12. Small areas and spatial correlation

PASCAL ARDILLY
INSEE

PAUL BOUCHE
ENSAI - Sciences Po

WENCAN ZHU
ENSAI

Abstract

When we want to circulate the results of a survey on small populations, particularly if we are dealing with small geographical areas, the low sample size matching these populations can lead to estimates that are not accurate enough. The classical sampling theory does not provide a satisfactory solution to this problem and specific estimation techniques must therefore be used, based on using auxiliary information and on models of varying complexity. All these models are a formal link between the variable studied and auxiliary variables. The simplest form is a linear link but there are other non-linear models (Poisson model, logistics model). Most of the models isolate local area-specific effects. Correlations between these effects can be introduced, all the stronger given that the areas are geographically close. This spatial correlation is then likely to improve the quality of localised estimates.

This chapter is devoted to a general introduction to the issue known as the "small-area estimation", with particular attention to considering spatial correlation in the models.

12.1 Setting up the model

12.1.1 Background and objectives

Survey statisticians have a particular interest in estimating unknown θ parameters, defined in a finite and generally large population. Most parameters are totals or immediate derivatives of totals such as means or proportions. More rarely, we find non-linear functions that can still be expressed as total functions (ratios, variances in the population, correlation or regression coefficients). Depending on the survey topic, we may also want to estimate highly non-linear parameters, such as quantiles or inequality indicators, which are not written as total functions.

The parameters are defined using one (or more) variable(s) of interest and are formally stated using expressions generally involving all individuals in population U. We will designate Y as the variable of interest, which will subsequently be considered as unique. The individuals of U are identified by index i, and if parameter θ is a total T then $T = \sum_{i \in U} Y_i$.

When individual value Y_i for each individual i of U is not available, we can estimate T by sampling, *i.e.* from information Y_i obtained from a responding sample, designated s, included in U. The sample is usually drawn from a complex sampling design, for example combining stratification, unequal probability sampling and several sub-samples. Some parameters of interest are not defined for the whole population U but on a sub-population, designated d. Such a sub-population is called a *domain* (or an *area*), and we then have to deal with an area estimate. In this case, the parameter of interest θ may be the total over the area, *i.e.* $T_d = \sum_{i \in d} Y_i$, which must be estimated from collected data. The classical sampling theory assigns each sampled unit i a sampling weight w_i, a positive real coefficient that depends on the sampling method and how non-responses are handled, which "expands" the value of the variable of interest Y_i. To estimate a defined total T over the complete population U, the estimator takes a linear form $\hat{T} = \sum_{i \in s} w_i Y_i$. To estimate a total over an area d, we simply restrict the sum to the elements of d without modifying their weighting, *i.e.* $\hat{T}_d = \sum_{i \in s \cap d} w_i Y_i$. If parameter θ is a mean over d, now designated $\bar{Y}_d$ (including proportions, which are means of Boolean variables), we estimate the size N_d of the area using $\hat{N}_d = \sum_{i \in s \cap d} w_i$ (a size is a total of constant individual values equal to 1) and we form ratio $\hat{\bar{Y}}_d = \hat{T}_d / \hat{N}_d$. But if we know N_d, we can also use the alternative estimate $\hat{\bar{Y}}_d = \hat{T}_d / N_d$.

In all cases, sampling leads to a specific error of estimators $\hat{T}_d$ and $\hat{\bar{Y}}_d$, summarised by means of two indicators respectively called *bias* and *sampling variance*. Consider the case of $\hat{\bar{Y}}_d$. The bias means the difference between the expected value of $\hat{\bar{Y}}_d$, *i.e.* the expected "on average" estimate given the uncertainty that leads to the creation of s, and the $\bar{Y}_d$ parameter, while sampling variance measures the sensitivity of the estimate $\hat{\bar{Y}}_d$ to responding sample s. A precise sampling design results in a low bias and a low variance. The estimators derived from the sampling theory and used by survey statisticians are generally bias-free or have negligible bias. The sampling variance is a decreasing function of n_d, where n_d is the size of the responding sample matching area d, *i.e.* size of $s \cap d$. When n_d is small enough that the quality targets for estimate $\hat{\bar{Y}}_d$ are not reached, there is a *small area* estimation problem.

To address this difficulty, when it is no longer possible to increase the value of size s, designated n, we have to create a new theoretical context to make the final estimate of parameter θ (total T_d or mean $\bar{Y}_d$) less sensitive to responding sample s (or $s \cap d$, which is equivalent). This is done using a *modelling* technique. It means putting oneself within a hypothetical framework that simplifies reality (this is the general definition of a model). The usual approach is to consider that Y_i is

essentially explained by a set of known individual variables X_i for each unit i of the population while involving a few δ *a priori* unknown quantities - parameters of the model. It will be enough to estimate these quantities δ to be able to deduce any unknown value Y_i (corresponding to cases $i \notin s$), and therefore *ultimately* the value of parameter θ.

The use of modelling essentially requires auxiliary information to be available. Of course, we are thinking of variables known at individual level over the entire population U. Let's assume that the auxiliary information about individual i consist in p individual variables, designated $X_{i,1}, X_{i,2}, ..., X_{i,p}$, and start from the principle that there is a "sufficiently reliable" link between such values and the variable of interest Y_i. This link is by construction considered valid when applied to the entire population U, without knowing anything other than $X_{i,1}, X_{i,2}, ..., X_{i,p}$. It must remain valid if we are limited to the responding sample s, which means that the information provided by belonging to the responding sample should not lead the statistician to change the formal expression of this relationship (so-called "uninformative" sampling design). In an ideal world where everything is simple, there would be a certain function f such that for any individual i of U we have $Y_i = f(X_{i,1}, X_{i,2}, ..., X_{i,p}; \delta)$ where δ is a vector parameter unknown at this stage, known as a parameter of the model. In this perfect context, the functional form of the f function is fully known but it is nonetheless configured by δ. If, using information collected during the survey, we manage to estimate the δ parameter satisfactorily, we will be able to predict the values Y_i of all individuals i not sampled (or sampled but not responding) and therefore predict θ.

The traditional framework of sampling statistics is that the sampling theory is not based on any modelling and considers that the variable of interest Y is not random (it is therefore deterministic). It is the sample selection procedure and the non-response mechanism that introduce uncertainty and this uncertainty allows any estimator, such as mean estimator $\bar{Y}_d$, to be considered as a random variable. However, if some mathematical modelling describe Y, since the reality is not that of an ideal and simple world, it would not be reasonable to assume that there is equality between Y_i value and any value of the type $f(X_{i,1}, X_{i,2}, ..., X_{i,p}; \delta)$, because the relationship between Y_i and $X_{i,1}, X_{i,2}, ..., X_{i,p}$ would be too restricted and therefore not credible. Therefore, function f must be considered as including a random component U_i, the first characteristic of which is to be guided by chance. We should now abandon the traditional environment of the sampling theory and consider that the Y *variables are random variables*, such as $Y_i = f(X_{i,1}, X_{i,2}, ..., X_{i,p}, U_i; \delta)$.

12.1.2 Standard individual linear model

The structure of the model therefore uses specific and explicit uncertainties that have no relation to sampling uncertainty. In certain circumstances, it is customary to introduce an individual random variable U_i which is zero on average and thus linked to Y_i, for all i of U (Equation 12.1) :

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p} + U_i \tag{12.1}$$

Auxiliary variables $X_{i,1}, X_{i,2}, ..., X_{i,p}$, which are perfectly deterministic, are called "fixed effects". The uncertainty of the model relating to Y_i values should not be confused with the sampling uncertainty that determines the sample composition s. At this stage, the special features of the context of small area estimation are revealed. As population U is partitioned into D areas, we consider that if i belongs to area d, uncertainty U_i - zero *on average* - is composed of an effect (random) specific to area d, designated τ_d, and an individual (random) residual designated e_i. We therefore have:

$$U_i = \tau_d + e_i. \tag{12.2}$$

In the simplest approach, the two components τ_d and e_i are assumed to be independent, the τ_d are mutually independent, just as the e_i are mutually independent. The expected value and variance associated with the model's uncertainty will be designated ε and v, so the simplest assumptions supporting this model are:

— for expected values, $\varepsilon(\tau_d) = 0$ and $\varepsilon(e_i) = 0$;
— for variances, $v(\tau_d) = \sigma_\tau^2$ and $v(e_i) = \sigma_e^2$.

Furthermore, all possible covariances involving these elementary components are zero. So, overall $\varepsilon(U_i) = 0$ and $v(U_i) = \sigma_\tau^2 + \sigma_e^2$. The format of this model makes it possible to create a correlation between the variables of interest associated with units of the same area since $\forall i \in U$, $\forall j \in U,\ j \neq i$. If $i \in d$ and $j \notin d$ then $cov(Y_i, Y_j) = cov(U_i, U_j) = 0$ and if $i \in d$ and $j \in d$ then $cov(Y_i, Y_j) = cov(U_i, U_j) = \sigma_\tau^2$. Thus, the variances-covariances matrix of the vector of the Y_i, where i covers U, has the form of one diagonal matrix per block, as each block is associated with an area and can be described using a diagonal including everywhere $\sigma_\tau^2 + \sigma_e^2$ while all the other elements of the block take the constant value σ_τ^2.

Due to assumptions covering the moments of random components, such a model can only strictly be applied to quantitative and continuous variables Y - which in particular excludes any qualitative variable of interest (and therefore parameters defined as proportions). Random effect τ_d is a local effect interpreted as being the component of the variable of interest explained by belonging to the area beyond the information contained in individual variables $X_{i,1}, X_{i,2}, ..., X_{i,p}$. The areas are often geographical areas and τ_d intends to reflect the purely explanatory part due to the geographical location of the unit. Assessing the true explanatory part of the location over a given area, and even defining a geographical effect, is a rather philosophical question. Indeed, because it is an easy and practical explanation, one can always consider a significant residual effect that would be due to inadequate consideration of the truly explanatory individual auxiliary variables as a geographical effect. In other words, if there are geographical elements that explain Y, they should ideally be translated in one way or another into fixed effects vector $X_{i,1}, X_{i,2}, ..., X_{i,p}$. Therefore, *a priori*, we have to see local effect τ_d as an "interference" effect and seek to minimise its importance. The smaller parameter σ_τ^2, *i.e.* the weaker τ_d numerical values, the more the explanatory nature will be based on fixed effects $X_{i,1}, X_{i,2}, ..., X_{i,p}$ and therefore the better the model. As the covariance structure is complex, we say that the model belongs to the family of general linear models.

With such a model, the expected value of random variable Y_i is a linear function of the β parameters. The largest explanatory component of Y_i consist in non-random effects $X_{i,j}$ (fixed effects) however the residual component τ_d attributed exclusively to the area is random (random effect). For these reasons, we are talking about *linear mixed model*.

If we use the designations in section 12.1.1, we confirm that $Y_i = f(X_{i,1}, X_{i,2}, ..., X_{i,p}, U_i; \delta)$, where vectorial parameter δ brings together all the unknown quantities appearing in the model, namely $\delta = (\beta_0, \beta_1, \beta_2, ..., \beta_p, \sigma_\tau^2, \sigma_e^2)$. It has a dimension $p+3$, distinguishing $p+1$ actual parameters associated with explanatory fixed effects and two real parameters associated with the variances-covariances structure attached to the model.

12.1.3 Individual linear model with spatial correlation

The standard linear mixed model expresses the assumption of zero correlation between uncertainties U_i associated with individuals belonging to two separate areas. This situation is not necessarily credible, because there is no reason why the limits of geographical zoning making up the areas should be a barrier that suddenly stops any propagation of the measured phenomena. In general, there is a form of natural spatial continuity of the behaviours of localised individuals and two geographically-close individuals on the ground are more likely to display similar Y values

than two distant individuals. From this viewpoint, a relationship between the geographical effects characterising nearby areas seems quite natural.

From a technical viewpoint, we can try to reflect this situation by introducing a correlation that only considers the distance between areas. The analytical form of the correlation is free, provided it decreases when distance increases. In this spirit, we can rely on a model that exactly keeps the expressions 12.1 and 12.2 but that ensures $\forall i \in d$, $\forall j \in d'$, if $i \neq j$:

$$cov(Y_i, Y_j) = cov(U_i, U_j) = cov(\tau_d, \tau_{d'}) = \sigma_\tau^2 exp\left(-\frac{1}{\rho} dist(d, d')\right) \tag{12.3}$$

where $dist(d, d')$ is a defined distance between areas d and d'. For example, we can take the normal Euclidean distance calculated from the coordinates of the centroids of the two areas involved. Coefficient ρ is a scale parameter that allows better adjustment of the model. The more the distance influences covariance, the closer ρ will be to zero. In the particular case where $d = d'$, and when $i \neq j$, then $cov(Y_i, Y_j) = cov(\tau_d, \tau_d) = \sigma_\tau^2 exp(0) = \sigma_\tau^2$. If $i = j$, the variance of the individual effect is added, *i.e.* $cov(Y_i, Y_j) = cov(U_i, U_j) = \sigma_\tau^2 + \sigma_e^2$. This time, the variances-covariances matrix is a full matrix, without zeros. Nonetheless we can consider, as an interesting variant, that the distance becomes infinite when it exceeds a certain threshold. This allows many zeros to be reintroduced into the matrix, so facilitating subsequent digital processing (especially by saving random access memory). Under such circumstances, there are slightly more model parameters since new parameter ρ must be taken into account, so $\delta = (\beta_0, \beta_1, \beta_2, ..., \beta_p, \rho, \sigma_\tau^2, \sigma_e^2)$.

Another approach is to introduce a simple relationship between local effects τ_d of the different areas, ensuring that this relationship is all the stronger as the areas are closer together. Thus, we can consider that the local effect associated with a given area is "almost" a linear combination of local effects of the areas surrounding it, with a linking intensity that diminishes as we move away from the given area. The intensity of the link between the effects τ_d is reflected by two elements, on the one hand a system of coefficients $\alpha_{d,d\prime}$ [1] that govern the relative influence of the different areas distinguished $d\prime$ over a given area d, on the other hand a parameter ρ between -1 and 1 which governs the absolute value of the linking intensity. For all i, we state: $\sum_{d\prime=1, d\prime\neq d}^{D} \alpha_{d,d\prime} = 1$. The proposed relationship between random effects is $\tau_d \approx \rho \sum_{d\prime=1, d\prime\neq d}^{D} \alpha_{d,d\prime} \tau_{d\prime}$. In matrix writing, this becomes:

$$\begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_D \end{pmatrix} = \rho . \begin{pmatrix} 0 & \alpha_{1,2} & \dots & \alpha_{1,D} \\ \alpha_{2,1} & 0 & \dots & \alpha_{2,D} \\ \vdots & \ddots & 0 & \vdots \\ \alpha_{D,1} & \dots & \dots & 0 \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_D \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_D \end{pmatrix} \tag{12.4}$$

by introducing an uncertainty vector u_d that follows a Gaussian distribution with variance $\sigma_u^2 I_D$. This is known as the SAR model (*Simultaneous AutoRegressive model*).

Arbitration between the latter method and the former one is not obvious *a priori*, which is why the only advice to be offered at this stage is to test both methods and then use the available quality assessment tools, particularly those mentioned in part 12.3.

1. Parameters $\alpha_{d,d\prime}$ are the weightings $w_{d,d\prime}$ of weighting matrix W used in the previous chapters. In this chapter, w means sampling weighting.

The introduction of a spatial correlation in the basic linear mixed model does not change any of the restrictive use conditions. Such a model can only be used to estimate θ parameters built from a quantitative and continuous variable of interest. Moreover, it loses much of its benefit if the areas are geographically large because the distance considered is measured between the centroids of the areas.

In practice, to limit the number of non-zero coefficients in the variance-covariance matrix of local effects (and thus speed up the calculations and/or insufficient memory problems), we completely neutralise influence $\alpha_{d,d\prime}$ of $d\prime$ areas located beyond a certain distance of d, or even possibly not in the immediate vicinity of reference area d. Nonetheless, it is difficult to avoid the problems posed by "edge effects" that arise when an area is on the edge of a larger territory, because all its neighbours cannot be taken into account. For example, this is almost systematically true for the border territories of states.

12.1.4 How to deal with qualitative variables using a generalised linear mixed individual model

The logistic model

The parameters for counting any sub-population are based on individual qualitative variables. Let us assume that we want to estimate the total number of individuals θ confirming a given property Γ - such as "being a woman" or "being a farmer under 50". If we define individual variable $Y_i = 1$ when i confirms Γ and otherwise $Y_i = 0$, it is easy to confirm that $\theta = \sum_{i \in U} Y_i$. Random variable Y defined in this way is a dummy variable that quantifies an initially-qualitative individual piece of information. By dividing θ by size U, we get the proportion of individuals in the population who confirm property Γ. Unfortunately, model 12.1 is not at all appropriate for this type of variable. We work around the difficulty by opting for a modelling fully compatible with the dummy variables where Bernoulli distribution will provide the distribution of Y_i. This is a distribution that loads the value 1 with a probability P_i and value 0 with a probability $1 - P_i$. We can therefore consider that for any individual i from the overall population U, variable Y_i is a random variable that obeys Bernoulli distribution $\mathscr{B}(1, P_i)$. The core of the model follows as we will link the P_i parameter to the individual characteristics of i summarised by auxiliary variables $X_{i,1}, X_{i,2}, ..., X_{i,p}$ and we will introduce a local random effect τ_d. The functional form that links P_i to $X_{i,1}, X_{i,2}, ..., X_{i,p}$ and to τ_d must be compatible with the constraint $P_i \in [0,1]$. There are various options, but the most common one is to state, for all i in d:

$$log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p} + \tau_d \tag{12.5}$$

We are talking about a *logistic model*. The expected value of random variable Y_i is P_i, which is obviously not a linear function of the β parameters (unlike in 12.1). For this reason, we say that the model represented by Equation 12.5 is a *generalised linear mixed model*. The class of models taking the form in 12.5 distinguishes between models where local effects τ_d are mutually independent, as in 12.2, and models with spatial correlation, as in 12.3 or 12.4.

The Poisson model

Qualitative information is sometimes aggregated when handling statistical units. If we take the previous example, in the case where the units are households and no longer physical individuals, for each household i we have the total number of individuals Y_i confirming property Γ (the number of women in the household, or the number of farmers under the age of 50 in the household). This variable is no longer a dummy variable but a variable that can take any integer value (in practice this value has always an upper bound). Under these conditions, the Poisson distribution is a fairly

simple natural distribution that can be associated with Y_i. It has a single real parameter λ_i (strictly positive) that will be made to depend on unit i through individual characteristics $X_{i,1}, X_{i,2}, ..., X_{i,p}$ and a local random effect τ_d. The λ_i parameter is often transformed using a simple function before being linked to explanatory factors. In practice, the logarithm function is mainly used, meaning that the complete - generalised linear mixed - model is expressed as follows:

$$\begin{aligned} &Y_i \sim Poisson(\lambda_i) \\ &log(\lambda_i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p} + \tau_d. \end{aligned} \tag{12.6}$$

Once again, local random effects τ_d can be considered as mutually independent, as in 12.2, or spatially correlated, as in 12.3 or 12.4.

12.1.5 Extension to models defined at area level

The Fay and Herriot model

Taking sampling into account, we can produce estimators of any parameter, particularly totals T_d (or means $\bar{Y}_d$) defined at area level d. These estimators are constructed with the individual sampling weights w_i (themselves a function of the sampling method used). They only use information about area d, which is why they are called *direct estimators*. It is possible to construct a model based on these estimators, designated $\hat{T}_d$ for totals and $\hat{\bar{Y}}_d$ for means. The statistical unit modelled is then no longer the individual but the area. The aim is to link the available information $\hat{T}_d$ or $\hat{\bar{Y}}_d$ to a set of explanatory variables, these being adapted to the level handled. Of course, they must characterise the areas and no longer the individuals. Local effects τ_d retain their nature and interpretation, just as in Equation 12.2.

A famous model is the so-called *Fay and Herriot* model, part of the family of linear mixed models. If the explanatory variables selected at domain level are designated $X_{d,1}, X_{d,2}, ..., X_{d,p}$, the most basic version of the model is written:

$$\bar{Y}_d = \beta_0 + \beta_1 X_{d,1} + \beta_2 X_{d,2} + ... + \beta_p X_{d,p} + \tau_d. \tag{12.7}$$

The variable explained here is the true mean in area d. Since this value is unknown, a step must be added to replace it with an estimate. At this stage, estimate $\hat{\bar{Y}}_d$ made from the survey is certainly not good quality since sample $s \cap d$ is small, nonetheless it exists and can be linked to the true value by introducing an error term err_d according to

$$\hat{\bar{Y}}_d = \bar{Y}_d + err_d. \tag{12.8}$$

Variable err_d is the sampling error. This last equation has nothing to do with a model, it is simply the definition of sampling error. Generally, estimator $\hat{\bar{Y}}_d$ is weighted so as to be unbiased or have negligible bias (if there was a calibration, for example, and however if we consider that the non-response was correctly handled) so that the sampling error has expected value zero when taking sampling uncertainty into account, *i.e.* $E(err_d) = 0$. The variance of the error depends on sampling but we know it varies as the inverse of n_d. From now on, we will designate this variance ψ_d. Combining the two previous equations leads to the operational formula:

$$\hat{\bar{Y}}_d = \beta_0 + \beta_1 X_{d,1} + \beta_2 X_{d,2} + ... + \beta_p X_{d,p} + \tau_d + err_d. \tag{12.9}$$

As we saw with the individual models, we can make an assumption of independence between local effects τ_d or instead propose spatial correlation, structured as in equations 12.3 or 12.4.

As assumptions on the expected value and variance of effects τ_d are strictly speaking only compatible with true means $\bar{Y}_d$ that have continuous distributions, it is needless to say that the individual variable of interest Y_i collected for individuals should be quantitative and continuous. Given that, if the variable of interest Y_i is qualitative and if area d has a large enough size N_d, we can consider - sometimes a little boldly! - than true mean $\bar{Y}_d$ may *a priori* take a large enough number of values for this set to be considered as continuous, *i.e.* without "hole". Size N_d is the essential parameter. For example, consider Y_i the indicative variable characterising the "woman" condition. Mean $\bar{Y}_d$ is then the proportion of women in the population of the area. If $N_d = 10$, this mean can take values $k/10$, where k is an integer between 0 and 10, which is a long way from creating a "continuous" situation. If $N_d = 10000$, the mean can take values $k/10\,000$, where k is an integer between 0 and 10 000, which makes the assumption of continuity much more plausible. This is why we can conclude that model 12.9 is acceptable for estimating proportions (qualitative variables of interest) when areas d are not too small.

The Poisson model

Although the Fay and Herriot model adapts well to qualitative variables, *i.e.* parameters that are defined as proportions by area of individuals confirming a property Γ (similar to a sub-population Γ) or as headcount by area of these same individuals, under certain circumstances it may be preferable to have a model more specifically adapted to the counts. Designate $N_{\Gamma,d}$ as the total number of individuals in the area d belonging to sub-population Γ. The sample is used to form the unbiased (or nearly) estimator $\hat{N}_{\Gamma,d} = \sum_{i \in s \cap d \cap \Gamma} w_i$. This estimator only uses area-related information, so it is a direct estimator, and it is poor quality since sample $s \cap d$ is small. Nonetheless, it is a calculable random variable for which the distribution can be modelled using a Poisson distribution. This distribution, which is dependent on a single real parameter λ_d, a function of the area, is particularly suitable for counts. We can show that λ_d is the expected value of $\hat{N}_{\Gamma,d}$ and it should therefore be numerically quite close to this estimate. At this stage, this is a first assumption and not a property that would arise from the sampling theory. Nonetheless, the risk taken remains small because the asymptotic behaviour of direct estimators is close to a Gauss distribution, which the Poisson distribution itself is close to if its parameter is large enough.

The core of the model comes from the following. We generally consider that the logarithm of the λ_d parameter is written in this way

$$log(\lambda_d) = \beta_0 + \beta_1 X_{d,1} + \beta_2 X_{d,2} + ... + \beta_p X_{d,p} + \tau_d \tag{12.10}$$

using the designations in the previous sections. Random variable τ_d keeps the same interpretation. The aim is to distinguish the effect of the location of statistical units beyond what fixed effects $X_{d,1}, X_{d,2}, ..., X_{d,p}$ are capable of causing. The assumptions about correlations between local effects τ_d are identical to those for the models already discussed. More precisely, either we consider that these effects are mutually independent, which is simpler but perhaps sometimes inconsistent with the reality in the field, or we introduce spatial correlations, for example by using expressions 12.3 or 12.4. In both cases, it is a generalised linear mixed model.

12.2 Forming the "small area" estimator

Defining the model to be used is only a first step in the process. At this stage, we still only see the benefit of the model qualitatively, which reduces the scale of the problem by considerably simplifying reality. Indeed, it is much easier to make estimates in an environment where all the relevant information is assumed to be explained by a few well-known variables and by a few parameters rather than working in an undefined system that would consequently depend on an infinite number of uncontrolled components..., as assumed by classical sampling theory!

The next step is to choose the estimation strategy - we should also now talk about prediction since the parameter of interest has become a random variable following modelling.

12.2.1 BLUP estimation strategy: the standard individual model

In this section, we only consider linear models. In this context, several strategies for estimating/predicting the parameter of interest can be used but we now discuss what is probably the most common, the *Best Linear Unbiased Predictor* (BLUP) strategy. We consider the case where the parameter is mean $\bar{Y}_d$. Its predictor is generally a function of the data collected, *i.e.* Y_i where i describes the overall responding sample s. Above all, the statistician seeks a linear predictor of type $\sum_{i\in s} a_i Y_i$ where a_i are real unbiased coefficients, *i.e.* its expected value equals that of $\bar{Y}_d$. Finally, the statistician seeks to minimise the mean square error which is the expected value of the square of the difference between the predictor and the value $\bar{Y}_d$ it has to predict. The solution to this mathematical problem is the BLUP estimator (or predictor), also called the Henderson estimator in the literature. We will designate it $\tilde{\bar{Y}}_d^H$.

In the specific case of the standard individual linear mixed model (see section 12.1.2), when the sampling fraction is negligible, we confirm that the BLUP estimator is written:

$$\tilde{\bar{Y}}_d^H = \gamma_d[\bar{y}_d + (\bar{X}_d - \bar{x}_d)^T \tilde{\beta}] + (1-\gamma_d)\bar{X}_d^T \tilde{\beta} \tag{12.11}$$

All vectors are column vectors, the transposed vector being identified by exponent T. By designating D the total number of areas of interest, by designating $X_i = (X_{i,1}, X_{i,2}, ..., X_{i,p})^T$ the vector of auxiliary variables, $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ the vector of the model parameters associated with these variables, $\bar{x}_d = \frac{1}{n_d}\sum_{i\in s\cap d} X_i$, $\bar{y}_d = \frac{1}{n_d}\sum_{i\in s\cap d} Y_i$ and $\bar{X}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} X_i$, we have:

$$\gamma_d = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \frac{\sigma_e^2}{n_d}} \tag{12.12}$$

$$\tilde{\beta} = \left(\sum_{d=1}^{D}\left(\sum_{i\in s\cap d} X_i X_i^T - \gamma_d n_d \bar{x}_d \bar{x}_d^T\right)\right)^{-1} \cdot \left(\sum_{d=1}^{D}\left(\sum_{i\in s\cap d} X_i Y_i - \gamma_d n_d \bar{x}_d \bar{y}_d\right)\right). \tag{12.13}$$

The coefficients vector $\tilde{\beta}$ does not have a familiar expression here, but we can confirm that it is the classical and well-known estimator called "generalised least squares", frequently encountered in the linear regression model theory. It optimally estimates the vector of unknown parameters β of the model.

It is important to note that we need to know the true means by area $\bar{X}_d$. In practice, this means that individual variables X_i are available in a certain comprehensive file covering the scope of the survey (which does not mean that these individual values are accessible to the statistician responsible for the estimate, who perhaps only has $\bar{X}_d$). However, this file may not be the sampling database and the X_i information used to calculate $\tilde{\beta}$ may come from the survey collection file, in exactly the same way as Y_i. In this case, which is common in practice, it should be ensured that variable X derives from the same concepts in both sources (comprehensive file and collection file). For example, to calculate $\tilde{\beta}$ from a labour force survey where X represents the employment status collected in the survey and to form $\tilde{\bar{Y}}_d^H$ using $\bar{X}_d$ representing the employment status declared in the census would be very hazardous.

Formally, Henderson's estimator consist in two elements that are combined using real coefficient γ_d. The first element - located in the square brackets of Equation 12.11 - is a circumstance estimator that is a little complicated to interpret but that has the same statistical performance as $\bar{y}_d$, the estimator constructed from sub-sample $s \cap d$: it has a sampling variance as a decreasing function of n_d, so *a priori* large. Because this characteristic is associated with the direct estimators, and because at the same time the presence of coefficient $\tilde{\beta}$ - formed from the complete sample - does not allow it to be qualified strictly as a direct estimator, we will talk about a pseudo-direct estimator. The second element is an estimator constructed by multiplying regression coefficient $\tilde{\beta}$ by the true mean of auxiliary variable $\bar{X}_d$, which intuitively should give a value close to the true mean of the variable of interest if the model is appropriate. This estimator $\bar{X}_d^T \tilde{\beta}$ is called a *synthetic estimator*. Its statistical properties are totally dependent on those of $\tilde{\beta}$ since mean $\bar{X}_d$ is not random. We can see for ourselves that $\tilde{\beta}$ is made up of terms involving the entire responding sample s and not just the $s \cap d$ part. By its nature, this makes it very stable, in other words weakly dependent on responding sample s. If we consider only the sampling uncertainty, we can therefore say that the synthetic component offers a low sampling variance. The flipside to this stability is the existence of a sampling bias, which may be numerically strong if the model is inappropriate.

Coefficient γ_d, which is always between 0 and 1, is a remarkable coefficient because it optimally weights (remember that it minimises the MSE) the two separate components, which have completely opposed behaviours in terms of both bias and sampling variance. In this, we say that $\tilde{\bar{Y}}_d^H$ is a *composite estimator* (or *mixed estimator*). The BLUP strategy therefore leads to an expression of γ_d that gives priority to the most efficient of the two components. We'll take the case where σ_τ^2 is small, which corresponds to small local effects τ_d, *i.e.* to an efficient model, since it carries the true explanatory character on controlled auxiliary variables X_i and not on the "catch-all" residual term τ_d. Under such circumstances, we tend to trust the model and build the final estimator based as much as possible on the model, *i.e.* the synthetic estimator. This is actually what happens since γ_d is small. Now let's take the case where the responding sample size n_d is large. Such a context gives confidence in the pseudo-direct estimator, which doesn't (or scarcely) uses the model and therefore by construction that is unlikely to be hindered by the model lacking relevance (the pseudo-direct estimator has weak bias, and in this case low variance since n_d is large). This is what we conclude since γ_d is large, close to 1.

We add that, with this theory, we can easily predict each local effect τ_d. After simple but nonetheless tedious calculations, we get:

$$\tilde{\tau}_d = \gamma_d \left(\bar{y}_d - \bar{x}_d \tilde{\beta} \right) \tag{12.14}$$

which allows the Henderson estimator to be written in a more intuitive form:

$$\tilde{\bar{Y}}_d^H = \bar{X}_d^T \tilde{\beta} + \tilde{\tau}_d. \tag{12.15}$$

There is still one step to be completed to reach the operational stage. Indeed, the BLUP estimator $\tilde{\bar{Y}}_d^H$ has a complex expression that at this stage depends on certain components of the vector of the parameters for model δ introduced at 12.1.2. Indeed, applying the BLUP strategy made it possible to produce estimators $\tilde{\beta}$ of β that have reduced the scale of the problem which means the vector of initial parameters δ is now limited to variance components, *i.e.* two real values σ_τ^2 and σ_e^2. They will be summarised by vector $\Sigma = (\sigma_\tau^2, \sigma_e^2)$. In fact, what we call - commonly but incorrectly - estimator $\tilde{\bar{Y}}_d^H$ is not one since this expression is not calculable and we should therefore strictly designate it $\tilde{\bar{Y}}_d^H(\Sigma)$ and talk about a "pseudo estimator". As components of Σ are unknown, they will have to be estimated using the data collected. Once parameter Σ has been estimated by

$\hat{\Sigma}$, we will substitute $\hat{\Sigma}$ with Σ in $\tilde{\tilde{Y}}_d^H(\Sigma)$ to arrive at a new expression, *i.e.* $\tilde{\tilde{Y}}_d^H(\hat{\Sigma})$, which this time deserves the name of estimator since it can be calculated. We gives the estimator/predictor obtained in this way the name *Empirical Best Linear Unbiased Predictor* (EBLUP).

We frequently estimate Σ by the maximum likelihood method. We also have a variant called restricted maximum likelihood, which can be recommended as it reduces the bias of estimators when sample sizes are modest. Nonetheless this approach imposes an additional assumption on the distribution of random variables τ_d and e_i, which is almost systematically considered as variables following a Gauss distribution. There are no analytical expressions giving $\hat{\sigma}_\tau^2$ and $\hat{\sigma}_e^2$, but numerical algorithms are able to produce estimates matching the theory. Based on these estimates, we get $\hat{\gamma}_d = \dfrac{\hat{\sigma}_\tau^2}{\hat{\sigma}_\tau^2 + \frac{\hat{\sigma}_e^2}{n_d}}$ then:

$$\hat{\beta} = \left(\sum_{d=1}^{D} \left(\sum_{i \in s \cap d} X_i X_i^T - \hat{\gamma}_d n_d \bar{x}_d \bar{x}_d^T \right) \right)^{-1} \left(\sum_{d=1}^{D} \left(\sum_{i \in s \cap d} X_i Y_i - \hat{\gamma}_d n_d \bar{x}_d \bar{y}_d \right) \right) \tag{12.16}$$

and finally the EBLUP estimator:

$$\hat{\tilde{Y}}_d^H = \hat{\gamma}_d [\bar{y}_d + (\bar{X}_d - \bar{x}_d)^T \hat{\beta}] + (1 - \hat{\gamma}_d) \bar{X}_d^T \hat{\beta}. \tag{12.17}$$

Note that we can avoid any assumption relating to the distribution of Y_i by using a "method of moments", but conversely it proves theoretically less effective if the distribution of random variables τ_d and e_i is indeed Gaussian.

12.2.2 Application to the individual linear model with spatial correlation

The BLUP strategy, with its natural extension EBLUP, is applied in exactly the same way when spatial correlations between local effects are introduced. The difference with the standard linear model lies solely in the mathematical expressions of the various estimators, which are obviously much more complicated, but the principles do not change. Detailing the formal expression of the Henderson estimator in the presence of spatial correlations can only reasonably be done by using matrix-based notation, which is burdensome and has no added educational value.

The BLUP (or EBLUP) estimator remains a combination of a direct estimator and a synthetic estimator, with an optimal weighting calculated taking the context into account, depending on the confidence that can be given to the model and the responding sample size n_d. The coefficient σ_τ^2 introduced into Equation 12.3 retains an essential role, but the calculations must now be done also taking the additional coefficient ρ into account, which adjusts the intensity of spatial correlation. The model parameter to be estimated is therefore $\Sigma = (\rho, \sigma_\tau^2, \sigma_e^2)$.

The algorithms for calculating maximum likelihood (restricted, if applicable) adapt to the introduction of an additional parameter, and they produce an estimate of ρ, σ_τ^2 and σ_e^2. The complexity of the variances-covariances structure does not seem to allow methods for estimating Σ other than maximum likelihood or restricted maximum likelihood.

12.2.3 Application to the Fay and Herriot model

The Fay and Herriot model is very important as, in practice, it is widely used. In many cases, it fits well and produces satisfactory estimates, which are preferable to direct estimates. Although it involves a higher degree of aggregation than in the previous models, the BLUP strategy is also implemented within this model. In the expression of Henderson's optimum estimator with the standard model, the σ_e^2 terms have obviously disappeared but on the other hand we find the values

of true sampling variances by area ψ_d. It is important to note that in the standard theory, true sampling variances are assumed to be known. This is obviously not the case in reality, and *in fine* we have to replace theoretical expressions ψ_d by estimators $\hat{\psi}_d$ obtained by applying traditional methods to calculate sampling variance. At this stage, it is recommended to finish by smoothing the $\hat{\psi}_d$ values. This operation protects against including abnormally weak or abnormally strong $\hat{\psi}_d$ estimates, thus avoiding a highly adverse impact on the quality of final estimates by area. We end up at

$$\tilde{\hat{Y}}_d^H = \gamma_d \hat{\hat{Y}}_d + (1-\gamma_d)\bar{X}_d^T \tilde{\beta} \tag{12.18}$$

with

$$\begin{aligned} \gamma_d &= \frac{\sigma_\tau^2}{\sigma_\tau^2 + \hat{\psi}_d} \\ \tilde{\beta} &= \left[\sum_{d=1}^{D} \frac{\bar{X}_d.\bar{X}_d^T}{\sigma_\tau^2 + \hat{\psi}_d}\right]^{-1} \cdot \left[\sum_{d=1}^{D} \frac{\bar{X}_d.\hat{\bar{Y}}_d}{\sigma_\tau^2 + \hat{\psi}_d}\right]. \end{aligned} \tag{12.19}$$

Estimator $\tilde{\hat{Y}}_d^H$ retains a composite form and the BLUP strategy produces the ideal γ_d weighting, shared between direct estimate $\hat{\hat{Y}}_d$, which is independent of the model but unstable, and synthetic estimate $\bar{X}_d^T \tilde{\beta}$ that is totally dependent on the model but, conversely, fairly insensitive to the composition of the responding sample. In rare cases, when estimating a proportion, estimate $\tilde{\hat{Y}}_d^H$ may fall outside the interval $[0,1]$. In this case, the initial model must be adapted.

If spatial correlations are introduced, the above expressions change as a result - by becoming considerably more complicated - but none of the main principles are altered. In all cases, with or without spatial correlations, the software is able to produce estimator σ_τ^2 using maximum likelihood (restricted if necessary), from which we immediately deduce $\hat{\gamma}_d$ and $\hat{\beta}$, then the final EBLUP estimator $\hat{\hat{Y}}_d^H$. Note that in the absence of spatial correlation, there are other methods for estimating parameter σ_τ^2 besides maximum likelihood.

12.2.4 Strategy for non-linear models

The world of non-linear models is technically much more complicated than that of linear models. In particular, the BLUP strategy is not directly suited to this context because there is no satisfactory mathematical solution. Nonetheless, it remains a basic technique and that is why one way to deal with non-linear models, such as the logistic model or the Poisson model, is to replace them with approximate models that have a linear structure. What an approximate model is refers to a complicated but nonetheless operational theory. The BLUP strategy is applied by starting from the approximate linear model.

The initial model may or may not use spatial correlations. The developments presented in the preceding sections are then applied to the approximate linear model.

However, the most compelling approach is to use a strategy better suited to this non-linear context, such as the *Empirical Bayes* strategy, which produces optimal estimates, or the *Hierarchical Bayes* strategy, which corresponds to the classical Bayesian approach.

12.3 The quality of estimators

The model approach will obviously result in making the estimate dependent on the choice of the model and will therefore raise the question of the relevance of the model used. Indeed, simplification has a cost in terms of quality and one may question how correctly this model represents reality.

What are we talking about?

In terms of assessing the quality of small area estimations, it is more necessary than ever to specify the concept of quality. Indeed, the context suffers from a very specific complication due to the coexistence of different kinds of uncertainties, on the one hand, the sampling uncertainty that decides on the composition of the sample, and on the other hand, the uncertainty of the model which handles the variable of interest as a random variable. Quality can be assessed with or without taking the model uncertainty into account.

If the modelling does not include any uncertainty, one is dealing with the survey statistician's classical approach placed in finite population and handling with deterministic individual variables. From this viewpoint, the situation is extremely simple because all the "small area" estimators presented up to now are biased. This is the natural consequence of the failure to take sampling weightings into account (when sampling is not at equal probabilities in all cases), or only partial consideration of these weightings. For example, in the individual standard linear model, the weighting reflecting sampling is always missing. In the Fay and Herriot model, it is certainly found in the direct component $\hat{\bar{Y}}_d$ but not at all in the synthetic part $\bar{X}_d^T \hat{\beta}$. On the other hand, the model provides a decisive advantage in terms of sampling variance because the β parameters that are estimated use the entire responding sample and hence have a weak sampling variance. The estimated local random effect $\hat{\tau}_d$ is unstable but if the model is well suited, it will be numerically small and its variance will therefore have limited influence. The Henderson estimator should ultimately have limited sampling variance and *a priori* less than for the direct estimator if the model has good explanatory power.

When taking model uncertainty into account, if the model is linear by construction, the BLUP estimator is unbiased. Switching to EBLUP only incurs negligible bias. If the model is not linear, the context is much more complicated, but modest bias is expected.

12.3.1 An iterative process

Quality assessment can be designed using a cyclical mechanism (Figure 12.1).

Having, on the one hand, certain selection criteria for explanatory variables, and on the other, having a set of auxiliary variables X potentially explanatory for Y, we adjust a model. At this stage, we have statistical tools to assess the quality of this adjustment. Combined with a prediction strategy, this model produces a theoretical estimator $\tilde{\bar{Y}}_D$. This estimator depends on parameters that contribute to defining the model (at least parameter σ_τ^2, σ_e^2 if applicable and ρ if there is a spatial correlation). These parameters are estimated using an *ad hoc* method. At the end of the cycle, we assess the quality of the final predictor (bias, MSE; see sections 12.3.3 and 12.3.4). If it is not acceptable, a new cycle is initiated by re-examining the relevance of the model, or even the relevance of the prediction strategy or the estimation of the model parameters. Quality assessment also involves checking the relevance of the model's distribution assumptions, if necessary. This is why we will confirm the Gaussian nature of estimated local effects $\hat{\tau}_d$ when a maximum likelihood (whether restricted of not) technique has been used.

12.3.2 The problem of bias

Survey statisticians are sometimes reluctant to use a model-dependent estimator (although it is essential to handle non-response). Their main fear is of substantial bias if we confine ourselves to sampling uncertainty. This risk is inevitable since the model simplifies, and therefore distorts, reality. The important thing is not to escape the bias but to obtain limited bias that is more than offset by the gain in terms of variance. Unless we are working on artificial populations, calculations of bias due to sampling cannot be done, but two simple tools can be used to assess the situation, however without providing evidence.

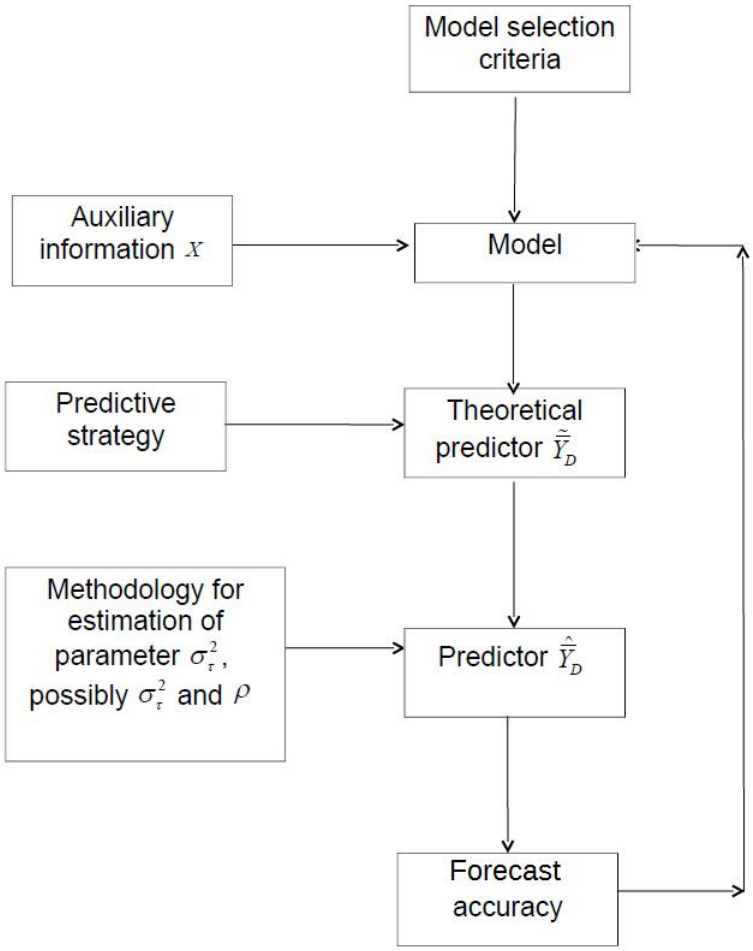

Figure 12.1 – Diagram of the iterative process for assessing the quality of estimators

The first tool is purely graphical and consist in constructing a point cloud where each point represents one of the D areas handled. One of the axes plots the direct estimate (therefore obtained without a model), the other axis plots the "small area" estimate (therefore from a model). If the resulting point cloud is not symmetrical around straight line $y = x$ (first bisector), there is a strong suspicion of bias due to sampling. Nonetheless, there is no inevitability (remember the situation, obviously idealised, of a model reflecting a reality in which all means by area are equal). The converse situation is more convincing in the sense that if the point cloud is symmetrical, there will probably be no significant bias due to sampling. Most often, in practice, we see that scatter points form an angle with the first bisector that, when projected onto the axis representing the "small area" estimate, is smaller than the projection onto the axis representing the direct estimate. This phenomenon is called *shrinkage*, and it is therefore rather an indication of bias due to sampling. It reflects a form of *essentially* excessive concentration of estimates. It arises mechanically from the simplifying model, which has a normalising effect and therefore more or less tends to standardise estimates by area. We stress that this graphical approach offers no evidence but only creates suspicions. In practice, because it cannot accurately reflect reality, any model inevitably creates a theoretical bias due to sampling and the possible symmetry of the point cloud only indicates the probable weakness of this bias.

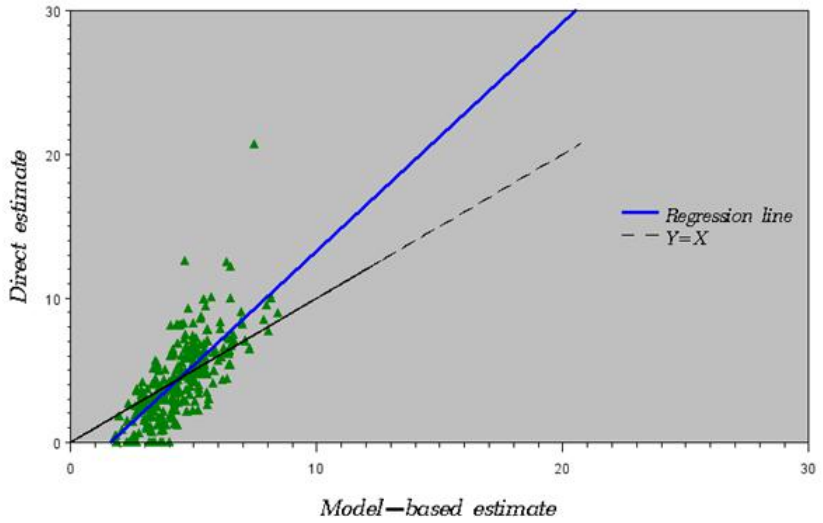

Figure 12.2 – Example of relationship between direct estimates and small area estimates

The second technique is even simpler and more intuitive as it involves summarising estimates of totals $\hat{T}_d^H$ obtained on D small areas and comparing the result with the direct estimate of the total $\hat{T}$ covering the whole population that results from classical sampling theory. In fact, the latter is by construction unbiased (the uncertainty here is exclusively sampling uncertainty). If there is a bias due to the model and this bias is somewhat systematic, a discrepancy will be seen between the two values. On the other hand, a bias without a systematic component cannot be detected since compensation may occur during summation.

It is customary to use the above-mentioned difference in order to increase quality. Indeed, if $\hat{T}_d^H$ is the "small area" estimator of true total T_d in area d, if $\hat{T}$ is the unbiased direct estimator from the overall responding sample s representing the whole population U, we very often adopt the

following final estimate:

$$\hat{\hat{T}}_d^H = \hat{T}_d^H \frac{\hat{T}}{\sum_{d=1}^{D} \hat{T}_d^H} \tag{12.20}$$

which makes it possible to calibrate the estimate of the total in U on $\hat{T}$. This operation is called *benchmarking* and helps to limit the bias of $\hat{\hat{T}}_d^H$ while ensuring consistent distribution.

Furthermore, it is always interesting to map the mean estimates by area $\hat{\hat{Y}}_d^H$, which provides a visual check on the consistency of the estimation system as a whole. Normally, two areas with similar and related characteristics on a map should correspond to two similar estimated means $\hat{\hat{Y}}_d^H$ (in practice, the colours representing their respective values should be in the same range).

12.3.3 Mean square error

In an environment where bias is possible, probable, or even inevitable, the correct error concept is that of mean square error (MSE). This indicator designates the expected value of the square of the difference between the estimator and the parameter. Taking into account both sampling uncertainty and model uncertainty, the theoretical framework offered by the model makes it possible to obtain the expression for MSE and then estimate it without or almost without bias. The expression for MSE and its estimation are very complicated, even with the linear model, and the calculation is therefore entrusted to a software package. Nonetheless, without spatial correlation, we can confirm, if there is a large number of areas D, that the numerically more important term in estimating the MSE of $\hat{\hat{Y}}_d^H$ is $\hat{\gamma}_d \hat{\psi}_d$ for the Fay and Herriot model, and $\hat{\gamma}_d \frac{\hat{\sigma_e}^2}{n_d}$ for the standard individual linear model. Regarding all these error calculations, the results obtained essentially assume that the model is specified in a way that perfectly matches reality (the model can be qualified as "exact"). This is certainly not strictly true! Introducing spatial correlation obviously creates an additional technical difficulty, but general theory makes it possible to succeed, which does not mean that currently available information technology tools are able to use it. Note that under certain particularly favourable circumstances, we have an external source that can provide the true value of the parameter (*e.g.* after a census). This makes it possible to assess the estimation error made directly.

12.4 Implementation with R

■ **Example 12.1 — Dissemination of census on squares.** In 2021, the European Union Statistical Office EUROSTAT wishes to produce statistics (gender, age range, activity, etc.) covering the entire population of each Member State in one-by-one kilometre squares. In addition, in France, INSEE aims to disseminate data from the Population Census using squares whose side could measure some hundred meters. Since 2004, the French census has been carried out by sampling in the municipalities with more than 10,000 inhabitants [2]. Therefore, the targeted area contains too few observations to obtain good direct estimators of the parameters of interest. This is why *small area* estimation is an appropriate statistical technique for using this type of data.

Introducing a spatial correlation in this context makes it possible to reflect the phenomenon of natural continuity of the socio-demographic characteristics of individuals populating geographically contiguous areas. Indeed, moving from any square to neighbouring squares, one cannot reasonably assume there is independence between the behaviours of the statistical units – households or individuals – which formed it. ■

2. The census is complete (exhaustive) in municipalities of fewer than 10,000 inhabitants

The *sae* package in R is used to calculate small area estimates at "area" and "individual" levels, in the case of models respectively not taking and taking into account spatial autocorrelation. This package, which was implemented by Molina and Marhuenda, was described in *The R Journal* (Molina et al. 2015).

The main functions that have been used to handle data from the French census, using a model developed at area level, are from the *sae* package. They are `eblupFH()`, `eblupSFH()`, `mseFH()` and `mseSFH()`.

To produce estimates using a model at individual level, we used functions `eblupBHF()` and `pbmseBHF()` from the *sae* package, as well as function `corrHLfit()` from the *spaMM* package.

Area level modelling: eblupFH() and mseFH() base functions

The first estimates are based on the Fay and Herriot model without spatial autocorrelation. Function `eblupFH()` provides as output:

i) Fay and Herriot estimates for each area;
ii) an estimate of variance σ_τ^2 of the random effect specific to the areas.

In addition, function `mseFH()` produces the calculation of mean square errors associated with each estimate (see section 12.3.3).

The arguments for these functions are the same. The standard syntax is as follows:

```
mod_FH <- eblupFH(formula = Y ~ X1+...+Xp, vardir = varech, method =  ,
    maxiter = m, precision = e, B = 0, data = )
```

First, the `formula` parameter specifies the variable of interest Y as well as the explanatory variables selected $X_1, ..., X_p$. The numerical values of all these variables must be contained in a table that associates one line with each area, specified in the argument `data`. The $\hat{\gamma}_d$ parameters involved in the Henderson estimators are calculated using (estimated) sampling variances per area $\hat{\psi}_d$, which are available in a variable specified in the argument `vardir`.

The estimate of the sole variance parameter of the model is obtained using an *ad hoc* technique. In practice, we use an iterative method that should converge towards value σ_τ^2. The `maxiter` and `precision` parameters are technical parameters (defined either by the user or by default) that govern this iterative process. The algorithm calculates an estimate of σ_τ^2 at each stage of iteration. The role of the `precision` parameter is as follows. As soon as the difference between two consecutive values is less than this (e in our example), the algorithm stops. Otherwise, as long as the maximum number of iterations `maxiter` is not reached, iterations continue. The output indicates whether or not the algorithm converges. The method must also be specified. We can choose from three methods, including the maximum likelihood method and the restricted maximum likelihood method (respectively `method = "ML"` and `method = "REML"`). The third method (`method = "FH"`) is a "method of moments".

The reader's attention is drawn to the need not to add dummy variables identifying the areas to the regressors. Indeed, as the constant is already a part of one of the standard regressors, this practice would lead to create a non-invertible matrix. The following command therefore leads to a failure:

```
mod_FH <- eblupFH(formula = Y ~ X1+...+Xp+as.factor(Carreau) , vardir =
    varech, method =  , maxiter = m, precision = e, B = 0, data = )
```

Spatial correlation at area level: the eblupSFH() and mseSFH() functions

The software estimates a SAR model (see section 12.1.3) of the type

$$\tau = \rho.A.\tau + u \tag{12.21}$$

The parameters of these functions are the same as for the previous functions, except that the proximity matrix **A** (coefficient matrix $\alpha_{i,j}$; see section 12.1.3) must also be specified. The R command is as follows:

```
mod_SFH <- eblupSFH(formula = Y ~ X1+...+Xp, vardir = varech, method =  ,
    maxiter = m, precision = e, proxmat = A, B = 0, data = )
```

The proximity matrix is described by the `proxmat` parameter. It has standardised lines in that the sum of the elements of each line always equals 1.

A similar iterative process to that for use without spatial correlation makes it possible to calculate:

i) Fay and Herriot estimates for each area;
ii) an estimate of the variance of the random effect specific to areas;
iii) an estimate of the spatial autocorrelation parameter ρ.

Modelling at individual level, without spatial correlation: functions eblupBHF() and pbmseBHF()

Using individual level modelling, the `eblupBHF()` function from the *sae* package allows direct estimates and "small area" estimates to be calculated without spatial correlation. The syntax is as follows:

```
mod_BHF <- eblupBHF(formula = Y ~ X1+...+Xp, dom = ,
meanxpop = , popnsize = Popn, data = adr_est)
```

It uses the following configuration — `formula` for the formal expression of the model, `dom` to designate the variable identifying the areas, `popnsize` for the size of the population N_d in each area, `meanxpop` for the means of explanatory variables $\bar{X}_d$ calculated in the whole population of the area. The `data` parameter designates the data table.

The `pbmseBHF()` function estimates the errors (MSE) of "small area" estimators using a *bootstrap* technique. The parameters of this function are the same as for the previous function, to which the number of re-samplings of the *bootstrap* defined by the `B` parameter is added (B=1000 for example).

```
mse_BHF <- pbmseBHF(formula = Y ~ X1+...+Xp, dom = ,
meanxpop = , popnsize = , B = 1000, data = )
```

Taking spatial correlation into account in the individual model

The *spaMM* package can be used to take spatial correlation into account. It can manage several types of models, in particular the Poisson model (see section 12.1.4). Function `corrHLfit()` handles the Poisson model at the individual level with spatial correlation.

```
library(spaMM)
mod_spa <- corrHLfit(formula = Y ~ X1+...+Xp+Matern(1|x+y),
HLmethod = "REML", family = "poisson", ranFix = list(nu=0.5), data = )
```

For configuring this function, `formula` designates the formal expression of the model. The *Matern(1|x+y)* component, which is specific to the function used, takes into account the coordinates x and y of the areas (here, the centres of the squares), which should therefore be contained in the

data table, in order to calculate the distances used in the spatial correlation function. Furthermore, `HLmethod` specifies the method for estimating variance and spatial correlation parameters (here, restricted maximum likelihood), `family` chooses the distribution of the variable of interest (here, a Poisson distribution). The functional form of the spatial correlation can be selected from a configured family of complicated functions called Matérn functions. Parameter `ranFix` specifies the configuration of this family of functions. If we indicate `list(nu=0.5)`, we get the exponential form of Equation 12.3, which is the expression traditionally used - except that the estimated parameter is $\frac{1}{\rho}$ and not directly ρ. The `data` parameter designates the data table.

As output, we obtain, among other things, the estimated coefficients of the model, including the coefficient ρ involved in the exponential spatial correlation (in fact, the reverse if we refer to expression 12.3), optimal predictions $\hat{\tau}_d$ of random local effects, and the estimated variance of random effect $\hat{\sigma}^2_\tau$.

Conclusion

The small area estimate is based on the use of stochastic models. It is the counterpart to a certain shortage of information collected using the sample dividing the area when it is small. In order to limit inaccuracy, unsurprisingly, we have to make assumptions covering the entire population and compensating for the lack of information obtained at local level. Models explicitly involve local geographical effects, the interpretation of which is delicate, in that we can always consider it as a last resort to conceal inadequate consideration of explanatory fixed effects of the phenomenon studied. Basically, the first question is knowing to what extent there is a purely geographical effect. Moreover, these models, whatever they are, always create bias in relation to sampling uncertainty. The main aim is to limit its extent, rather than measure sampling variance, which becomes a secondary aim for the sampling statistician. Of course, we have statistical tools to assess the quality of the adjustment to a model, but this does not guarantee the selected model is suitable for the particular situation of a given area, which may be very specific without the statistician being aware of it. There is no reliable estimate for sampling bias, and we currently have only a few qualitative tools available, which are convincing to varying degrees and which only lead to an assessment of an overall situation. In general, the theory of linear models (*Linear Mixed Models* or LMM) is much simpler than that traditionally used (*Generalised Linear Mixed Models* or GLMM) for non-linear models, which are still really difficult to access. The presence of spatial correlation always complicates the context and then raises the question of the availability of the computer code to make the estimates. The development of R is promising and, in the future, we should move towards extending the range of models accepting spatial correlation.

References - Chapter 12

Battese, George E, Rachel M Harter, and Wayne A Fuller (1988). « An error-components model for prediction of county crop areas using survey and satellite data ». *Journal of the American Statistical Association* 83.401, pp. 28–36.

Chandra, Hukum, Ray Chambers, and Nicola Salvati (2012). « Small area estimation of proportions in business surveys ». *Journal of Statistical Computation and Simulation* 82.6, pp. 783–795.

Coelho, Pedro S and Luıs N Pereira (2011). « A spatial unit level model for small area estimation ». *REVSTAT–Statistical Journal* 9.2, pp. 155–180.

Fay III, Robert E and Roger A Herriot (1979). « Estimates of income for small places: an application of James-Stein procedures to census data ». *Journal of the American Statistical Association* 74.366a, pp. 269–277.

Molina, Isabel and Yolanda Marhuenda (2015). « sae: An R package for small area estimation ». *R Journal, in print*.

Pratesi, Monica and Nicola Salvati (2008). « Small area estimation: the EBLUP estimator based on spatially correlated random area effects ». *Statistical methods and applications* 17.1, pp. 113–141.

Rao, John NK (2015). *Small-Area Estimation*. Wiley Online Library.