Estimating the prevalence of anemia rates among children under five in Peruvian districts with a small sample size

Anna Sikov1,3 · José Cerda-Hernández2,3

Accepted: 10 April 2023 / Published online: 2 May 2023 © Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
In this paper we attempt to answer the following question: “Is it possible to obtain reliable estimates for the prevalence of anemia rates in children under five years in the districts of Peru?” Specifically, the objective of the present paper is to understand to which extent employing the basic and the spatial Fay–Herriot models can compensate for inadequate sample size in most of the sampled districts, and whether the way of choosing the spatial neighbors has an impact on the resulting inference. Furthermore, we explore the question of how to choose an optimal way to define the neighbors. As such, our research focuses on studying the prediction accuracy of the aforementioned models, and on the sensitivity of the results to the definition of “neighbor”. We use the data from the Demographic and Family Health Survey of the year 2019, and the National Census carried out in 2017.

Keywords Direct estimate · Spatial autocorrelation · Fay–Herriot model · Mean square error · Bootstrap

1 Introduction

The prevalence of anemia in young children is an important public health problem. According to the World Health Organization (WHO), anemia is a condition in which the number of red blood cells or the hemoglobin concentration within them is lower than normal, which can cause symptoms such as fatigue, weakness,
dizziness and shortness of breath, among others (Organización Mundial de la Salud 2011; World Health Organization 2004). For this reason, reduction of prevalence of anemia is one of the priorities of the health policies of the Peruvian state. According to “The National Plan for reduction and control of Maternal and Child Anemia and Chronic Child Malnutrition in Peru: 2017–2021”, presented by the Ministry of Health, the target level was the reduction to 19% of anemia in children by the end of 2021. Nonetheless, the prevalence of anemia, reported in 2018 was still 43.5%, which corresponds to a reduction of 3.3%, compared to the rates, observed in 2014 (Ministerio de Salud, Peru 2014, 2017). Evidently, at the current rate of reduction the targeted level of 19% will be attained only by the year 2050. In order to combat the problem of anemia in childhood, the Peruvian Government has implemented various social programs, such as “Vaso de leche”, “Juntos” and “Qali Warma”, the objective of which is to reduce the prevalence of anemia and malnutrition in childhood. One of the most important aims of these programs is to quantify their impact on the reduction of the prevalence of anemia and malnutrition so as to optimize their costs and benefits (see Alcázar 2012 for details). In order to evaluate this impact, good estimates of the percentage of anemic children are needed. However, in the case of Peru, obtaining these estimates, typically presents the most challenges, since there are many remote districts, especially in mountainous regions, which are generally not included in the sample of the surveys due to logistic problems and limited budget; others have a very small sample size (see Fig. 1). We will see below that a possible remedy to this problem would be to use spatial models, which exploit spatial correlations between the neighboring areas. However, populated areas in Peru are mostly located in mountainous regions, and therefore their location can be represented by three coordinates (longitude, latitude and altitude), in contrast to the proposed methods in the literature that use only the first two coordinates. Another problem is that application of the spatial Fay–Herriot model requires definition of the spatial neighbors which is completely subjective. In this study we address the question: “Is it possible to obtain reliable estimates for the prevalence of anemia rates in children under five years in the districts of Peru?” in the presence of the above-mentioned problems.

In this article we utilize the two following sources of data: 1—the data provided by the Demographic and Health Survey—the ENDES, carried out by the National

![Fig. 1 ENDES data: sample size in the districts of the departments of La Libertad (the left panel) and Arequipa (the right panel), where the blank districts do not have available data](image-url)

Springer
Institute of Statistics and Informatics in 2019 (INEI 2019) and 2—the data, obtained from the national census, carried out in 2017. The main objective of the national surveys like the ENDES is to describe some selected population characteristics such as health, employment and unemployment, education, household income and expenses, poverty etc. However one of the common problems of these surveys is that their corresponding sampling design is usually more appropriate for representing characteristics of the entire population, or of large subgrups, such as urban or rural population, the population of major geographical regions, etc. Nonetheless, as noted by Rao and Molina (2015), more and more policy makers are demanding estimates for small domains to use them in the elaboration of policy decisions. In the case of the ENDES, inference at more disaggregated levels, such as provinces or districts is generally not reliable, since at these levels the areas may have small or null sample size. Namely, some of the areas of interest are usually not included in the sample, while the others do not have a sufficient number of observations in order to provide reliable direct estimates, based only on the area-specific sample data. As noted previously, in the case of Peru, the problem is even more pronounced due to limited logistics support and resources. For instance, in Puno region, only 34.5% of the districts data regarding the prevalence of anemia is available. Furthermore, 65.8% of these districts have less than 10 observations.

In order to solve the problem of small sample sizes, the governmental entities like the statistical office of the European Union, United States Census Bureau among many others, utilize the basic Fay–Herriot model (Fay and Herriot 1979), which is the area level model (district-level in our case). Based on this approach, the area level predictions are constructed as a linear combination of standard design-based estimates and indirect model-dependent estimates, where the corresponding regression model incorporates the auxiliary information, which is generally available from the census, administrative records or some other source of data, thus “borrowing strength” across other areas. Thereby, the basic Fay–Herriot model allows the areas to be linked through the vector of the regression coefficients, compensating for their small sample sizes. The variation, which is not explained by the auxiliary variables, is accounted for by the corresponding area-specific random effects. In the case of the basic Fay–Herriot model, these effects are assumed to be independent. A limitation of the basic model is that it is not designed to handle the data that exhibit spatial dependence (Moran 1950) between the areas, which is the typical problem, arising in the data, collected from socio-economic surveys like the ENDES. In such situations, many authors (see for example, (Cressie 1993; Marhuenda et al. 2013; Petrucci and Salvati 2006; Pratesi and Salvati 2008; Singh et al. 2005)) advocate the use of the natural extension of the basic model: the spatial Fay–Herriot model, which incorporates the information about geographical proximity of the areas which, in turn, is utilized to determine the covariance structure of the random effects of the spatially linked areas. More specifically, the random effects are modelled by a simultaneously autoregressive model (SAR), which is characterized by a spatial autoregressive coefficient and a proximity matrix (see (Anselin 1992; Banerjee et al. 2004; Cressie and Chan 1989) and Cressie (1993) for more details). In this way, the expected value of a random effect of a specific area is defined as a linear combination of random effects of the neighboring areas. A drawback of this model is that it
contains some degree of subjectivity, since it depends on the definition of the neighbours, which is apparently not unique. In addition, it should be noted, that including spatial correlation into the model will not result in considerable gain in efficiency if this correlation is not substantially strong (Pfeffermann 2002).

In order to predict the area-specific characteristic of interest, Fay and Herriot (1979) develop the Best Linear Unbiased Predictor (BLUP). As mentioned above, this predictor constitutes a composite estimator, which is derived as the weighted average of the direct area-specific estimator and a corresponding synthetic regression estimator. However, the BLUP can only be obtained if the variances of the random area-specific effects are known. In real applications, this is generally not the case. If the variances are unknown, they are substituted by their corresponding estimates, obtained by maximum likelihood, restricted maximum likelihood or by a method of moments (Fay and Herriot 1979; Kackar and Harville 1984; Prasad and Rao 1990; Rao 2003 and Rao and Molina 2015). The resulting predictor is the empirical BLUP (EBLUP) (Fay and Herriot 1979). In the case of a spatial Fay–Herriot model, a Spatial Best Linear Unbiased Predictor (SBLUP) is used (see Pratesi and Salvati 2008 for details). Replacing the unknown variance and autoregressive parameters by their corresponding estimates in the SBLUP leads to the empirical SBLUP (SEBLUP).

In this article we apply the basic and the spatial Fay–Herriot model in order to predict the percentage of anemic children under 5 years in the districts in Peru. Our main interest is to compare and to evaluate the performance of district-level predictors EBLUP and SEBLUP of the prevalence of anemia rates in the situation where the sampling design is inadequate in the sense that most districts are either not sampled or have a very small sample size, which is a typical problem in emerging and developing countries. As already mentioned, application of the spatial Fay–Herriot model is associated with some degree of subjectivity, introduced by definition of neighboring districts. In order to address this issue we conduct a sensitivity analysis of the results to various definitions of neighbors (see Sect. 4.4). This analysis is helpful to identify the optimal choice of neighboring districts. An obvious strategy is to use the estimator of the residual variance of the model as the criterion of optimality, which identifies the definition of neighbors that produces the lowest value of the estimator as the optimal choice. Another complication that arises in our case is that each district has an additional dimension, namely the altitude. In Sect. 4 we discuss how this additional coordinate can be aggregated in the definition of neighboring districts. Next, we compute the mean square error for the aforementioned predictors. In the case of the basic Fay–Herriot model, we use the Prassad and Rao estimate (Prasad and Rao 1990) for the means square error, and in the case of the spatial Fay–Herriot we implement the parametric and non-parametric bootstrap, developed in Molina et al. (2009).

The rest of the paper is organized as follows. In Sect. 2 the basic and the spatial Fay–Herriot models are presented. In Sect. 3 we briefly describe the problem of estimation of the MSE and provide some references to the most important works in this area. Section 4 illustrates application of the models described in Sect. 2 to our data. This involves estimation of the model parameters, computing the EBLUP and SEPLUB, and estimating their respective MSEs. Furthermore, in this section,
we address the issue of subjectivity in selecting neighboring districts, as well as the challenge posed by the three-dimensional coordinates problem. Finally, Sect. 5 provides some conclusions.

2 Small area estimation models

2.1 Basic Fay–Herriot model

Let \( Y_i \) denote the direct area-level estimate of the characteristic of interest in the \( i \)-th area, where \( i = 1, \ldots, D \) and \( D \) is the total number of the areas with available data, and \( \theta_i \) denotes the corresponding true value of this characteristic. We suppose that \( Y_i \) is design unbiased for \( \theta_i \). Denote by \( X_i = (x_{i1}, \ldots, x_{ip}) \) the vector of \( p \) auxiliary area-level covariates, which can usually be obtained from census or administrative sources. Then, the Fay–Herriot model is defined as follows

\[
Y_i = \theta_i + e_i; \quad \theta_i = X_i \beta + u_i,
\]

(1)

Here \( e_i \sim N(0, \sigma^2_e) \) are the errors of the direct estimates and \( u_i \sim N(0, \sigma^2_u) \) are the area-level random effects, that represent the variability of the \( \theta_i \)'s that is not explained by auxiliary variables, where \( \text{cov}(e_i, e_j) = \text{cov}(u_i, u_j) = 0 \) if \( i \neq j \) and \( \text{cov}(e_i, u_j) = 0 \) \( \forall i, j \); \( \beta \) is the vector of the coefficients that expresses the association between \( \theta = (\theta_1, \ldots, \theta_D)' \) and \( X = (X_1, \ldots, X_D)' \). It is assumed that the sampling error variances \( \sigma^2_e \) are known. This assumption is customary, as it is generally possible to estimate the design variance of sampling errors from the observed data.

Note that the coefficients \( \beta \) do not depend on the area. Specifically, the association between \( X_i \) and \( \theta_i \) is the same for all the areas, and hence the model-based estimate for the characteristic of interest in the \( i \)-th area will incorporate the information about the other areas through the vector of coefficients \( \beta \).

The model (1) can be rewritten as follows:

\[
Y = X\beta + u + e,
\]

(2)

where \( Y = (Y_1, \ldots, Y_D)' \), \( u = (u_1, \ldots, u_D)' \sim N(0, \Sigma_u) \), \( e = (e_1, \ldots, e_D)' \sim N(0, \Sigma_e) \), such that \( \Sigma_u = \sigma^2_u I_D \) and \( \Sigma_e_{ij} = \sigma^2_e \) if \( i = j \) and 0 otherwise, where \( i, j = 1, \ldots, D \).

If \( \sigma^2_u \) were known, \( \theta_i, \quad i = 1, \ldots, D \) could be estimated using the Best Linear Unbiased Predictor (BLUP), developed in Fay and Herriot (1979), as follows.

\[
\hat{\theta}_{BLUP}^i(\sigma^2_u) = X_i \hat{\beta}(\sigma^2_u) + \hat{u}_i(\sigma^2_u),
\]

(3)

Here,

\[
\hat{\beta}(\sigma^2_u) = \left( X'[V(\sigma^2_u)^{-1}X]^{-1}X'\right)^{-1}X'[V(\sigma^2_u)^{-1}]Y,
\]

(4)
where

$$V(\sigma_u^2) = \text{Var}(u + e) = \Sigma_u + \Sigma_e$$

and

$$\gamma_i(\sigma_u^2) = \frac{\sigma_u^2}{\sigma_i^2 + \sigma_u^2}.$$ 

Alternatively, the predictor (3) can be presented as

$$\hat{\theta}_{BLUP}^i(\sigma_u^2) = \gamma_i(\sigma_u^2)Y_i + (1 - \gamma_i(\sigma_u^2))X_i\hat{\beta}(\sigma_u^2)$$

(6)

Note that the predictor (6) constitutes a convex combination of the direct estimate $Y_i$ and the model-based estimate $X_i\hat{\beta}$. Obviously, if the $i$th area does not have available data, its corresponding value of $\gamma_i$ is equal to zero, and therefore the prediction of $\theta_i$ for this area is equal to the model-based estimator.

In most real data applications, the value of the parameter $\sigma_u^2$ is unknown. In this case, $\sigma_u^2$ is usually estimated by means of maximum likelihood (ML) or restricted maximum likelihood (REML).

The log-likelihood function is obtained as

$$l_{ML}(\beta, \sigma_u^2) = c - \frac{1}{2} \log |V| - \frac{1}{2}(Y - X\beta)V^{-1}(Y - X\beta)'$$

(7)

where $c$ is some constant and $V = V(\sigma_u)$. Given function is maximized with respect to $\sigma_u^2$, whereas the parameters $\beta$ are estimated as (4).

The restricted log-likelihood function is defined as

$$l_{REML}(\sigma_u^2) = c' - \frac{1}{2} \log |V| - \frac{1}{2} \log |X'V^{-1}X| - \frac{1}{2} Y'PY,$$

(8)

where $c'$ is some constant, $V = V(\sigma_u)$ and $P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$.

Contrary to the ML, the REML takes into account the loss of degrees of freedom due to estimation of the parameters $\beta$, and consequently, it is advantageous in the case of small sample sizes (Molina et al. 2009; Rao 2003; Rao and Molina 2015).

Alternatively, the method of moments, developed by Prasad and Rao (see Prasad and Rao 1990), or the method, proposed by Fay and Herriot (see Fay and Herriot 1979 for details) can be employed.

It is important to emphasize that all mentioned estimators for $\sigma_u^2$ are translation invariant, that is, have the following properties (see Kackar and Harville 1984 for more details):

1. $\hat{\sigma}_u^2(Y) = \hat{\sigma}_u^2(-Y)$
2. $\hat{\sigma}_u^2(Y - Xa) = \hat{\sigma}_u^2(Y), \forall a \in R^p$ and $\forall Y$.

In Kackar and Harville (1984), the authors show that the empirical BLUP $\hat{\theta}_{EBLUP}^i$, which is defined in Fay and Herriot (1979) as
is unbiased for \( \theta_i \) if a consistent estimator \( \hat{\sigma}_u^2 \) is translate invariant.

As discussed previously, if the data present strong spatial correlations, a spatial Fay–Herriot model is a natural way to proceed. This model is described in the following subsection.

### 2.2 Spatial Fay–Herriot model

The spatial Fay–Herriot model is defined as follows (see Pratesi and Salvati 2008 for more details):

\[
Y = X\beta + u + e; \quad u = \rho Wu + \epsilon,
\]

where \( \epsilon = (\epsilon_1, \ldots, \epsilon_D)' \sim N(0, \Sigma_e) \) such that \( \Sigma_e = \sigma_e^2 I_D \), \( \rho \) is the spatial autoregressive coefficient (see Banerjee et al. 2004; Cressie and Chan 1989; Cressie 1993), and \( W \) is a matrix of non-negative spatial weights, the elements \( w_{ij} \) of which define the spatial measure of proximity between the areas \( i \) and \( j \), such that \( \forall i = 1, \ldots, D, w_{ii} = 0 \) and \( \sum_{j=1}^D w_{ij} = 1 \). As noted above, the weights \( w_{ij} \) can be defined in a variety of ways. Typically, \( w_{ij} \) depend on the definition of the neighboring areas. However, it must be noted that, it is hard to formulate specific criteria to choose the "best" definition. Here we present a few common approaches to define neighboring areas of a specific area \( i \) (the interested readers can refer to Anselin (1992) and Cressie (1993) for more details).

1. Those areas, whose distance between their corresponding centroids and the centroid of the area of interest is within \( L \) miles. For example, Cressie and Chan (1989) define two areas as neighbors if the distance between their centroids is within 30 miles.
2. The \( k \) nearest areas to the area of interest.
3. Areas that share a common boundary with the area of interest.

Clearly, it is important to use caution when defining the neighbors, since different definitions may produce different results.

Now, the model (10) can be written as:

\[
Y = X\beta + (I - \rho W)^{-1}e + e = X\beta + v, \quad v \sim N(0, G),
\]

where

\[
G = \sigma_e^2 [(I - \rho W)'(I - \rho W)]^{-1} + \Sigma_e = \Omega + \Sigma_e.
\]

Note that the matrix \( G \) exists only if \( (I - \rho W) \) is non-singular.

Next, let \( \phi = (\sigma_e^2, \rho) \) index the unknown model parameters, and \( b_i = (0, \ldots, 0, 1, 0, \ldots, 0)' \) be a D-dimensional vector with value 1 in the \( i \)th position and 0 in all other positions. Therefore, the spatial BLUP (SBLUP) for \( \theta_i \) is obtained as:
The estimates of the unknown parameters \( \phi \) can be obtained using ML or REML, where the covariance matrix \( V \) in (7) or (8) is replaced by the matrix \( G(\phi) \). Molina et al. (2009) warn about possible numeric problems, associated with optimization of the functions (7) and (8) in this case.

Replacing the parameters \( \phi \) with their corresponding estimates, \( \hat{\phi} \) in (13) and in (14), we obtain the empirical SBLUP (SEBLUP) for \( \theta_i \), which is given by

\[
\hat{\theta}_i^{SEBLUP}(\hat{\phi}) = X_i\hat{\beta}(\hat{\phi}) + \hat{u}_i(\hat{\phi}),
\]

The predictor (15) is unbiased for \( \theta_i \) if \( \hat{\sigma}_u^2 \) and \( \hat{\rho} \) are derived using ML or REML (see Kackar and Harville 1984 for more details).

### 3 Estimation of the mean square error of EBLUP and SEBLUP

In real applications, a natural question of interest is how to estimate the mean square error (MSE) of the predictors (9) and (15). In this section we present a brief review of the main estimation methods that have been proposed in the literature to address this problem. We start with analyzing the MSE of the BLUP (3). It can be easily shown that

\[
MSE(\hat{\theta}_i^{BLUP}(\sigma_u^2)) = \gamma_i(\sigma_u^2)\sigma_i^2 + (1 - \gamma_i(\sigma_u^2))^2 X_i Var(\hat{\beta}(\sigma_u^2))X_i^t
\]

\[
= g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2),
\]

where \( X_i \) is the \( i \)th line of the matrix \( X \) and \( \hat{\beta}(\sigma_u^2) \) is the estimate for \( \beta \), defined in (4).

The estimate for the MSE defined in (16) is obtained by replacing \( \sigma_u^2 \) with the estimate \( \hat{\sigma}_u^2 \), as follows.

\[
\text{mse}(\hat{\theta}_i^{BLUP}(\sigma_u^2)) = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2),
\]

It should be noticed that (16) and (17) do not account for the error associated with the estimation of the parameter \( \sigma_u^2 \). It can be demonstrated that (see Kackar and Harville 1984; Harville and Jeske 1992) if the sampling errors and the area-level random effects have a normal distribution, and the estimator for \( \sigma_u^2 \) is translation invariant, the MSE can be decomposed as:
\[ \text{MSE}(\hat{\theta}_i^{\text{SEBLUP}}(\sigma_u^2)) = \text{MSE}(\hat{\theta}_i^{\text{BLUP}}(\sigma_u^2)) + E(\hat{\theta}_i^{\text{SEBLUP}}(\sigma_u^2) - \hat{\theta}_i^{\text{BLUP}}(\sigma_u^2))^2 \]  

(18)

The second term in the expression (18) represents the additional error which is the result of the estimation of the parameter \( \sigma_u^2 \). Contrary to the first term, the second term can not be expressed analytically, and therefore, can only be obtained by approximation, see for example (Datta and Lahiri 2000; Datta et al. 2005; Fay and Herriot 1979; Prasad and Rao 1990).

Alternatively, the MSE can be estimated utilizing resampling methods, such as the bootstrap and jackknife (see Chen and Lahiri 2003; Hall and Maiti 2006; Jiang et al. 2002 among many others).

If the spatial Fay–Herriot model is used, an additional parameter \( \rho \) is to be estimated. As noted previously, unknown parameters \( \phi = (\sigma_u^2, \rho) \) can be estimated using the method of ML or REML. As in the previous case the MSE of \( \hat{\theta}_i^{\text{SEBLUP}} \) can be decomposed as Molina et al. (2009), Pratesi and Salvati (2008) and Singh et al. (2005):

\[ \text{MSE}(\hat{\theta}_i^{\text{SEBLUP}}(\phi)) = \text{MSE}(\hat{\theta}_i^{\text{SEBLUP}}(\phi)) + E(\hat{\theta}_i^{\text{SEBLUP}}(\phi) - \hat{\theta}_i^{\text{SEBLUP}}(\phi))^2 \]

(19)

where the term \( g_1(\phi) \) represents the error produced by the estimation of the random effects, and the term \( g_2(\phi) \) represents the error produced by the estimation of the parameters \( \rho \) (see Singh et al. 2005). If the parameters \( \phi \) are estimated by means of REML, the estimate for the MSE is approximately unbiased and is given by

\[ \text{mse} \left( \hat{\theta}_i^{\text{SEBLUP}}(\phi) \right) \approx g_1(\hat{\phi}) + g_2(\hat{\phi}) + 2g_3(\hat{\phi}) \]

(20)

If ML is used for estimation of \( \phi \), the expression for the estimate of the MSE includes an extra term, which corrects for the additional bias of \( g_1(\phi) \) (see Molina et al. 2009; Pratesi and Salvati 2008, 2009; Singh et al. 2005 for details).

The expressions of \( g_1(\phi) \) and \( g_2(\phi) \) can be obtained analytically (computational details can be found in Singh et al. (2005)), whereas for the term \( g_3(\phi) \) which represents the error due to estimating the parameters \( \phi \), no analytic form can be derived. In Pratesi and Salvati (2009) the authors propose a heuristic approximation for \( g_3(\phi) \). Alternatively, a bootstrap method can be adopted in order to estimate \( g_3(\phi) \) (see Pfeffermann and Tiller 2005), or the whole expression for \( \text{MSE} \left( \hat{\theta}_i^{\text{SEBLUP}}(\phi) \right) \) (see Molina et al. 2009).

4 A case study

4.1 Objectives of the study

In this section we illustrate and study the performance of the basic and spatial Fay–Herriot models using data collected as part of the Demographic and Health Survey - ENDES, carried out by the National Institute of Statistics and Informatics.
in 2019. The survey collects information on the topics such as anemia, nutrition, education, domestic violence among many others. The sampling units in this survey are households, which were sampled by a two-stage sampling design: at the first stage, a sample of localities was selected; at the second stage, a sample of dwellings was chosen within each of the selected localities. A household is defined as a group of people living in the same dwelling and sharing the same budget for food expenditure. In this study we focus on modeling the prevalence of anemia rates in children under five, per district. As it has been pointed out previously, direct estimates are unreliable for most of the sampled districts. Our main aim is to study gain in precision of the estimates obtained by employing the aforementioned models. Specifically, we focus on the following two points. First, we address the question of choosing the neighbor criterion to be used. Second, we compare the MSE and the coefficient of variation of the predictors EBLUP and SEBLUP obtained by application of the basic and the spatial Fay–Herriot model, respectively. The auxiliary covariates used in the models are the characteristics of the district, obtained from the National Census carried out in 2017, as displayed in Table 1.

Initially, we fitted the basic (1) and the spatial (11) Fay–Herriot models to all districts with available direct estimates. However, for both models, the standardized residuals exhibited a far from normal distribution. In order to remedy this problem, we divided all the districts into the following three groups: 1—the districts, where less than 30% of the population live in poverty (a total of 585 districts, 281 sampled districts), 2—the districts where 30–55% of the population live in poverty (a total of 671 districts, 297 sampled districts) and 3—the districts where more than 55% of the population live in poverty (a total of 618 districts, 234 sampled districts), and fit the aforementioned models in each of the specified groups separately. This stratification can be justified by noting that most of the districts with poverty levels below 30% are predominantly located in relatively large cities, whereas districts with poverty rates of 55% or higher are primarily situated in mountainous villages. Consequently, these sub-populations are markedly dissimilar in numerous aspects. Table 2 displays the means and standard deviations of the auxiliary variables in each of the

| Table 1 | Description of the auxiliary variables |
|---------|---------------------------------------|
| Variable | Description of the variable           |
| Altitude | The altitude of the district (height above sea level) |
| Water   | % of dwellings with access to centralized water supply |
| Water-days | % of dwellings with access to potable water only several days per week |
| Floor   | % of dwellings that have non-dirt flooring |
| Internet | % of dwellings with access to internet |
| SIS     | % of the population that is affiliated with the Comprehensive Health Insurance (SIS) |
| Uninsur | % of the population that do not have health insurance |
| Refrig  | % of households that have a refrigerator |
| Spanish | % of native Spanish speakers |
| Rural   | % of rural dwellings |
defined groups. Inspecting the results presented in Table 2 it can be concluded that
the variables that represent the major differences between the districts with high,
average and low poverty rates are Altitude, Floor, SIS, Refrig., Spanish and Rural. It
is noteworthy that the variable Water does not exhibit any conspicuous differences
between the groups. This phenomenon may be attributed to a lack of adequate water
supply in the majority of districts in Peru, regardless of their poverty status.

Next, we compare the estimators for the MSE of the EBLUP and SEBLUP,
defined by (9) and (15) correspondingly. In the case of the EBLUP, we utilize
the estimator proposed by Prasad and Rao (1990). In order to obtain the estimator for
the MSE of the SEBLUP we use the parametric and non-parametric bootstrap,
proposed in Molina et al. (2009).

### 4.2 Definition of neighboring districts

In what follows the neighbors of a specific district are defined in two steps. In
the first step, $K_1$ nearest neighbors are chosen, using districts’ latitude and longi-
tude, where $K_1 = 3, 4, \ldots, 10$. It should be noticed that another two ways to define
neighbors, mentioned in Sect. 2.2 are inapplicable in our case due to a large num-
ber of non-sampled districts. In the second step we use difference in altitude as
the measure of proximity between each of the $K_1$ previously selected districts and
the district of interest. In this step we choose $K_2 \leq K_1$ "closest" districts. The spa-
tial weights of each of the $K_2$ districts selected in the second step is equal to $1/K_2$,
while the spatial weights of all other district are equal to 0. For this study we
use $K_2 = 1, \ldots, K_1$. Then, for each pair $(K_1, K_2)$ we analyze the fit of the spatial
Fay–Herriot model. Specifically, we study the behavior of the estimator for the variance of the model errors, $\hat{\sigma}^2$ as a function of $(K_1, K_2)$. Obviously, the optimal
definition of neighbors corresponds to the values of $K_1$ and $K_2$ which results in
the lowest value of $\hat{\sigma}^2$. It should be noted that in the second step the proximity (or
similarity) between neighboring districts can be expressed using other variables,
for example, the poverty level or human development index in the district. This additional information can be potentially useful, especially in the case where many areas have small or very small sample size. In this study in addition to the variable “Altitude” we use the variables “Poverty” and “Extreme Poverty” which stand for the percentage of the population living in poverty and extreme poverty, respectively.

### 4.3 Fitting the basic Fay–Herriot models

Initially, we present the results of fitting the basic Fay–Herriot model. Table 3 shows the estimated coefficients $\hat{\beta}$ of the model and their corresponding $p$-values, as obtained when fitting the model separately to each of the three defined groups of the districts.

Table 3 indicates that the prevalence of anemia in a district is apparently associated with the variables that reflect the poverty level of that district. It should be noted that many other auxiliary variables that also reflect the poverty level in a district, such as the percentage of dwellings with concrete walls, the percentage of dwellings with access to centralized hygiene system, the percentage of illiterate population etc., were initially included in the model. However, their respective coefficients were non-significant in all three cases. It is important to emphasize that only significant variables were retained in the models for the analyses performed in Sects. 4.4 and 4.6. It is worth noting that the estimates of the coefficients of some variables vary across the groups.

|                | Less than 30% | 30%-55% | More than 55% |
|----------------|---------------|----------|---------------|
| (281 districts)| (297 districts)| (234 districts) |
| **Estimator** | **p-value** | **Estimator** | **p-value** | **Estimator** | **p-value** |
| Water         | 0.01464      | 0.7657   | –            | 0.12398      | 0.0086   | –            | 0.16061      | 0.0028 |
| Water-days    | 0.13637      | 0.0049   | –            | 0.04564      | 0.4965   | 0.19523      | 0.0297 |
| Floor         | 0.24208      | 0.0027   | –            | 0.03398      | 0.5654   | –            | 0.03302      | 0.6755 |
| Refrigerator  | – 0.13299    | 0.0967   | –            | 0.23934      | 0.0123   | –            | 0.35556      | 0.0043 |
| Internet      | – 0.20771    | 0.0061   | –            | 0.25977      | 0.0692   | 0.09958      | 0.6244 |
| Spanish       | – 0.22165    | <0.0001  | –            | 0.20084      | <0.0001  | –            | 0.23646      | <0.0001 |
| SIS           | 0.03799      | 0.6848   | –            | 0.29200      | 0.0124   | –            | 0.28159      | 0.0651 |
| Uninsure      | – 0.16627    | 0.0341   | –            | 0.10550      | 0.5181   | 0.11754      | 0.5379 |
| Altitude      | – 0.01516    | 0.0011   | –            | 0.01034      | 0.1281   | 0.01605      | 0.0765 |
| Rural         | 0.07947      | 0.0478   | –            | 0.05782      | 0.1363   | 0.02077      | 0.6591 |
Table 4 Values of $\hat{\sigma}^2_e$ as a function of $K_1$ and $K_2$: districts with less than 30% of population living in poverty (a total of 281 districts)

| $K_1$ | $K_2 = 1$ | $K_2 = 2$ | $K_2 = 3$ | $K_2 = 4$ | $K_2 = 5$ | $K_2 = 6$ | $K_2 = 7$ | $K_2 = 8$ | $K_2 = 9$ | $K_2 = 10$ |
|-------|-----------|------------|------------|------------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1     | 0.0047    | –          | –          | –          | –          | –          | –          | –          | –          | –          |
| 2     | 0.0052    | **0.0041** | –          | –          | –          | –          | –          | –          | –          | –          |
| 3     | 0.0054    | **0.0039** | 0.0045     | –          | –          | –          | –          | –          | –          | –          |
| 4     | 0.0052    | 0.0048    | 0.0045     | 0.0045     | –          | –          | –          | –          | –          | –          |
| 5     | 0.0048    | 0.0044    | 0.0043     | 0.0044     | 0.0048    | –          | –          | –          | –          | –          |
| 6     | 0.0044    | 0.0047    | 0.0047     | 0.0048     | 0.0050    | 0.0051    | –          | –          | –          | –          |
| 7     | 0.0049    | 0.0045    | 0.0046     | 0.0049     | 0.0051    | 0.0052    | 0.0051    | –          | –          | –          |
| 8     | 0.0051    | 0.0043    | 0.0047     | 0.0051    | 0.0051    | 0.0051    | 0.0051    | –          | –          | –          |
| 9     | 0.0047    | 0.0042    | 0.0046     | 0.0048    | 0.0050    | 0.0051    | 0.0051    | 0.0051    | 0.0052    | –          |
| 10    | 0.0053    | 0.0043    | 0.0047     | 0.0050    | 0.0051    | 0.0050    | 0.0051    | 0.0051    | 0.0052    | 0.0052    |

Numbers in bold correspond to the lowest values attained by the estimator $\hat{\sigma}^2_e$ in the two following cases: $1 - K_2 < K_1$ and $2 - K_2 = K_1$

Table 5 Values of $\hat{\sigma}^2_e$ as a function of $K_1$ and $K_2$: districts where 30% - 55% of population living in poverty (a total of 297 districts)

| $K_1$ | $K_2 = 1$ | $K_2 = 2$ | $K_2 = 3$ | $K_2 = 4$ | $K_2 = 5$ | $K_2 = 6$ | $K_2 = 7$ | $K_2 = 8$ | $K_2 = 9$ | $K_2 = 10$ |
|-------|-----------|------------|------------|------------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1     | 0.0033    | –          | –          | –          | –          | –          | –          | –          | –          | –          |
| 2     | 0.0037    | **0.0027** | –          | –          | –          | –          | –          | –          | –          | –          |
| 3     | 0.0022    | **0.0021** | 0.0035     | –          | –          | –          | –          | –          | –          | –          |
| 4     | 0.0026    | 0.0032    | 0.0039     | 0.0036     | –          | –          | –          | –          | –          | –          |
| 5     | 0.0039    | 0.0036    | 0.0038     | 0.0039    | 0.0041    | –          | –          | –          | –          | –          |
| 6     | 0.0049    | 0.0043    | 0.0044     | 0.0043    | 0.0045    | 0.0045    | –          | –          | –          | –          |
| 7     | 0.0036    | 0.0038    | 0.0042     | 0.0042    | 0.0047    | 0.0048    | 0.0049    | –          | –          | –          |
| 8     | 0.0047    | 0.0037    | 0.0040     | 0.0044    | 0.0045    | 0.0047    | 0.0049    | 0.0049    | –          | –          |
| 9     | 0.0048    | 0.0040    | 0.0044     | 0.0045    | 0.0046    | 0.0048    | 0.0048    | 0.0048    | 0.0049    | –          |
| 10    | 0.0042    | 0.0039    | 0.0046     | 0.0047    | 0.0049    | 0.0049    | 0.0049    | 0.0050    | 0.0051    | –          |

Numbers in bold correspond to the lowest values attained by the estimator $\hat{\sigma}^2_e$ in the two following cases: $1 - K_2 < K_1$ and $2 - K_2 = K_1$

Table 6 Values of $\hat{\sigma}^2_e$ as a function of $K_1$ and $K_2$: districts with more than 55% of population living in poverty (a total of 234 districts)

| $K_1$ | $K_2 = 1$ | $K_2 = 2$ | $K_2 = 3$ | $K_2 = 4$ | $K_2 = 5$ | $K_2 = 6$ | $K_2 = 7$ | $K_2 = 8$ | $K_2 = 9$ | $K_2 = 10$ |
|-------|-----------|------------|------------|------------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1     | 0.0078    | –          | –          | –          | –          | –          | –          | –          | –          | –          |
| 2     | 0.0079    | 0.0077    | –          | –          | –          | –          | –          | –          | –          | –          |
| 3     | 0.0077    | 0.0082    | 0.0076     | –          | –          | –          | –          | –          | –          | –          |
| 4     | 0.0083    | 0.0082    | 0.0075     | 0.0075     | –          | –          | –          | –          | –          | –          |
| 5     | 0.0082    | 0.0082    | 0.0066     | 0.0053    | 0.0046    | –          | –          | –          | –          | –          |
| 6     | 0.0081    | 0.0065    | 0.0046     | 0.0048    | 0.0045    | 0.0043    | –          | –          | –          | –          |
| 7     | 0.0077    | 0.0046    | **0.0040** | 0.0045    | 0.0041    | 0.0042    | **0.0042** | –          | –          | –          |
| 8     | 0.0078    | 0.0063    | 0.0046     | 0.0048    | 0.0042    | 0.0043    | 0.0045    | 0.0045    | –          | –          |
| 9     | 0.0082    | 0.0043    | 0.0044     | 0.0048    | 0.0046    | 0.0046    | 0.0046    | 0.0046    | 0.0048    | –          |
| 10    | 0.0082    | 0.0051    | 0.0044     | 0.0045    | 0.0045    | 0.0046    | 0.0046    | 0.0047    | 0.0048    | 0.0049    |

Numbers in bold correspond to the lowest values attained by the estimator $\hat{\sigma}^2_e$ in the two following cases: $1 - K_2 < K_1$ and $2 - K_2 = K_1$
4.4 Sensitivity analysis

In this Section we conduct a sensitivity analysis to investigate the impact of selecting neighboring districts. To this end, the spatial Fay–Herriot model was fitted with $K_1 = 1, \ldots, 10$ and $K_2 = 1, \ldots, K_1$ neighbors, as explained in Sect. 4.2. The figures in Tables 4, 5 and 6 suggest that the results are sensitive to the way in which the neighbours were defined. Furthermore, it should be noted that the estimators of the parameter $\rho$ vary quite widely with the choice of $K_1$ and $K_2$ (from 0.15 to 0.87). These results demonstrate that the way of choosing of the neighbors can dramatically alter inferences. In this situation we recommend using the values of $K_1$ and $K_2$ that correspond to the minimal value of $\hat{\sigma}_\epsilon^2$. The results displayed in the tables, illustrate that the optimal choice of neighbors in the case of the districts of the first two groups is $K_1 = 3$ and $K_2 = 2$, whereas for the third group the optimal values are $K_1 = 7$ and $K_2 = 3$. At the same time, the tables show that if the variable “Altitude” is not utilized, which implies $K_2 = K_1$, the optimal value of $K_1$ in the case of the first two groups is $K_1 = 2$, while for the third group $K_1 = 7$. Comparing the corresponding magnitudes of $\hat{\sigma}_\epsilon^2$, it can be observed that incorporating the variable “Altitude” leads to a minor reduction of 5% (from 0.0041 to 0.0039 and from 0.0042 to 0.0040) in the first and the third group, and of 22% (from 0.0027 to 0.0022) in the second group.

As we have already mentioned, for the purpose of selecting $K_2$ districts, out of the $K_1$ previously selected districts, the variable “Altitude” is not the only variable that can be utilized in order to establish the degree of similarity between the districts. In Table 7 the results obtained in the case of utilizing the variables “Altitude”, “Poverty” and “Extreme Poverty” are summarized. Furthermore, Table 7 presents the outcomes of the optimal choice of $K_1$ if only the first step was utilized, namely, $K_2 = K_1$.

It can be concluded from Table 7 that the use of the variable “Extreme Poverty” had some beneficial effect in the case of the first and the third group, while in the second group we would recommend to use the variable “Altitude”. Notably, the variable “Extreme Poverty” was not significant in the models presented in Table 3. In summary, it can be inferred that choosing $K_2$ districts in

| Variable   | Less than 30% (281 districts) | 30%-55% (297 districts) | More than 55% (234 districts) |
|------------|--------------------------------|--------------------------|-------------------------------|
|            | Opt. $K_1, K_2$ | $\hat{\sigma}_\epsilon^2$ | Opt. $K_1, K_2$ | $\hat{\sigma}_\epsilon^2$ | Opt. $K_1, K_2$ | $\hat{\sigma}_\epsilon^2$ |
| Altitude   | (3, 2)          | 0.0039                    | (3, 2)          | 0.0021                    | (7, 3)          | 0.0040                    |
| Poverty    | (3, 2)          | 0.0045                    | (2, 1)          | 0.0029                    | (7, 2)          | 0.0034                    |
| Ex. Poverty| (3, 2)          | 0.0033                    | (2, 1)          | 0.0027                    | (7, 3)          | 0.0032                    |
| –          | (2, 2)          | 0.0041                    | (2, 2)          | 0.0027                    | (7, 7)          | 0.0041                    |
Estimating the prevalence of anemia rates among children under…

Table 8 Estimates for the parameters $\rho$, $\sigma^2_s$, $\beta_s$ and $\beta_w$ for various definitions of neighbors

| District | Altitude | Ex. Poverty |
|----------|----------|-------------|
|          | Altitude | Ex. Poverty |
|          | (3, 2)   | (6, 4)   | (10, 6) | (3, 2)   | (6, 4)   | (10, 6) |
|          | $\hat{\rho}$ | $\hat{\sigma}_c^2$ | $\hat{\beta}_s$ | $\hat{\beta}_w$ | $\hat{\beta}_s$ | $\hat{\beta}_w$ |
|          | 0.0385 | 0.0082 | -0.236 | -0.106 | 0.0082 | 0.0046 | -0.217 | -0.081 | 0.0082 | 0.0046 | -0.197 | -0.071 | 0.0082 | 0.0046 | -0.197 | -0.071 |
|          | 0.6966 | 0.0048 | -0.217 | -0.081 | 0.0048 | 0.0046 | -0.224 | -0.071 | 0.0048 | 0.0046 | -0.197 | -0.071 | 0.0048 | 0.0046 | -0.197 | -0.071 |
|          | 0.8199 | 0.0073 | -0.234 | -0.101 | 0.0073 | 0.0046 | -0.224 | -0.071 | 0.0073 | 0.0046 | -0.197 | -0.071 | 0.0073 | 0.0046 | -0.197 | -0.071 |
|          | 0.3025 | 0.0045 | -0.234 | -0.101 | 0.0045 | 0.0046 | -0.224 | -0.071 | 0.0045 | 0.0046 | -0.197 | -0.071 | 0.0045 | 0.0046 | -0.197 | -0.071 |
|          | 0.7253 | 0.0044 | -0.197 | -0.071 | 0.0044 | 0.0046 | -0.197 | -0.071 | 0.0044 | 0.0046 | -0.197 | -0.071 | 0.0044 | 0.0046 | -0.197 | -0.071 |
|          | 0.8191 | 0.0044 | -0.197 | -0.071 | 0.0044 | 0.0046 | -0.197 | -0.071 | 0.0044 | 0.0046 | -0.197 | -0.071 | 0.0044 | 0.0046 | -0.197 | -0.071 |
|          | 0.3641 | 0.0068 | -0.233 | -0.103 | 0.0068 | 0.0046 | -0.212 | -0.076 | 0.0068 | 0.0046 | -0.200 | -0.081 | 0.0068 | 0.0046 | -0.200 | -0.081 |
|          | 0.7420 | 0.0045 | -0.200 | -0.081 | 0.0045 | 0.0046 | -0.200 | -0.081 | 0.0045 | 0.0046 | -0.200 | -0.081 | 0.0045 | 0.0046 | -0.200 | -0.081 |
|          | 0.8231 | 0.0045 | -0.200 | -0.081 | 0.0045 | 0.0046 | -0.200 | -0.081 | 0.0045 | 0.0046 | -0.200 | -0.081 | 0.0045 | 0.0046 | -0.200 | -0.081 |

Table 9 Estimates for the prevalence of anemia rates in six different districts, for various definitions of neighbors

| District | Altitude |
|----------|----------|
|          | Ex. Poverty |
|          | (3, 2) | (6, 4) | (10, 6) | (3, 2) | (6, 4) | (10, 6) |
| Maino Amazonas | 0.197 | 0.177 | 0.165 | 0.204 | 0.173 | 0.177 |
| Amashca Ancash | 0.493 | 0.449 | 0.462 | 0.459 | 0.465 | 0.480 |
| Huasmin Cajamarca | 0.261 | 0.221 | 0.223 | 0.255 | 0.236 | 0.234 |
| Cusipata Cusco | 0.532 | 0.581 | 0.585 | 0.558 | 0.600 | 0.594 |
| L. Prado Lima | 0.288 | 0.335 | 0.345 | 0.298 | 0.347 | 0.343 |
| Itauta Puno | 0.740 | 0.689 | 0.673 | 0.705 | 0.669 | 0.673 |

Table 10 Mean Square Errors of the estimates for the prevalence of anemia rates in six different districts, for various definitions of neighbors

| District | Altitude |
|----------|----------|
|          | Ex. Poverty |
|          | (3, 2) | (6, 4) | (10, 6) | (3, 2) | (6, 4) | (10, 6) |
| Maino Amazonas | 0.00616 | 0.00474 | 0.00439 | 0.00596 | 0.00451 | 0.00443 |
| Amashca Ancash | 0.00697 | 0.00515 | 0.00506 | 0.00619 | 0.00498 | 0.00513 |
| Huasmin Cajamarca | 0.00694 | 0.00500 | 0.00450 | 0.00643 | 0.00466 | 0.00459 |
| Cusipata Cusco | 0.00660 | 0.00503 | 0.00490 | 0.00593 | 0.00495 | 0.00488 |
| L. Prado Lima | 0.00551 | 0.00426 | 0.00386 | 0.00493 | 0.00379 | 0.00378 |
| Itauta Puno | 0.00706 | 0.00569 | 0.00559 | 0.00679 | 0.00605 | 0.00538 |

the second step using an additional variable to measure similarity between the $K_1$ previously selected districts, can potentially produce more powerful predictors (see Sect. 4.6). Next, we evaluate the sensitivity of the resulting inference to the method of defining neighbors. To this end, we fit the spatial Fay–Herriot model using different definitions of neighborhood, estimate the model parameters $\beta$, $\rho$ and $\sigma^2_s$ and compute the SEBLUP estimates for the prevalence of anemia rates along with corresponding MSEs. The neighbors were determined using the procedure described in Sect. 4.2 with the variables Altitude, Poverty and Ex. Poverty.
employed in the second step. For each variable we consider the following values of \((K_1, K_2): (3,2),(6,4)\) and \((10,6)\), which produce 9 different neighbor definitions. Tables 8, 9 and 10 display results for districts with poverty levels of 55% or more, while implications for the two other groups are similar. Table 8 presents parameter estimates for \(\rho, \sigma^2, \beta_S\) and \(\beta_W\), for 9 different ways of choice of neighboring districts described above, where \(\beta_S\) and \(\beta_W\) represent the coefficients of the variables Spanish and Water respectively. The coefficients of other variables are not shown, but have very similar behavior. Tables 9 and 10 summarize the SEBLUP estimates and their corresponding MSEs for various definitions of neighboring districts, respectively. Here, we do not demonstrate the results obtained in the case where the variable Poverty is utilized in the second step due to their similarity to those reported if the variable Ex. Poverty is employed. The results are presented for six districts belonging to six different departments of Peru.

The results displayed in the Tables 8, 9 and 10 may indicate that, if two different definitions of neighboring districts produce similar estimates of \(\hat{\sigma}^2\), then the resulting inferences should also be similar. This implies that searching for neighboring district definitions that yield the lowest values of \(\hat{\sigma}^2\) may be a reasonable strategy. However, in order to draw more precise conclusions, a more thorough analysis is required.

### 4.5 Spatial Fay–Herriot model

In what follows we fit the Spatial Fay–Herriot model for the following two scenarios.

| Table 11 | Estimates of the coefficients \(\beta\) and \(\rho\) when fitting the spatial Fay–Herriot model under the first scenario |
|----------|-------------------------------------------------|
|          | Less than 30% (281 districts) | 30%-55% (297 districts) | More than 55% (234 districts) |
|          | Estimator | \(p\)-value | Estimator | \(p\)-value | Estimator | \(p\)-value |
| Water    | 0.02746 | 0.5601 | – 0.11775 | 0.0154 | – 0.10223 | 0.0537 |
| Water-days | 0.14083 | 0.0040 | 0.00120 | 0.9849 | 0.17271 | 0.0309 |
| Floor    | 0.20302 | 0.0006 | 0.03356 | 0.5613 | 0.07551 | 0.3485 |
| Refrig   | – 0.10396 | 0.1909 | – 0.23956 | 0.0091 | – 0.40809 | 0.0005 |
| Internet | – 0.16100 | 0.1336 | – 0.18821 | 0.1525 | 0.12318 | 0.4997 |
| Spanish  | – 0.21124 | < 0.0001 | – 0.18363 | < 0.0001 | – 0.17702 | < 0.0001 |
| SIS      | 0.03798 | 0.6849 | – 0.24657 | 0.0286 | – 0.44381 | 0.0017 |
| Uninsur  | – 0.13185 | 0.0737 | – 0.09021 | 0.5689 | – 0.18672 | 0.2952 |
| Altitude | 0.01453 | 0.0395 | 0.01317 | 0.1245 | – 0.02271 | 0.0662 |
| Rural    | 0.09343 | 0.0158 | 0.07523 | 0.0483 | 0.03962 | 0.3671 |
| \(\rho\) | 0.4508 | < 0.0001 | 0.6557 | < 0.0001 | 0.8711 | < 0.0001 |
1. The neighbors are chosen using only the first step (the $K_1$ nearest neighbors), where $K_1 = 2$ for the districts with poverty level of less than 30%, and for the districts with poverty level between 30% and 55%, and $K_1 = 7$ for the districts with poverty level of more than 55%.

2. The neighbors are chosen using both steps, where in the second step we use the variable “Extreme Poverty” for the districts with poverty level of less than 30% ($K_1 = 3, K_2 = 2$), and for the districts with poverty level of more than 55% ($K_1 = 7, K_2 = 3$); for the districts with poverty level between 30% and 55%, the variable “Altitude” was utilized with $K_1 = 3$ and $K_2 = 2$.

Tables 11 and 12 display the estimators for the coefficients $\beta$ and $\rho$ obtained by fitting the spatial Fay–Herriot model under the first and the second scenarios. The results illustrate that the spatial correlations are substantially high, especially for the poorer districts, being higher under the second scenario as opposed to the first scenario. This suggests that ignoring the spatial correlation structure between the districts may increase the potential for greater MSE. It is important to emphasize that, in comparing the results reported in Tables 11 and 12 with those presented in Table 3, there is no drastic difference in the estimates for the parameters $\beta$.

This application shows that including the variable “Extreme Poverty” as a criterion for defining neighboring districts for the first and third groups and using the variable “Altitude” for the second group does not appear to have a significant impact on the coefficients $\beta$. However, further research is necessary to examine the potential effect of including various district-level covariates in definition of neighboring districts, on the model coefficients $\beta$.

In the following section we compare the EBLUP and the SEBLUP as well as their corresponding MSEs.
4.6 EBLUP, SEBLUP and MSE

First, we compare the predictions EBLUP and SEBLUP for the prevalence of anemia rates among children under five years, with the corresponding direct estimates. In the following figures, SEBLUP1 and SEBLUP2 refer to the SEBLUP predictors obtained under the first and the second scenarios defined above. For the purpose of these comparisons the following three groups of districts are used: the districts that only have 5 observations (a total of 19 districts), the districts with 15 observations (a total of 16 districts) and the districts with 40–49 observations (a total of 24 districts).

As expected, the results presented in Figs. 2, 3 and 4 illustrate that the differences between SEBLUP1, SEBLUP2, EBLUP and the corresponding direct estimate decreases as the sample size increases. Interestingly, the discrepancies between SEBLUP1 and SEBLUP2 are generally minor: the mean absolute differences...
between EBLUP1 and EBLUP2 is 0.014 in the first case, 0.010 in the second case and 0.009 in the third case. The corresponding relative differences amount to 3.6%, 2.4% and 3.1%, respectively.

Next, we present the MSEs of the discussed predictors. Figures 5, 6 and 7 display the MSEs obtained by application of the parametric bootstrap. The MSEs derived from application of the non-parametric bootstrap are somewhat larger, however, the conclusions reached are very similar to those reported below. The results indicate very clearly that in our case application of the spatial Fay–Herriot model yields better MSEs than the basic Fay–Herriot model. The results also provide evidence that except for several districts, the MSEs of SEBLUP2 have had better performance than SEBLUP1 and EBLUP, especially if the sample size is small. Specifically, the relative difference in MSE between SEBLUP1 and SEBLUP2 are 12.9%, 8.8%
and 6.0% in the first, second and third case, respectively. The direct estimates’ sampling errors are not displayed since they are much larger than the MSEs depicted in Figs. 5, 6 and 7. Specifically, in the first, second, and third cases, the relative difference in MSE between the direct estimate and EBLUP is 695 %, 190 %, and 66 %, respectively.

Finally, we compute the coefficients of variation (CV) for all predictors discussed above. As one can observe from Table 13, the direct estimator has very large CV in the districts where the sample size is smaller than 50. If the sample size is larger than 50, only for 61 districts (out of 104 districts) the CV of the direct estimator is smaller than 20%. Comparing this result to the corresponding numbers for EBLUP (87 districts), SEBLUP1 (92 districts) and SEBLUP2 (92 districts), it can be concluded that, in this study, for large samples employing the basic Fay–Herriot as well
as the spatial Fay–Herriot considerably improve the precision of the predictors, where the SEBLUP1 and SEBLUP2 slightly outperform the EBLUP. For smaller sample sizes a similar pattern can be observed, where the difference is that in these cases the performance of SEBLUP1 and SEBLUP2 is considerably better than that of the EBLUP, especially if the sample size is less than 7 or between 7 and 10. Moreover, the performance of SEBLUP2 is evidently better for all sample sizes. As an important byproduct of our study, we present a few maps of the estimates SEBLUP2 for districts belonging to different departments in Peru (see Appendix B). It should be noted that when calculating estimates for districts with null sample size, only the first part of the right-hand side of (15) was used.

### Table 13 Distribution of coefficient of variation for Direct estimate, EBLUP, SEBLUP1 and SEBLUP2 by sample size

| Sample size | Predictor | < 10% | 10–20% | 20–30% | > 30% | Total |
|-------------|-----------|-------|--------|--------|-------|-------|
| Less than 7 | Direct    | 0     | 1      | 4      | 120   | 125   |
| Less than 7 | EBLUP     | 0     | 57     | 56     | 12    | 125   |
| Less than 7 | SEBLUP1   | 0     | 78     | 40     | 7     | 125   |
| Less than 7 | SEBLUP2   | 0     | 82     | 40     | 3     | 125   |
| 7–10        | Direct    | 0     | 11     | 20     | 164   | 195   |
| 7–10        | EBLUP     | 0     | 76     | 89     | 30    | 195   |
| 7–10        | SEBLUP1   | 1     | 110    | 61     | 23    | 195   |
| 7–10        | SEBLUP2   | 1     | 120    | 57     | 17    | 195   |
| 11–20       | Direct    | 0     | 16     | 50     | 161   | 227   |
| 11–20       | EBLUP     | 1     | 106    | 90     | 30    | 227   |
| 11–20       | SEBLUP1   | 2     | 128    | 79     | 18    | 227   |
| 11–20       | SEBLUP2   | 3     | 145    | 63     | 16    | 227   |
| 21–50       | Direct    | 0     | 27     | 65     | 69    | 161   |
| 21–50       | EBLUP     | 0     | 84     | 65     | 12    | 161   |
| 21–50       | SEBLUP1   | 2     | 104    | 46     | 9     | 161   |
| 21–50       | SEBLUP2   | 6     | 113    | 34     | 8     | 161   |
| More than 50| Direct    | 10    | 51     | 39     | 4     | 104   |
| More than 50| EBLUP     | 10    | 77     | 15     | 1     | 104   |
| More than 50| SEBLUP1   | 11    | 81     | 11     | 1     | 104   |
| More than 50| SEBLUP2   | 13    | 79     | 12     | 0     | 104   |
| All Districts| Direct | 10    | 106    | 178    | 518   | 812   |
| All Districts| EBLUP    | 11    | 400    | 315    | 86    | 812   |
| All Districts| SEBLUP1  | 16    | 501    | 237    | 58    | 812   |
| All Districts| SEBLUP2  | 23    | 539    | 206    | 44    | 812   |

### 5 Conclusion

Considering the results obtained in Sect. 4 it can be concluded that utilizing the basic Fay–Herriot model results in considerably smaller MSEs (and therefore, the CVs) of the predictors as opposed to the direct estimates. However, the obtained
CVs in most of the districts are still substantially large. If the spatial Fay–Herriot model is applied, an additional reduction in MSEs is attained. This is due to incorporating information about the spatial structure of the data, which is ignored by the basic model. The reduction in MSE is more substantial if we select the neighbors using the two-step procedure which allows to employ additional information about the districts (see Sect. 4.2). Interestingly, the estimates for the coefficients \( \beta \) did not show drastic difference when fitting the basic and the spatial Fay–Herriot model (see Tables 3, 11 and 12). Regarding the question about reliability of the EBLUP and SEBLUP, the magnitudes of the corresponding CVs indicate that in the first case the percentage of unreliable estimates (the estimates with the CV larger than 20%) is considerably large, especially if the sample is small. Specifically, if the sample size is smaller than 7, the percentage of unreliable estimates is 54%. For larger sample sizes a very modest reduction can be achieved (48% if the sample size is between 21 and 51). If the sample size is larger than 50, the percentage of unreliable estimators reduces to 15%. In the case of the SEBLUP the corresponding percentages are as follows: 34% if the sample size is smaller than 7, 26% if the sample size is between 21 and 50 and 12% if the sample size is larger than 50. However, it should be noticed that the percentage of the estimates whose CV is larger than 30% is relatively small: in the case of the EBLUP it oscillates between 7 and 15% (for SEBLUP the range is between 2 and 9%) if the sample size is smaller than 50. If the sample size is larger than 50, the CV of only 1 predictor EBLUP (out of 104) is larger than 30%. In the case of the SEBLUP, the CVs of all predictors are smaller than 30%. Apart from comparing the performance of the basic and the spatial Fay–Herriot models we explore the sensitivity of choice of the neighbors to the resulting inference. It follows from the results that the conclusions drawn can significantly depend on the definition of the neighbors. We recommend that, in practice, one chooses the definition that achieves the smallest variance, \( \hat{\sigma}^2 \). There is no theoretical basis for this choice, however it may be advantageous from the perspective of reduction of the MSEs of the predictors. In this paper we do not discuss the problem of prediction in the nonsampled district. This could be a topic for future research.

**Appendix A**

For this study, we use R software (see R Core Team 2020). The initial step was creating a `SpatialPointsDataFrame` object that encompasses both the geographic information and survey data. To accomplish this, we employed the `SpatialPointsDataFrame` function, which is available in the `sp` package (see Bivand et al. 2013), as follows.

```r
proj=CRS("+proj=longlat +datum=WGS84 +no_defs")
data.sp=SpatialPointsDataFrame(coords=
  cbind(data$Latitude, data$Longitude),
data=data, proj4string=proj)
```

To establish neighboring districts based on their geographical coordinates, it was employed the `spdep` package (see Pebesma and Bivand 2023). More precisely, it was utilized the `knearneigh` function which returns a list of the districts along with their corresponding \( K \) nearest neighbors.

Springer
neighbors<-knnearneigh(coordinates(data.sp), k=K)

We use the list of the neighbors obtained in the output, in order to produce the matrix of spatial weights, $W$ by assigning a weight of $1/K$ to each neighbor. Finally, the $fh$ function from the emdi package (see Kreutzmann et al. 2019) was used to fit both the basic and spatial Fay Herriot models.

```
fh(fixed=formula, vardir="VR", combined_data=data,domains="District",method = "reml", correlation="spatial",corMatrix=W, MSE=TRUE, mse_type = "spatialparbootbc") -> spatial.fh
```

Here, the variables "VR" and "District" refer to the variance of the direct estimator and the identification district code, respectively. Note that omitting the third line would result in fitting the basic Fay Herriot model.

**Appendix B**

See Figs. 8 and 9.

![Fig. 8](image)

*Fig. 8* The prevalence of anemia rates in children under five, per district, in varios departments in Peru
Acknowledgments  The authors thank the reviewers for very thoughtful and helpful comments.

Funding  This research is supported by a grant from the Unidad de Investigación de la FIEECS-UNI.

Declarations

Conflict of interest  We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome. We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our
Estimating the prevalence of anemia rates among children under…

institutions concerning intellectual property
Authors have no conflicts of interest to disclose.

Consent to publication This manuscript has not been published anywhere and is not being considered for publication elsewhere.

References

Alcázar L (2012) Impacto económico de la anemia en el Perú. Grupo de Análisis para el Desarrollo (GRADE)  
Anselin L (1992) Spatial econometrics. Methods and models. Kluwer, Boston  
Banerjee S, Carlin B, Gelfand A (2004) Hierarchical modeling and analysis for spatial data. Chapman and Hall, New York  
Bivand RS, Pebesma E, Gomez-Rubio V (2013) Applied spatial data analysis with R, 2nd edn. Springer, New York  
Chen S, Lahiri P (2003) A comparison of different MSPE estimators of EBLUP for the Fay-Herriot model. In: Proceedings of the section on survey research methods. Washington, DC, American Statistical Association, pp 903–911  
Cressie N (1993) Statistics for spatial data. Wiley, New York  
Cressie N, Chan NH (1989) Spatial modeling of regional variables. J Am Stat Assoc 84:393–401  
Datta GS, Lahiri PS (2000) A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Stat Sinica 10:613–627  
Datta GS, Rao JNK, Smith DD (2005) On measuring the variability of small area estimators under a basic area level model. J Am Stat Assoc 87:724–731  
Fay RE, Herriot RA (1979) Estimates of income for small places: an application of James-Stein procedures to census data. J Am Stat Assoc 74:269–277  
Hall P, Maiti T (2006) On parametric bootstrap methods for small area prediction. J Roy Stat Soc B 68:221–238  
Harville D, Jeske D (1992) Mean squared error of estimation or prediction under a general linear model. J Am Stat Assoc 87:724–731  
INEI, Perú (2019) Encuesta Demográfica y de Salud Familiar-ENDES. https://proyectos.inei.gob.pe/microdatos/  
Jiang J, Lahiri PS, Wan SM (2002) A unified jackknife theory for empirical best prediction with M-estimation. Ann Stat 30:1782–1810  
Kackar RN, Harville DA (1984) Approximations for standard errors of estimators for fixed and random effects in mixed models. J Am Stat Assoc 79:853–862  
Kreutzmann A, Pannier S, Rojas-Perilla N, Schmid T, Templ M, Tzavidis N (2019) The R Package EMDI for estimating and mapping regionally disaggregated indicators. J Stat Softw 91(7):1–33. https://doi.org/10.18637/jss.v091.i07  
Marhuenda Y, Molina I, Morales D (2013) Small area estimation with spatio temporal Fay–Herriot models. Comput Stat Data Anal 58:308–325  
Ministerio de Salud (2014) Plan Nacional para la Reducción de la Desnutrición Crónica Infantil y la Prevención de la Anemia en el país: 2014-2016. RM No 258-2014 Lima, Minsa  
Ministerio de Salud (2017) Plan Nacional para la Reducción y Control de la Anemia Materno Infantil y la Desnutrición Crónica Infantil en el Perú: 2017-2021. RM No 249-2017 Lima, Minsa  
Molina I, Salvati N, Pratesi M (2009) Bootstrap for estimating the MSE of the spatial EBLUP. Comput Stat 24:441–458  
Moran PAP (1950) Notes on continuous stochastic phenomena. Biometrika 37(1):17–23  
Organización Mundial de la Salud (2011) Concentraciones de hemoglobina para diagnosticar la anemia y evaluar su gravedad. Ginebra, OMS. (WHO/NMH/NHD/ MNM/11.1)  
Pebesma E, Bivand R (2023) Spatial data science with applications. Chapman & Hall  
Petrucci A, Salvati N (2006) Small area estimation for spatial correlation in watershed erosion assessment. J Agric Biol Environ Stat 11(2):169–182  
Pfeffermann D (2002) Small area estimation- new developments and directions. Int Stat Rev 70:125–143
Pfeffermann D, Tiller RB (2005) Bootstrap approximation to prediction MSE for state-space models with estimated parameters. J Time Ser Anal 26:893–916
Prasad NGN, Rao JNK (1990) New important developments in small area estimation. J Am Stat Assoc 85(409):163–171
Pratesi M, Salvati N (2008) Small area estimation: the EBLUP estimator based on spatially correlated random area effects. Stat Method Appl 17(1):113–141
Pratesi M, Salvati N (2009) Small area estimation in the presence of correlated random area effects. J Off Stat 25(1):37–53
R Core Team (2020) R: language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Rao JNK (2003) Small area estimation. Wiley, London
Rao JNK, Molina I (2015) Small area estimation. Wiley series in survey methodology, 2nd edn. Wiley, Hoboken
Singh BB, Shukla K, Kundu D (2005) Spatial-temporal models in small area estimation. Surv Methodol 31(2):183–195
World Health Organization (2004) Centers for Disease Control and Prevention. Assessing the Iron Status of Populations, Ginebra, WHO

Publisher's Note. Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.