Taking advantage of sampling designs in spatial small-area survey studies

This document corresponds to the accepted manuscript of the paper, with the same title, published in Statistical Modelling (https://doi.org/10.1177/1471082X231226287)

Vergara-Hernández, C; Mari-Dell'Olmo, M; Oliveras, L; Martinez-Beneito, MA

Abstract

Spatial small area estimation models have become very popular in some contexts, such as disease mapping. Data in disease mapping studies are exhaustive, that is, the available data are supposed to be a complete register of all the observable events. In contrast, some other small area studies do not use exhaustive data, such as survey based studies, where a particular sampling design is typically followed and inferences are later extrapolated to the entire population. In this paper we propose a spatial model for small area survey studies, taking advantage of spatial dependence between units, which is the key assumption used for yielding reliable estimates in exhaustive data based studies. In addition, and in contrast to most survey-based spatial studies, we also take into account information on the sampling design and additional supplementary variables to obtain estimates in small areas. This makes it possible to merge spatial and sampling models into a common proposal.

1 Introduction

Spatial statistics has historically devoted considerable effort to deriving reasonable statistical estimates which take into account the geographic nature of the units of study. When sparse information is available, or simply when spatial dependence is strong, spatial models usually yield substantially improved estimates, which overcome many of the small-area problems that non-spatial estimates entail. Hereafter, we will refer to small areas, our target of interest, as those commonly used in disease mapping models (Martinez-Beneito and Botella Rocamora 2019), with very few observations (or even no observations) per area. In this context, the use of spatial methods is really necessary since, with such scarce information, spatial dependence is an invaluable second source of information that could change statistical quantities for some areas from not estimable to estimable, as a consequence of the spatial dependence between areas. In this sense, we will estimate those quantities for which little or no information is available, as geostatistical models usually do, according mainly to spatial dependence and the information in their neighbors.

Frequently, survey studies follow particular sampling designs in order to achieve a desired precision for their estimates over the entire region of study or for some geographic/demographic groups. Nevertheless, if small area inference is pursued, these groups rarely coincide with those small areas since this would typically require huge sample sizes in order to achieve the desired precision for each area. This produces several problems when these surveys are used for inference at a small area level (smaller than that used for the sampling design), such as areas without any sampled unit or areas with unacceptably high variability on (direct) estimates of interest. In this context the use of hierarchical models, and spatial models in particular, is crucial, although taking the sampling design into account in those models is not a trivial task.

In addition, when a sampling design has been followed, ignoring that design for inference seems a clear error since the available sample (and estimates) may thus no longer be representative of the overall population, over- or under-representing some population groups and thereby producing potentially biased or misleading results. As a solution, sampling theory proposes alternative estimates and methods that take the followed sampling scheme into account. Therefore, taking the sampling design into account seems, in principle, as important as considering spatial dependence in small-area studies; thus, model proposals which consider both features are very welcome. Unfortunatelly, however, sampling designs in small-area studies are typically planned for larger spatial units than those considered, so most of the

sampling theory available in the literature does not apply, at least directly, with such small areas (smaller than those used during the design of the sample).

The spatial literature has already dealt with sampling-based estimates for small areas. Nevertheless, as mentioned by Lawson (2018) "this problem has been studied by many authors, albeit without much reference to the spatial nature of the geographical problem", despite of the evident benefits that spatial dependence brings in the small-area context. Spatial proposals to survey-based small-area estimation have usually followed model-based approaches in which data are aggregated by areas. Within this approach, (non-spatial) design-based estimates are derived at a first stage, which take the sampling scheme followed into account and, at a second stage, spatial dependence is later induced on these estimates by means of a (spatial) smoothing model (Marhuenda, Molina, and Morales 2013; Porter et al. 2014; Chen, Wakefield, and Lumely 2014; Mercer et al. 2015; Mercer, Lu, and Proctor 2019; Li et al. 2019; Paige et al. 2020). Although the uncertainty in the first stage of these models is usually considered for the second stage, the different processes modelled in both phases could interact; however, that interaction can hardly be corrected in these two-phase separate proposals. Moreover, for very small areas without sampled units, the first stage of such a procedure may not even be feasible so, in some contexts, this approach may not be an option. In addition, many of these works usually consider Normal likelihood models, as proposed by Fay and Herriot (1979) which, for areas such a small as ours, that choice can only be justified if the original variable also follow that distribution, at the individual level, since the Central Limit Theorem does not apply at all to such small areas.

On the other hand, individual-level models have also been proposed that estimate the effect of any variable potentially related to the outcome of interest. In a second stage these estimates are poststratified to reproduce these effects on the domains of interest according to their population compositions. These proposals have been frequently posed as Bayesian hierarchical models, where the two processes mentioned above can be merged into a single model. This procedure is known in the literature as *Multilevel Regression and Poststratificaction (MRP)* (Little 1993; Gelman and Little 1997; Park, Gelman, and Bafumi 2004; Gelman 2007). Although the flexibility of this approach makes it a very natural framework for inducing spatial dependence, and related survey design issues, spatial dependence has rarely been considered within this approach. In this paper we will follow the MRP approach, with the particular objective of including both sampling design information and spatial dependence together in our estimates.

Our general goal with this paper is to propose a sensible inference procedure for small-area survey studies. As specific goals, we also intend to take spatial dependence and the sampling design jointly into account in order to take advantage of these important features for our estimates. Furthermore, as a second specific goal, we also intend that this procedure will allow us to make inferences at a very small area level, smaller than that originally planned for the sampling design. Specifically, we will try to take advantage of the information used for the sampling plan and, if possible, all the related additional information that allowed us to yield enhanced estimators, including as much information as possible. In this paper we will focus on the specific, but frequently observed, case of estimating proportions of a binary variable in a collection of geographic units. Additionally, we will also assume that we will have some categorical auxiliary variables that will be helpful, or that simply should be considered by sampling design in order to improve those proportion estimates. The case of numerical (non-binary) output variables or covariates is left for future work, although the studied setting covers a high proportion of real settings in practice.

The paper is organized as follows. Section 2 begins by introducing our approach in the trivial case of simple random sampling, where no particular sampling design is used to draw the survey sample. We then generalize the mentioned simple framework to more complex sampling designs. In particular, we show how to proceed in the very common setting of having a stratified random sample or if we would like to use additional auxiliary variables for our inference. In section 3, we show an illustrative case study with real data, comparing several estimation proposals. Finally, Section 4 presents some final conclusions.

2 Spatial small-area inference with different sampling designs

In this section we present our proposal for spatial modeling of small areas. First, we introduce the use of spatial models when simple random sampling has been used. Subsequently, we consider the inclusion of auxiliary variable(s) information in survey-based inference, regardless of whether this information is used for designing the sample of the study (stratified sampling) or not, respectively. Finally, we discuss how to combine these tools in those sampling designs that allow it.

2.1 Small-area inference under simple random sampling

Let us assume a study region composed of I (small) spatial units, neighborhoods of a city from now on, in accordance with the following case study. Let $Y \in \{0, 1\}$ be a random variable, whose mean (proportion of 1s in the corresponding population) we would be interested to know for each of the mentioned spatial units. Let us also assume that Y has been observed for just a sample of individuals for each spatial unit, sampled under simple random sampling (SRS) for all of them. We will denote the corresponding sample size for each neighborhood as n_i $(i = 1, \ldots, I)$ of a total of N_i $(i = 1, \ldots, I)$ individuals living in each of them. In that case, we would like to estimate from our sample the proportions $\pi_i = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}$, where Y_{ij} stands for the observed value of Y for the j-th individual of the i-th neighborhood.

Under SRS, those proportions may be estimated as simply the sample mean of Y at each neighborhood $\hat{\pi}_i^{SRS} = \bar{Y}_{i\cdot} = (n_i)^{-1} \sum_{j=1}^{n_i} Y_{ij}$. This elemental result can be found, for example, on page 35 of Lohr (2010), among many other sampling theory books. Nevertheless, we will reference sampling theory results according to Lohr's book (adding the corresponding book page) to be more precise. The sample mean in that scenario can be shown to be an unbiased estimator of each π_i (see page 52 of Lohr (2010)), with variance (page 36 of Lohr (2010)):

$$\widehat{Var}(\hat{\pi}_i^{SRS}) = \frac{\hat{\pi}_i^{SRS}(1 - \hat{\pi}_i^{SRS})}{n_i}(1 - \frac{n_i}{N_i}). \tag{1}$$

Additionally, due to the hypothetical Normality of sample means for large sample sizes, a $100(1 - \alpha)\%$ confidence interval for π_i is frequently derived as simply $\hat{\pi}_i^{SRS} \pm t_{1-\alpha/2} \sqrt{\widehat{Var}(\hat{\pi}_i^{SRS})}$, where $t_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a t distribution with $n_i - 1$ degrees of freedom (page 43 of Lohr (2010)).

The $\hat{\boldsymbol{\pi}}^{SRS} = (\hat{\pi}_1^{SRS}, \ldots, \hat{\pi}_I^{SRS})$ estimator may also be justified from other points of view beyond sampling theory. For example, it can be easily shown that $\hat{\pi}_i^{SRS}$ is also the maximum likelihood estimator (MLE) of π_i for the frequentist model:

$$Y_{ij} \sim Bernoulli(\pi_i),\ i = 1, \ldots, I;\ j \in \partial_i,$$

and is also the MLE for the abridged version of this model:

$$O_i \sim Binomial(\pi_i, n_i),\ i = 1, \ldots, I,$$

where $O_i = \sum_{j=1}^{n_i} Y_{ij}$. In addition, under a Bayesian perspective, $\hat{\boldsymbol{\pi}}^{SRS}$ may be also viewed as the posterior mode of the models with either of the two previous data likelihoods and with π_i, $i = 1, \ldots, I$, independently distributed with $Uniform(0, 1)$ prior distributions. This is a consequence of the equivalence of the frequentist MLE and the Bayesian posterior mode when independent uniform prior distributions are used for the components of $\boldsymbol{\pi}$. These results further support the convenience and use of $\hat{\boldsymbol{\pi}}^{SRS}$ as an (unbiased) estimator of $\boldsymbol{\pi}$ under SRS.

According to all this, $\hat{\boldsymbol{\pi}}^{SRS}$ may be considered an unquestionable estimator of $\boldsymbol{\pi}$ under SRS. Nevertheless, this estimate may sometimes be inappropriate. Specifically, when the spatial units of study are small, $\hat{\boldsymbol{\pi}}^{SRS}$ may become unreliable (excesive variance) and yield inappropriate estimates.

Numerous papers have been published in the disease mapping area that try to fix these small-area estimation problems. The main proposals to overcome the problems inherent in that approach come from Bayesian hierarchical models, where the independent uniform prior distributions for the components of $\boldsymbol{\pi}$ are substituted by a spatial prior distribution. That distribution shares information between neighboring units and therefore uses neighbors' information for each area as a secondary information source, therefore yielding more reliable risk estimates. Different conditional auto-regressive (CAR) distributions have been proposed for this purpose, such as the intrinsic CAR (Kunsch 1987) or lCAR distributions (Leroux, Lei, and Breslow 1999), or even the combination of spatial and non-spatial processes (Besag, York, and Mollié 1991).

Something similar could be done to estimate $\boldsymbol{\pi}$ in survey sampling contexts, as introduced above, to improve the $\hat{\boldsymbol{\pi}}^{SRS}$ traditional estimates. Being more precise, we could estimate the proportions $\boldsymbol{\pi}$ as the estimates arising from the following Bayesian hierarchical model:

$$\begin{aligned} Y_{ij} &\sim Bernoulli(\pi_i^*),\ i = 1, \ldots, I;\ j = 1, \ldots, n_i \\ logit(\pi_i^*) &= \mu + \theta_i \\ &\ldots \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I)$ is a vector of spatially structured random effects following, for example, one of the CAR-based spatial processes mentioned (Vandendijck et al. 2016; Watjou, Faes, and Vandendijck 2020). For this spatial model the population proportions π_i could be estimated as:

$$\hat{\pi}_i^{Spatial} = \frac{n_i}{N_i}(\bar{Y}_{sampled})_i + \frac{N_i - n_i}{N_i}(\bar{Y}_{unsampled})_i, \tag{2}$$

where $(\bar{Y}_{sampled})_i$ and $(\bar{Y}_{unsampled})_i$ correspond, respectively, to the mean for the sampled and unsampled populations for the i-th spatial unit (page 205 of Gelman et al. (2014)). The first term on the right hand side of this expression is known for each neighborhood, it can be directly calculated from the available sample, and the second term can be sampled from the corresponding predictive distribution of the model for each step of the MCMC process. This estimator takes into account the finite sample nature of each proportion since when the sample size grows, the first component of the previous expression, which has no uncertainty, increases its weight, reducing therefore the variability of $\hat{\pi}_i^{Spatial}$. In any case, if the sampled population fraction (n_i/N_i) is small, $\hat{\pi}_i^{Spatial}$ will be very similar to the model parameters π_i^*, so these values could alternatively be used as π_i estimates for simplicity. In that case, the finite sample correction (2) will not be needed in practice.

In principle, this model could improve the small-area estimation problems that $\hat{\boldsymbol{\pi}}^{SRS}$ could show. Unfortunately, in contrast to $\hat{\boldsymbol{\pi}}^{SRS}$, the proportion estimates arising from this model will no longer be unbiased, since these proportion estimates will be shrunk towards those of their neighbors. Nevertheless, the more efficient use of information in spatial models, which is shared between nearby units, should yield more accurate proportion estimates than raw sample means, which will compensate for the (supposedly small) bias introduced by these estimators.

2.2 Small-area inference under stratified sampling designs

Despite the simplicity of SRS, which is the simplest scheme followed for collecting survey samples, this is not such a common sampling scheme in practice. Instead, more complex designs are usually followed for these studies. Stratified sampling, for example, is a much more frequent choice, where samples are divided into several population groups and their estimates are then merged in order to derive compound indicators with a given precision. This process could be done for either the whole population or particular geographical domains. Unfortunately, when a sample is planned, it is rarely planned for a small-area division of the region of study but rather for a coarser spatial division, or even the whole region, since otherwise the required sample size will typically be unattainable. Our goal will now be to use those stratified samples to make inferences at a higher spatial disaggregation level (small areas) than that originally planned, taking into account the specific stratified sampling scheme followed.

From this point on, we will consider that a stratified sampling design has been used to draw a sample of surveys and a SRS scheme has been used in order to collect the corresponding sample within each stratum. More specifically, and without loss of generality, we will assume that the strata of the sample correspond to age groups. This framework generalizes the previous SRS scheme to a much more common and general setting in practice. Additionally, from now on we assume that the stratified sample was originally collected for the whole region of study in order to estimate one proportion for that whole region with a given precision, but we want to now use that sample to estimate that same proportion, as accurately as possible, for the units of a small-area division of the region of study. If an intermediate spatial division, coarser than the small areas of interest, was considered for the sampling design, all the following would remain valid; nevertheless, for clarity we will focus on the simplest case of a single stratified sample for the whole population. That is, the following is valid regardless of the spatial division considered, if any, to design the sample of the study.

Sampling theory states (see page 73 of Lohr (2010)) that under stratified sampling, with SRS within each age group $k(= 1, \ldots, K)$, an unbiased estimate of $\pi = N^{-1}\sum_{j=1}^{N} Y_j$, the overall proportion of $Y = 1$ over the whole population, could be derived as:

$$\hat{\pi}^{Str} = N^{-1}\left(\sum_{k=1}^{K} \bar{Y}^{Str=k} N^{Str=k}\right), \tag{3}$$

where $\bar{Y}^{Str=k}$is the sample mean of Y on age group k and $N^{Str=k}$ the population size of that same age group, with $N = N^{Str=1} + \ldots + N^{Str=K}$. Note that each of the terms in the sum on the right-hand-side of this expression ($N^{Str=k}\bar{Y}^{Str=k}$) can be seen as a ratio estimator (See Chapter 4 of Lohr (2010)) for the total of Y for each stratum, where that ratio ($\bar{Y}^{Str=k}$) compares the total people with $Y = 1$ against

the total sampled people in that stratum. Afterwards, this ratio is combined with the auxiliary design information (population size $N^{Str=k}$) in order to derive a ratio estimate (page 119 of Lohr (2010)) of the total of people with $Y = 1$ in the corresponding age group. Therefore, the stratified estimator $\hat{\pi}^{Str}$ may be considered as a particular case of ratio estimator, which will be of importance in the following.

Unfortunately, stratified sampling on small areas poses additional estimation problems. On the one hand, stratified sampling collects several samples (strata) for each unit of analysis, instead of a single sample, so this increases the sample size requirements to derive estimates of this kind. Furthermore, having a sampling design at a coarser spatial level than the small areas of interest, could easily lead to some of the age groups not being sampled in some of those small areas, which would make the direct estimate of the parameter of interest $\hat{\pi}^{Str}$ unfeasible for those areas. Additionally, when small-area inference is undertaken in that setting, the original stratification weights, which were initially planned for larger units, may no longer be valid for deriving estimates for small areas. Our goal now will be to solve these problems when small-area inference is pursued on a stratified sample designed for a coarser geographical division.

Let us denote $N_i^{Str=k}$ and $n_i^{Str=k}$ the total and sampled populations, respectively, for the i-th neighborhood and k-th age group, $i = 1, \ldots, I;\ k = 1, \ldots, K$. Note that $N_i^{Str=k}$ will generally be known prior to sample selection, since the sampling design is organized according to these values, but not $n_i^{Str=k}$, which will depend on the sample drawn. Let us also denote $Y_{ij}^{Str=k};\ i = 1, \ldots, I;\ j = 1, \ldots, n_i^{Str=k};\ k = 1, \ldots, K$, the observed value for the binary variable of interest Y, for the j-th individual of the k-th age group on the i-th neighborhood. Let us also assume that our goal will now be to estimate $\pi_i = N_i^{-1} \sum_{k=1}^{K} (\sum_{j=1}^{N_i^{Str=k}} Y_{ij}^{Str=k})$, which is equivalent to $N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}$ if the stratified design/notation is ignored.

Expression (3) shows how to use the information in the stratified sample to compound estimates for the whole population. Unfortunately, that expression would in principle be useless for deriving areal estimates for a small-area division of the region of study which does not coincide with that used for the sampling design. Nevertheless, we mentioned that Expression (3) could be viewed simply as a combination of ratio estimators, beyond its stratified nature, and that combination could now be used as a post-stratification procedure to derive stratified small-area estimators. Thus, let $\bar{Y}_i^{Str=k} = (n_i^{Str=k})^{-1} \sum_{j=1}^{n_i^{Str=k}} Y_{ij}^{Str=k}$ be the sample mean for stratum k and the (small) spatial unit i. That estimate can be also expressed as: $(\sum_{j=1}^{N_i^{Str=k}} X_{ij}^{Str=k} Y_{ij}^{Str=k}) / (\sum_{j=1}^{N_i^{Str=k}} X_{ij}^{Str=k})$, where $X_{ij}^{Str=k} = 1$ if individual j of stratum k and spatial unit i is included in the sample, and 0 otherwise. For this last expression the denominator is also stochastic, since SRS is used over an upper geographical level and we do not know in advance how many of the elements of the k-th age group in the sample will correspond to spatial unit i. Therefore, $\bar{Y}_i^{Str=k} N_i^{Str=k}$ would simply be a ratio estimator, supported by sampling theory, of the Y total for age group k. Therefore, the sum of these ratio estimators over all the age groups $\sum_{k=1}^{K} \bar{Y}_i^{Str=k} N_i^{Str=k}$ could be used to estimate the total of Y over the entire spatial unit i, and

$$\hat{\pi}_i^{Str} = N^{-1} \sum_{k=1}^{K} \bar{Y}_i^{Str=k} N_i^{Str=k}, \tag{4}$$

could be used as a combined ratio estimator, or stratified estimator, of π_i suitable for a different (small area) spatial division than that considered for the stratified sampling design. This makes the available sample, originally planned for a coarser spatial division, also useful for making inference in a small-area division of the region of study.

Although Expression (4) allows π_i to be estimated for any division of region of the study, taking the stratified nature of the sample into account does not solve the small-area estimation problems that the size of the units of study could pose. These problems should be worse than those illustrated in Section 2. If independence was assumed at the inference stage for each age group and spatial unit, then $\bar{Y}_i^{Str=k}$ in Expression (4) may sometimes not be computed if no sampled units were available for that particular age group and spatial unit combination. Additionally, if the number of sampled units for that combination was simply small, this would make $\bar{Y}_i^{Str=k}$ inaccurate, and that variability would also be transferred to $\hat{\pi}_i^{str}$. Thus, small-area modeling becomes even more necessary in stratified sampling than for SRS designs. In SRS, small-area problems were alleviated by the modeling of the vector $\boldsymbol{\theta}$, considering their components as spatially dependent variables. Now we have a matrix $\mathbf{E}(\mathbf{Y}) = (\pi_i^{Str=k})_{i,k=1}^{I,K}$ to estimate, instead of a single vector, which we could model in several different manners, possibly considering spatial dependence among its cells.

Being more precise, let us denote the outcome variable for the j-th sampled individual of spatial unit i and age group k as $Y_{ij}^{Str=k}$. We could then model $Y_{ij}^{Str=k}$ as a Bernoulli variable of probability $(\pi^*)_i^{Str=k}$,

where that probability could be modeled, for example, as:

$$logit((\pi^*)_i^{Str=k}) = \mu + \psi_i + \phi_k.$$

In this proposal, age ($\boldsymbol{\phi}$) and spatial effects ($\boldsymbol{\psi}$) could be modeled as usual, thus ϕ_1 could be fixed to 0 while an improper uniform prior could be considered for $\phi_2, \ldots, \phi_K$ and a spatial CAR pior, such as the Leroux et al.'s lCAR, could be used to induce spatial dependence on $\boldsymbol{\psi}$. Once this model is fitted, we could estimate the probabilities $\hat{\pi}_i^{Str=k}$ for each neighborhood and stratum, either taking a finite sample correction into account, or not, like for SRS. Once estimated $\hat{\pi}_i^{Str=k}$, the overall probability for each neighborhood could be defined as:

$$\hat{\pi}_i^{StrSmall} = N^{-1} \sum_{k=1}^{K} (\pi^*)_i^{Str=k} N_i^{Str=k}. \tag{5}$$

We could use $\hat{\pi}_i^{StrSmall}$ as an alternative estimate to $\hat{\pi}_i^{Str}$, which, in principle, could solve the small-area estimation problems that this second estimator could exhibit. Note that the above expression poststratifies the model estimates $(\pi^*)_i^{Str=k}$, in this sense incorporating the stratified nature of the sample in the estimates leads to an MRP model (Gelman and Little 1997), with spatial dependence in our particular case.

Alternatively, an abridged version of the model above could be proposed, as already mentioned for the SRS case, with important computational benefits if computation was an issue due to the dimensions of the data set of study. Thus, the Bernoulli likelihood of the model above could be substituted by a Binomial distribution on the number of observed people $O_i^{Str=k}$ with $Y_{ij}^{Str=k} = 1$ (i.e. $O_i^{Str=k} = \sum_{j=1}^{n_i^{Str=k}} Y_{ij}^{Str=k}$) for the corresponding age group and neighborhood as follows:

$$O_i^{Str=k} \sim Binomial((\pi^*)_i^{Str=k}, n_i^{Str=k}),\ i = 1, \ldots, I,\ k = 1, \ldots, K$$

and once again:

$$logit((\pi^*)_i^{Str=k}) = \mu + \psi_i + \phi_k.$$

In this case, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ could be modeled in the same manner as for the previous model. Both models can be shown to have the same data likelihood and thus their equivalence. Although this abridged model could be more convenient from a computational point of view, the consideration of additional auxiliary variables, as introduced below, would mean a major change in its formulation since $O_i^{Str=k}$ and $n_i^{Str=k}$ would need to be recomputed for all the groups of the auxiliary variables. In contrast, considering an additional auxiliary variable for the original Bernoulli implementation would pose no particular problem in terms of the model implementation so, in that case, the original model could be more convenient in practice. In any case, both formulations should be borne in mind and used depending on their appropiateness for each particular problem.

2.3 Small-area inference with auxiliary variables

In addition to stratified sampling, a similar proposal to that above could be used to take into account the information of additional auxiliary variables, even though these were not originally used to design the sample. In fact, as mentioned, stratification is just a particular case of the use of auxiliary variables, specifically in the sampling design phase of the survey. We have seen that poststratification has been the key to deriving stratified estimators for units of study that were not originally considered in the survey design. This same idea can also be used to include auxiliary variables in the inference, in which case it is also possible to consider spatial dependence.

Let us assume now that our survey collected, in addition to our variable of interest Y, a second categorical variable $Z \in \{1, \ldots, K\}$ for each individual in the sample. Let us also assume that we knew the population $N_i^{Z=k}$ belonging to each of the K categories of Z for each small area i that makes up the region of study, i.e. $N_i^{Z=k} = \sum_{j=1}^{N_i} 1_{Z_{ij}=k};\ i = 1, \ldots, I;\ k = 1, \ldots, K$, where $1_{Z_{ij}=k}$ is equal to 1 if $Z = k$ for individual j of spatial unit i, and 0 otherwise. In that case, similarly to the stratified setting, we could use a ratio estimator to estimate the population with $Y = 1$ for each spatial unit and group of the auxiliary variables as follows:

$$(\hat{\tau}_Y)_i^{Z=k} = \left(\frac{\sum_{j=1}^{n_i} (Y_{ij}^k \cdot 1_{Z_{ij}=k})}{\sum_{j=1}^{n_i} 1_{Z_{ij}=k}} \right) N_i^{Z=k} = (\bar{Y}|Z_{ij} = k) N_i^{Z=k},$$

where $(\bar{Y}|Z_{ij}=k)$ stands for the sample mean of Y for those individuals in the sample with $Z_{ij}=k$. In addition, we could also estimate $\pi_i = N_i^{-1}(\tau_Y)_i$, the percentage of people with $Y=1$ in spatial unit i as:

$$\hat{\pi}_i^{Ratio} = N_i^{-1}(\hat{\tau}_Y)_i = N_i^{-1}\left(\sum_{k=1}^{K}(\hat{\tau}_Y)_i^{Z=k}\right) = N_i^{-1}\sum_{k=1}^{K}(\bar{Y}|Z_{ij}=k)N_i^{Z=k}. \tag{6}$$

This estimator would take advantage of the information on the total population for each value of Z, $N_i^{Z=k}$, that we know for each spatial unit. Once again, this estimator allows us to include information from auxiliary variables in our estimates in a similar manner to how Expression (4) did it with the stratification variable. Nevertheless, this new estimate would pose the same small-area estimation problems as those posed by the estimator in Expression (4). Luckily, they could be solved in exactly the same way as done there, by means of small-area hierarchical modeling that considers spatial dependence. Thus, the hierarchical model proposed to take advantage of stratified sampling would once again be valid also for the poststratification of auxiliary variables by simply changing the $N_i^{Str=k}$ used in Expression (4) for the totals $N_i^{Z=k}$ of each group for the auxiliary variable.

2.4 Small-area inference with stratification and auxiliary variables

We have just seen that accounting for stratified designs in small-area divisions unplanned during the sampling design and using auxiliary variables, also ignored in the sampling design, does not pose any difference in terms of their statistical analysis. We are now going to consider how to combine these two tools in order to obtain enhanced estimators of use in many real settings. We will consider the case of having two auxiliary variables, whether or not they are used to design the sample (stratification variables), since the case of considering more than two variables would be closely similar.

Let us consider that we have two auxiliary variables, Z_1 with values in $\{1,\ldots,K_1\}$ and Z_2 with values in $\{1,\ldots,K_2\}$. In that case, our quantities of interest will be:

$$\begin{aligned}\pi_i = N_i^{-1}\sum_{j=1}^{N_i}Y_{ij} = N_i^{-1}\sum_{j=1}^{N_i}\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}Y_{ij}1_{(Z_1)_{ij}=k_1}1_{(Z_2)_{ij}=k_2} = \\ = N_i^{-1}\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\pi_i^{Z_1=k_1,Z_2=k_2}N_i^{Z_1=k_1,Z_2=k_2},\end{aligned} \tag{7}$$

where $\pi_i^{Z_1=k_1,Z_2=k_2} = (N_i^{Z_1=k_1,Z_2=k_2})^{-1}\sum_{j=1}^{N_i}Y_{ij}1_{(Z_1)_{ij}=k_1}1_{(Z_2)_{ij}=k_2}$ stands for the proportion of times that $Y=1$ in neighborhood i when $Z_1=k_1$ and $Z_2=k_2$ and $N_i^{Z_1=k_1,Z_2=k_2}$ is the population corresponding to those values of the auxiliary variables. As introduced in the case of a single stratification or auxiliary variable, $\pi_i^{Z_1=k_1,Z_2=k_2}$ for $k_1=1,\ldots,K_1$ and $k_2=1,\ldots,K_2$ could be alternatively estimated by the corresponding means $\bar{Y}_i^{Z_1=k_1,Z_2=k_2}$ in the available sample. In that case, Expression (7) would produce an estimate $\hat{\pi}_i$ that would take into account both auxiliary variables. Nevertheless, that choice would show clear small-area problems, similar to the same problems mentioned above. As a consequence, $\pi_i^{Z_1=k_1,Z_2=k_2}$ could be estimated instead by means of a spatial hierarchical model that would solve the small-area problems that these indicators could exhibit. Therefore, once again, the case of 2 (or more) auxiliary variables would reduce to an MRP approach in which (spatial) model-based estimates are derived and then poststratified to integrate the effect of both variables.

In practice, one problem with the estimator in Expression (7) is that $N_i^{Z_1=k_1,Z_2=k_2}$ will not be so frequently available in population databases when two or more auxiliary variables are combined. In that case, if the distribution of Z_1 and Z_2 in the population are reasonably independent, that is:

$$N_i^{Z_1=k_1,Z_2=k_2} \approx \frac{N_i^{Z_1=k_1}\cdot N_i^{Z_2=k_2}}{N_i},$$

then that estimator could be restated as:

$$N_i^{-1}\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\pi_i^{Z_1=k_1,Z_2=k_2}\frac{N_i^{Z_1=k_1}\cdot N_i^{Z_2=k_2}}{N_i} = N_i^{-1}\sum_{k_1=1}^{K_1}\left(N_i^{-1}\sum_{k_2=1}^{K_2}\pi_i^{Z_1=k_1,Z_2=k_2}N_i^{Z_2=k_2}\right)\cdot N_i^{Z_1=k_1}. \tag{8}$$

The term within the parenthesis in this expression reproduces, for each value of Z_1, a ratio estimator for Z_2 conditioned to that particular value of Z_1. Similarly, the rest of that expression provides a second

ratio estimator, in this case for Z_1, which combines, according to this variable, the conditional ratio estimators previously calculated for Z_2. Therefore, this estimator may be viewed as a sequential ratio estimator of each of the auxiliary variables considered, where the information on each of the auxiliary variables is progressively included into the full estimator. Clearly, the order of the sums in Expression (8) could be reversed, thus the order used to include the auxiliary variables into the full estimator does not have any impact on it.

Both estimators in Expressions (7) and (8) are two separate alternatives, although the suitability of the second depends on the adequacy of the independence assumption for Z_1 and Z_2. Using one or the other in practice will depend on the availability of population information for the combination of both auxiliary variables, $N_i^{Z_1=k_1,Z_2=k_2};\ i=1,\ldots,I;\ k_1=1,\ldots,K_1;\ k_2=1,\ldots,K_2$ and/or the validity of the independence hypothesis between Z_1 and Z_2, if this may be assessed, otherwise it will just have to be simply assumed if population information on their combination is not available.

3 Estimating the proportion of low- and middle-income countries population for Barcelona neighborhoods: A real case study

In this section we are going to illustrate and compare the models presented above. First, we start be comparing estimators that ignore the sampling design of the data, which simply assume that the sample has been obtained by simple random sampling. Second, we take into account the complex sampling design of the data to obtain and compare more complex and realistic estimators.

3.1 Simple random sampling

In this illustrating case study we are going to compare different estimates assuming SRS, or simply ignoring the sampling design, for the estimation of the proportion of people from low- and middle-income countries (LMIC for short from now on) for the neighborhoods of Barcelona, Spain. The results of this study will serve as a reference to assess the advantages of the models introduced in the following subsection, which take into account the stratified nature of the sample. The outcome variable Y is equal to 1 if the surveyed person was born in a LMIC, and 0 otherwise. Barcelona is divided into ($I=$)73 neighborhoods with populations varying from 631 to 58,054 inhabitants, with a median of 20,533 people. The survey data used for this case study corresponds to the 2016 Barcelona Health Survey, with a sample size of around 4000 people. This survey was designed to achieve a given precision level for each of the 10 districts of the city, so our units of study are smaller than those corresponding to the originally planned design. The sample size for the different neighborhoods of the city vary between 1 and 226, with a median value of 46, yielding a surveyed population of 2.4 people for each 1000 inhabitants. Interestingly, also for 2016, the Barcelona register of inhabitants, which collects population information for the entire city (not just a sample) contains information on the country of birth of each city inhabitant. Therefore, we could alternatively obtain from this source the proportion of population from LMICs for each neighborhood throughout this exhaustive sample, that could be used as a gold standard (the *gold standard* $\boldsymbol{\pi}$ from now on) for assessing the quality of the different survey estimates. All the details (Rmarkdown document with the `R` code run) of the following and subsequent analyses in this paper can be found in the supplementary material of the paper.

By using the survey data, we have derived four different estimates of the proportion of LMIC population at the neighborhood level $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_I)$. First, we have derived *frequentist SRS* estimates of those proportions, $\hat{\pi}^{SRS}$ as defined above, which are simply the sample means of $\{Y_{ij}:\ j=1,\ldots,n_i\}$. Variance estimates ($\widehat{Var}(\hat{\boldsymbol{\pi}}^{SRS})$) and 95% confidence intervals (assuming Normality) were also calculated for this approach, as described above. Second, we have derived *Bayes SRS* estimates of those proportions, assuming a binomial Bayesian model above with independent $Uniform(0,1)$ prior distributions for the π_is. For this model, these prior distributions (which are equivalent to $Beta(1,1)$ prior distributions) are conjugate, so the posterior distribution of those proportions are known ($Beta(O_i+1, n_i-O_i+1)$, where O_i is the number of people from LMICs in the sample for the i-th areal unit). Therefore, Markov chain Monte Carlo is not required for inference on this model and analytic expressions of the estimates may be derived instead. Thus, the corresponding posterior modes for the proportion of people from LMICs for this model are:

$$\frac{(O_i+1)-1}{(O_i+1)+(n_i-O_i+1)-2}=\frac{O_i}{n_i},$$

which could be used once again as proportion estimates for this model. This Bayesian point estimate coincides with $\hat{\pi}_i^{SRS}$. Nevertheless, for the Bayesian case, credible intervals for π_i may be derived from those posterior distributions, which would avoid the Normality assumption of the frequentist estimators, possibly not so suited for small areas. These credible intervals would take into account the Binomial nature of the outcome variable, which should yield some advantages. Finally, we have derived two spatial proportion estimates corresponding to the Bayesian spatial model introduced previously, which considers those proportions as dependent quantities in contrast to the other two alternatives. Specifically, we consider the spatial estimates accounting for the finite population nature of the sample ($\hat{\boldsymbol{\pi}}^{Spatial}$ in Expression (2)) and, as an alternative, the proportion estimates which ignore that nature (simply $\boldsymbol{\pi}^*$). As point estimates we use the posterior mean of those variables. An lCAR (Leroux, Lei, and Breslow 1999) is used as prior distribution for $\boldsymbol{\theta}$ to induce spatial dependence in the proportion estimates.

Table 1 shows several statistics for these proportion estimates. To assess their accuracy we have calculated (the square root of) the mean squared error of each proportion estimate with respect to the gold standard, drawn from the Barcelona register of inhabitants. Furthermore, we have assessed the precision of those estimates by means of the length of their 95% confidence/credible intervals (CI). Additionally, we have also assessed the empirical coverage rate of those CIs as the proportion of those 73 intervals that contained the corresponding gold standard proportion π_i, which we would expect to be around 0.95. We have also measured the spatial dependence of the estimates by means of the Moran's I statistic (Moran 1948), which should be compared with the gold standard's ($\boldsymbol{\pi}$) Moran's I, equal to 0.202. Finally, we have assessed the divergence of the Moran's I statistics for these estimates with that of the gold standard $\boldsymbol{\pi}$. We have done this by evaluating the posterior probability of the Moran's I of these statistics is less than that of $\boldsymbol{\pi}$. For the Bayesian cases, this has been done by drawing samples of the posterior distribution of the proportions and then calculating the Moran's I posterior distribution from these. For the frequentist case, we have repeatedly sampled independent values of a t distribution with $n_i - 1$ degrees of freedom of mean $\hat{\pi}_i^{SRS}$ and variance $\widehat{Var}(\hat{\pi}_i^{SRS})$, the distribution implicitly assumed for deriving CIs. For these samples we have calculated the corresponding Moran's I, which have been used to estimate the probability of that statistic being less than that of $\boldsymbol{\pi}$.

Table 1: Assessment criteria for each proportion estimate assuming simple random sampling

Statistic	$\hat{\boldsymbol{\pi}}^{SRS}$	*Bayes SRS*	$\hat{\boldsymbol{\pi}}^{Spatial}$	$\hat{\boldsymbol{\pi}}^{*}$
$\sqrt{MSE}$	0.083	0.083	0.049	0.049
95% CI length	0.271	0.252	0.172	0.172
95% CI coverage rate	0.808	0.959	0.945	0.945
Moran's I	0.129	0.129	0.233	0.235
P(estimated I<real I)	0.986	0.969	0.671	0.663

The first line of Table 1 shows how the point estimates (posterior means) of the spatial model ($\hat{\boldsymbol{\pi}}^{Spatial}$ and $\hat{\boldsymbol{\pi}}^{*}$) outperform those of the independent SRS estimates in terms of MSE with respect to $\boldsymbol{\pi}$. This illustrates how sharing information between spatial units improves point estimates when dealing with small areas and how that increase in the information used outperforms the potential bias of those estimates. The benefits of spatial estimates are also evident in the CI lengths, as the CIs for this model are substantially narrower. Despite the enhanced precision of the spatial estimates, their 95% CI coverage rate is substantially better than that of the frequentist SRS estimate. Interestingly, the *Bayes SRS* estimate already shows an evident improvement in the CI coverage rate. We could conclude therefore that the coverage rate problems of the frequentist SRS estimates come from the Normality assumption, which, for low/moderate sample sizes does not hold in practice. Finally, the last two lines of Table 1 summarize the spatial dependence of the proportion estimates. The spatial model shows a clear increase in the spatial dependence of the estimates, with its Moran's I much closer to that of the gold standard (0.202). Moreover, for the spatial estimates the gold standard Moran's I falls around the 67% percentile of the posterior distribution, which highlights the compatibility of the spatial dependence of those estimates with that of $\boldsymbol{\pi}$. In contrast, the gold standard Moran's I falls around the 97% percentile of the estimates for the other two non-spatial alternatives, suggesting that these estimates display weaker spatial dependence than they ought to. Finally, Table 1 points out the equivalence in practice of the finite population ($\hat{\boldsymbol{\pi}}^{Spatial}$) and general ($\hat{\boldsymbol{\pi}}^{*}$) spatial estimates, which suggests that for the sampling fraction in this study (2.4 sampled individuals for each 1,000 inhabitants) no finite population correction is required in practice.

3.2 Taking into account the sampling design

The sampling used in the Barcelona health survey originally followed a stratified design, with strata corresponding to five age groups ($\{0-19, 20-39, 40-59, 60-79, 80-\ldots\}$) and sex, which was ignored in the previous case study. Our goal is now to take that sampling design into account for inference. In addition, although it was ignored for the sampling design, we have an additional auxiliary variable that could be taken into account for inference. This binary variable collects whether or not each individual owns their place of residence. The Barcelona health survey specifically asks this question in each survey and the city census contains the total of this same variable for each neighborhood of the city ($\{N_i^{home=k},\ i = 1,\ldots,73,\ k = 1,2\}$). Thus, this variable could be used to derive a ratio estimator that includes information on the ownership of the homes and its relationship with the birth country of Barcelona inhabitants, which could be possibly closely related. Additionally, a combined estimator could be also derived that takes into account the age stratification design of the sample, the home ownership information available, spatial dependence and integrates it all into a single model. This is our main goal now. For simplicity, and due to the mild association found between sex and the LMIC origin in Barcelona residents, we will ignore the sex stratification of the sampling design throughout the following analysis.

We have derived and compared three estimators in this case study. First, we have developed stratified estimates of the proportion of LMIC population for each neighborhood $\hat{\boldsymbol{\pi}}^{\boldsymbol{Str}}$ (see Expression (3)), taking the age stratification design into account and ignoring both spatial dependence and home ownership information. Second, we have developed a ratio estimator $\hat{\boldsymbol{\pi}}^{\boldsymbol{Ratio}}$ (see Expression (6)), ignoring both spatial dependence and age stratification, but considering the home ownership information available. Finally, we have developed a third estimator $\hat{\boldsymbol{\pi}}^{\boldsymbol{Full}}$ considering age strata, home ownership information and spatial dependence. This estimator has been derived by fitting a Bayesian hierarchical model which assumes:

$$Y_{ij}^{age=k_1\ home=k_2} \sim Bernoulli((\pi^*)_i^{age=k_1\ home=k_2})$$
$$logit((\pi^*)_i^{age=k_1\ home=k_2}) = \mu + \psi_i + \phi_{k_1} + \varphi_{k_2},$$

for $i = 1,\ldots,73$, $j = 1,\ldots,n_i$, $k_1 = 1,\ldots,5$ and $k_2 = 1,2$. In this case, $\boldsymbol{\psi}$ models the spatial variability between neighborhoods, $\boldsymbol{\phi}$ the effect of the age group and $\boldsymbol{\varphi}$ the effect of home ownership. Once this model has been fitted we poststratify all these probabilities for the different levels of the auxiliary variables into a single estimate per neighborhood ($\hat{\boldsymbol{\pi}}^{\boldsymbol{Full}}$), as described above (Expression (7)). Given the limited impact of the finite sample correction (Expression (2)) in the previous case study, we have not considered that correction when deriving $\hat{\boldsymbol{\pi}}^{Full}$ in this new study. Additionally, we have used Expression (8), which assumed independence between both auxiliary variables for deriving $\hat{\boldsymbol{\pi}}^{Full}$ since the population register of Barcelona does not publish population data for the combination of these variables at the neighborhood level, just the separate univariate frequencies (by neighborhood) for each of them. The comparison of the standardized, ratio and full estimates, among them and against the estimates derived in the previous section, will allow to assess their particular advantages. The `R` code used for this analysis is also supplied as annex supplementary material.

As a consequence of using auxiliary variables for inference, we find that some of their combinations with the neighborhood are not present in the sample. Therefore, the proportion of LMIC population may not be estimated in those neighborhoods for the $\hat{\boldsymbol{\pi}}^{Str}$ and $\hat{\boldsymbol{\pi}}^{Ratio}$ estimators. Specifically, for the age-stratified estimator $\hat{\boldsymbol{\pi}}^{Str}$ we find that 6.6% of the age group-neighborhood combinations are not sampled, which yields 17.8% of the neighborhoods with at least one of those groups unsampled; in other words, 17.8% (13) of the components of $\hat{\boldsymbol{\pi}}^{Str}$ may not be estimated with our sample. This makes it unfeasible to use two-phase estimators that include the sampling design in the first phase of the analysis and spatial dependence in the second. Similarly, for the ratio estimator, 2.7% of the home ownership-neighborhood combinations are not sampled, which means that $\hat{\boldsymbol{\pi}}^{Ratio}$ may not be estimated for 5.5% (4) of the neighborhoods. For the full approach, $\pi_i^{age=k_1\ home=k_2}$ estimates are derived according to a model which transfers information between age, home ownership groups and nearby neighborhoods. Therefore, $\pi_i^{age=k_1\ home=k_2}$ may be estimated for all values of i, k_1 and k_2 under this approach, which allows to estimate $\hat{\boldsymbol{\pi}}^{\boldsymbol{Full}}$ for all the neighborhoods. This is an initial evident advantage of this approach.

Table 2 contains the same statistics as those shown in Table 1 but for the new estimators developed in this reanalysis. In addition, this table also contains two columns corresponding to the $\hat{\boldsymbol{\pi}}^{\boldsymbol{SRS}}$ estimator, but considering as missing values those positions where $\hat{\boldsymbol{\pi}}^{\boldsymbol{Str}}$ (13 missing values) and $\hat{\boldsymbol{\pi}}^{\boldsymbol{Ratio}}$ (4 missing values) are also missing, respectively. This allows the stratified and ratio estimators to be compared with those that considered SRS over the same set of spatial units.

The comparison of the second and third columns of Table 2 shows that there is hardly any difference between the stratified and SRS estimators in terms of MSEs. Thus, although SRS estimates were in principle biased due to ignoring the stratified sample design followed, the effect of ignoring that design

Table 2: Assessment criteria for the stratified, ratio and full proportion estimates

Statistic	$\hat{\boldsymbol{\pi}}^{\boldsymbol{SRS}}$ (13 missing)	$\hat{\boldsymbol{\pi}}^{\boldsymbol{Str}}$	$\hat{\boldsymbol{\pi}}^{\boldsymbol{SRS}}$ (4 missing)	$\hat{\boldsymbol{\pi}}^{\boldsymbol{Ratio}}$	$\hat{\boldsymbol{\pi}}^{\boldsymbol{Full}}$
$\sqrt{MSE}$	0.061	0.062	0.069	0.062	0.043
95% CI length	0.206	0.101	0.232	0.115	0.141
95% CI coverage rate	0.867	0.600	0.826	0.667	0.959
Moran's I	0.048	0.044	0.158	0.150	0.305
P(estimated I<real I)	1	1	0.838	0.806	0.643

in $\hat{\boldsymbol{\pi}}^{\boldsymbol{SRS}}$ does not seem to be high. Differences are more evident in terms of confidence intervals since the lengths of the stratified estimates are substantially shorter than those assuming SRS. Although this seems like good news, this increased precision also reduces the coverage rate of the ICs to 0.600, far lower than expected. This seems a consequence of the small size of the units of study, which, under stratified sampling, means that many combinations of age groups and neighborhoods have really low sample sizes. Specifically, 28.5% of those combinations have a number of sampled units between 1 and 5, and for them it would be quite common that all the observed individuals have either a foreign or a national origin. In those cases, the variance estimates of the corresponding age-group and neighborhood combination (Expression (1)) are equal to 0, which could explain the lower coverage rate of these confidence intervals. Finally, regarding the spatial dependence of the stratified estimator, we do not find any big difference with respect to the unsatisfactory performance of the SRS estimator.

Columns 4 and 5 of Table 2 show similar results to the previous comparison. Interestingly, in our particular case, the ratio estimator seems to provide some advantages over the stratified estimator. This could simply be a consequence of the closer relationship between the LMIC origin and the ownership status of the inhabitants of Barcelona (LMIC people hardly ever own their homes (4.0% in the health survey sample) in contrast to people born in high income countries (48.9%)) than between the LMIC origin and the age group.

Finally, the 6th column of Table 2 illustrates $\hat{\boldsymbol{\pi}}^{\boldsymbol{Full}}$ results. This estimator is clearly better than $\hat{\boldsymbol{\pi}}^{\boldsymbol{Str}}$ and $\hat{\boldsymbol{\pi}}^{\boldsymbol{Ratio}}$ for all the criteria considered, except perhaps the CI lengths (which are wider), which, as shown, make them to be superior in terms of their coverage rate. According to the Moran's I statistic, $\hat{\boldsymbol{\pi}}^{\boldsymbol{Full}}$ also appears to have a spatial dependence comparable to that of $\boldsymbol{\pi}$. Additionally, $\hat{\boldsymbol{\pi}}^{\boldsymbol{Full}}$ also seems to display significant advantages in comparison to the SRS spatial estimator (Table 1) for all the criteria considered. Therefore, merging the advantages of stratified, ratio and small-area modeling into a single proposal seems to yield a clear reward for small-area modeling.

Figure 1 shows several choropleth maps with some of the estimated proportions (posterior means) in this work, as well as with the census gold standard. This figure illustrates the evident noise of the SRS estimates when no small-area method is used. This noise is filtered out for the spatial and full estimates where the proportion estimates show, in addition, a similar spatial dependence to that displayed by the gold standard. Differences between the spatial and full estimators are visually more subtle, indicating that the greatest benefit of the full estimator may come from the modeling of the spatial dependence. In any case, the main advantage of taking auxiliary variables into account could be the increase in the precision of the proportion estimates, which seems confirmed by the CI lengths of the spatial and full models in Tables 1 and 2. Maps with the length of the CIs for the Barcelona neighbourhoods for the three estimators in Figure 1 can also be found as supplementary material to the paper.

4 Conclusions

This paper proposes merging both survey sampling and spatial small-area approaches in order to derive sensible estimates from a survey sample over a collection of small areas. Our approach describes how to include information from auxiliary variables for making inference on the outcome of interest, whether these were used for the sampling design (stratification variables) or not (ratio estimators), merging this process with spatial methods in order to yield improved estimates which enjoy the advantages of both approaches. The use of spatial methods to avoid independence assumptions is the key to deriving sensible estimates in survey-based small-area studies since those assumptions are clearly harmful in the small-area context. As shown, the potential bias introduced by spatial dependence has many more advantages than drawbacks when dealing with small areas.

As illustrated in this paper, our proposal to incorporate the sampling design into the estimates leads

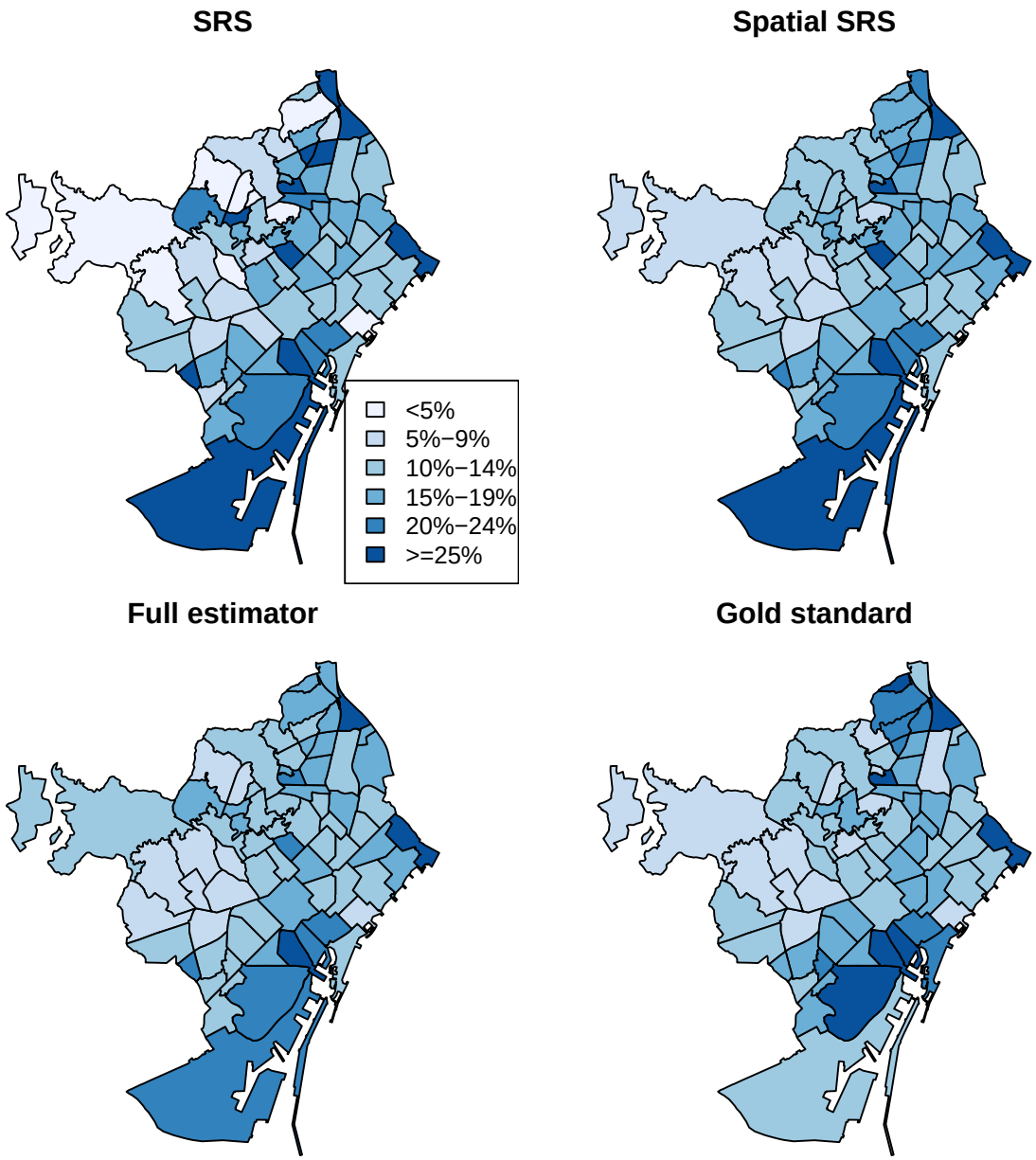

Figure 1: Estimated spatial patterns for the percentage of low- and middle-income countries population for several of the models in this paper, and that same percentage according to the Barcelona census gold standard.

to the MRP approach, which models the data at the individual level or a high level of disagregation, that arising from the combination of the auxiliary/stratifying variables. Subsequent poststratification, in addition to the inclusion of spatial dependence, is the key to derive sensible estimates. However, unlike MRP, we link the use of modeling plus poststratification to arguments from sampling theory. Following this line, similar approaches could be undertaken for different or more complex sampling designs in small-area studies, such as for proportional to size sampling, regression estimators ... These scenarios would deserve similar attention to that given in this paper to stratification or ratio estimators and, in principle, could be successfully modeled following similar procedures. This approach would allow merging spatial statistics, MRP and the statistical analysis of complex sampling designs into a common framework.

Typically, when a sampling design has been followed, the main goal of inference methods is to take that design into account in order to yield appropriate estimates. Nevertheless, the latest of our case studies shows that the use of auxiliary variables that were not used for sampling design (home ownership status in our case) could be even more advantageous in practice than the control of sampling design variables (age group). In addition, we have also shown that spatial dependence could also yield substantial benefits, even greater than the control of sampling design variables, since in our case the spatial estimates clearly outperform stratified and ratio estimates. Evidently, this will probably depend on each specific data set, in particular on whether or not the design variable(s) is(/are) closely related to the outcome variable. However, a clear message seems to arise here: when dealing with small areas, the modeling of the outcome variable, in particular its dependence structure, could be as important as taking the sampling design into account. Therefore, sampling theory alone does not seem to be sufficient to obtain appropriate estimates in small-area studies. In those cases, modeling resources, such as including spatial dependence or auxiliary variables which were not used for the sampling design, are important tools that should be used whenever possible in order to yield sensible estimates.

In this paper we propose modeling the underlying probabilities for each area and group of the auxiliary variable(s) by considering them as spatially dependent, which in principle seems to solve small area estimation problems. Nevertheless, more elaborate spatial modeling could be used in order to yield more flexible estimates for these probabilities. Thus, in the previous section, we proposed modeling those specific group-area probabilities as $logit((\pi^*)_i^{Z=k}) = \mu + \psi_i + \phi_k$, which cannot reproduce interactions between age groups and spatial units. We could alleviate this constraint by modeling $logit((\pi^*)_i^{Z=k}) = \mu_i + \xi_{ik}$, where the rows of $\boldsymbol{\xi}$ are modeled by means of a spatiotemporal (Martinez-Beneito, López-Quílez, and Botella-Rocamora 2008) or a multivariate CAR process (Botella-Rocamora, Martinez-Beneito, and Banerjee 2015), which allows the spatial patterns for the different age groups to be different, though dependent. In any case, the current proposal in this paper, although perhaps somewhat restrictive in comparison to the multivariate alternative, has proven to be interesting in itself, with quite promising results in practice.

Acknowledgements

The authors acknowledge the funding of Instituto de Salud Carlos III (ISCIII) through the project 'PI19/00219', Ministerio de Ciencia e Innovación through the project 'PID2022-136455NBI00/AEI/10.13039/501100011033/FEDER, UE', Dirección General de Ciencia e Investigación (Generalitat Valenciana) through the grant 'CIAICO/2022/165' and co-funding by FEDER grants from the European Union.

Bibliography

Besag, Julian, Jeremy York, and Annie Mollié. 1991. "Bayesian Image Restoration, with Two Applications in Spatial Statistics." *Annals of the Institute of Statistical Mathemathics* 43: 1–21. https://doi.org/10.1007/BF00116466.

Botella-Rocamora, Paloma, Miguel A. Martinez-Beneito, and Sudipto Banerjee. 2015. "A Unifying Modeling Framework for Highly Multivariate Disease Mapping." *Statistics in Medicine* 34 (9): 1548–59. https://doi.org/10.1002/sim.6423.

Chen, Cici, Jon Wakefield, and Thomas Lumely. 2014. "The Use of Sampling Weights in Bayesian Hierarchical Models for Small Area Estimation" 11 (October): 33–43. https://doi.org/10.1016/j.sste.2014.07.002.

Fay, 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 (366): 269–77. http://www.jstor.org/stable/2286322.

Gelman, Andrew. 2007. "Struggles with Survey Weighting and Regression Modeling" 22 (2). https://doi.org/10.1214/088342306000000691.

Gelman, Andrew, Jon B Carlin, Hal S Stern, David B. Dunson, Aki Vehtari, and Donald B Rubin. 2014. *Bayesian Data Analysis*. 3rd ed. Boca Raton: Chapman & Hall/CRC.
Gelman, Andrew, and Thomas C Little. 1997. "Poststratification into Many Categories Using Hierarchical Logistic Regression." *Survey Methodology* 23 (2): 127–35.
Kunsch, Hans R. 1987. "Intrinsic Autoregressions and Related Models on the Two-Dimensional Lattice." *Biometrika* 74: 517–24.
Lawson, Andrew B. 2018. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology (3rd Edition)*. CRC Press.
Leroux, Brian G., Xingye Lei, and Norman Breslow. 1999. "Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence." In *Statistical Models in Epidemiology, the Environment and Clinical Trials*, edited by M E Halloran and D Berry. Berlin Heidelberg New York: Springer.
Li, Zehang, Yuan Hsiao, Jessica Godwin, Bryan D. Martin, Jon Wakefield, and Samuel J. Clark and. 2019. "Changes in the Spatial Distribution of the Under-Five Mortality Rate: Small-Area Analysis of 122 DHS Surveys in 262 Subregions of 35 Countries in Africa." Edited by Olalekan Uthman. *PLOS ONE* 14 (1): e0210645. https://doi.org/10.1371/journal.pone.0210645.
Little, R. J. A. 1993. "Post-Stratification: A Modeler's Perspective" 88 (423): 1001–12. https://doi.org/10.1080/01621459.1993.10476368.
Lohr, Sharon L. 2010. *Sampling: Design and Analysis (2nd Edition)*. Brooks/Cole, Cengage Learning.
Marhuenda, Yolanda, Isabel Molina, and Domingo Morales. 2013. "Small Area Estimation with Spatio-Temporal Fay–herriot Models." *Computational Statistics & Data Analysis* 58 (February): 308–25. https://doi.org/10.1016/j.csda.2012.09.002.
Martinez-Beneito, Miguel A., and P. Botella Rocamora. 2019. *Disease Mapping from Foundations to Multidimensional Modeling*. CRC Press.
Martinez-Beneito, Miguel A., Antonio López-Quílez, and Paloma Botella-Rocamora. 2008. "An Autoregressive Approach to Spatio-Temporal Disease Mapping." *Statistics in Medicine* 27: 2874–89.
Mercer, Laina D., Fred Lu, and Joshua L. Proctor. 2019. "Sub-National Levels and Trends in Contraceptive Prevalence, Unmet Need, and Demand for Family Planning in Nigeria with Survey Uncertainty" 19 (1). https://doi.org/10.1186/s12889-019-8043-z.
Mercer, Laina D., Jon Wakefield, Athena Pantazis, Angelina M. Lutambi, Honorati Masanja, and Samuel Clark. 2015. "Space–time Smoothing of Complex Survey Data: Small Area Estimation for Child Mortality." *The Annals of Applied Statistics* 9 (4). https://doi.org/10.1214/15-aoas872.
Moran, P. A. P. 1948. "The Interpretation of Statistical Maps." *Journal of the Royal Statistical Society B* 10: 243–51.
Paige, John, Geir-Arne Fuglstad, Andrea Riebler, and Jon Wakefield. 2020. "Design- and Model-Based Approaches to Small-Area Estimation in a Low- and Middle-Income Country Context: Comparisons and Recommendations," September. https://doi.org/10.1093/jssam/smaa011.
Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls" 12 (4): 375–85. https://doi.org/10.1093/pan/mph024.
Porter, Aaron T., Scott H. Holan, Christopher K. Wikle, and Noel Cressie. 2014. "Spatial Fay-Herriot Models for Small Area Estimation with Functional Covariates." *Spatial Statistics* 10: 27–42. https://doi.org/https://doi.org/10.1016/j.spasta.2014.07.001.
Vandendijck, Y, C Faes, R A Kirby, A B Lawson, and N Hens. 2016. "Model-Based Inference for Small Area Estimation with Sampling Weights." *Spatial Statistics* 18: 455–73.
Watjou, Kevin, Christel Faes, and Yannick Vandendijck. 2020. "Spatial Modelling to Inform Public Health Based on Health Surveys: Impact of Unsampled Areas at Lower Geographical Scale." *International Journal of Environmental Research and Public Health* 17 (3): 786. https://doi.org/10.3390/ijerph17030786.