	0.2482	0.73	0.2305	0.39	0.0902
24	1 170	0.21	0.0209	0.22	0.0208	0.24	0.0355
99	1 241	10.76	0.9442	10.44	0.7215	9.38	0.0935

Source: Prepared by the author.

UNITED NATIONS

Series

ECLAC **Statistics**

Issues published

A complete list as well as pdf files are available at
www.eclac.org/publicaciones

97. Disaggregating data in household surveys: using small area estimation methodologies, Isabel Molina (LC/TS.2018/82/Rev.1), 2022.
96. ¿Cuál es el alcance de las transferencias no contributivas en América Latina?: discrepancias entre encuestas y registros, Pablo Villatoro, Simone Cecchini (LC/TS.2018/46), 2018.
95. Avances y desafíos de las cuentas económico-ambientales en América Latina y el Caribe, Franco Carvajal (LC/TS.2017/148), 2018.
94. La situación de las estadísticas, indicadores y cuentas ambientales en América Latina y el Caribe (LC/TS.2017/135), 2017.
93. Indicadores no monetarios de carencias en las encuestas de los países de América Latina: disponibilidad, comparabilidad y pertinencia, Pablo Villatoro (LC/TS.2017/130), 2017.
92. Un índice de pobreza multidimensional para América Latina, María Emma Santos, Pablo Villatoro, Xavier Mancero Pascual Gerstenfeld (LC/L.4129), 2015.
91. Ajuste de los ingresos de las encuestas a las Cuentas Nacionales Una revisión de la literatura, Pablo Villatoro (LC/L.4002), 2015.
90. La evolución del ingreso de los hogares en América Latina durante el período 1990-2008 ¿Ha sido favorable a los pobres?, Fernando Medina y Marco Galván (LC/L.3975), 2015.
89. ¿Qué es el crecimiento propobre?, Fundamentos teóricos y metodologías para su medición, Fernando Medina y Marco Galván (LC/L.3883), 2014.
88. Cuentas satélite y cuentas de salud: un análisis comparativo, Federico Dorin, Salvador Marconi y Rafael Urriola (LC/L.3865), 2014.

STATISTICS

Issues published:

97 Disaggregating data in household surveys
Using small area estimation methodologies
Isabel Molina

96 ¿Cuál es el alcance de las transferencias no contributivas en América Latina?
Discrepancias entre encuestas y registros
Pablo Villatoro y Simone Cecchini

95 Avances y desafíos de las cuentas económico-ambientales en América Latina y el Caribe
Franco Carvajal

94 La situación de las estadísticas, indicadores y cuentas ambientales en América Latina y el Caribe

ECLAC

Economic Commission for Latin America and the Caribbean (ECLAC)
Comisión Económica para América Latina y el Caribe (CEPAL)
www.eclac.org

LC/TS.2018/82/Rev.1