

Article

# Bayesian Bootstrap in Multiple Frames

Daniela Cocchi <sup>1,\*</sup> , Lorenzo Marchi <sup>2</sup>  and Riccardo Ievoli <sup>3</sup> <sup>1</sup> Department of Statistical Sciences, University of Bologna, 40126 Bologna, Italy; daniela.cocchi@unibo.it<sup>2</sup> KU Leuven, Research Centre Insurance, 3000 Leuven, Belgium; lorenzo.marchi@kuleuven.be<sup>3</sup> Department of Chemical, Pharmaceutical and Agricultural Sciences University of Ferrara, 44121 Ferrara, Italy; riccardo.ievoli@unife.it

\* Correspondence: daniela.cocchi@unibo.it

**Abstract:** Multiple frames are becoming increasingly relevant due to the spread of surveys conducted via registers. In this regard, estimators of population quantities have been proposed, including the multiplicity estimator. In all cases, variance estimation still remains a matter of debate. This paper explores the potential of Bayesian bootstrap techniques for computing such estimators. The suitability of the method, which is compared to the existing frequentist bootstrap, is shown by conducting a small-scale simulation study and a case study.

**Keywords:** resampling; complex surveys; Pólya's urn; variance estimation

## 1. Introduction

For several decades, Multiple Frame (MF) surveys, introduced by Hartley [1,2], and related estimators [3–6] have been receiving increasing attention. For these estimators, incorrect frame membership attribution is a potential drawback. The multiplicity estimator [7], further developed by Singh and Mecatti [8] and involving the number of frames to which units belong, bypasses the problem.

The topic is still relevant in statistical practice, as witnessed by the recent contributions of Lohr and Raghunathan [9], Wu and Thompson [10], and Lohr [11] and by targeted works regarding, for instance, calibration in dual frames [12], inference in the case of ordinal data [13], kernel-based methods [14], and empirical likelihood estimation in dual frames [15]. Moreover, variance estimation still represents one of the main challenges for scholars and practitioners [16,17]. In the special case of dual frames, useful proposals have been elaborated [18–20] based on linearization and jackknife methods. Another option in dual and multiple frames consists in applying bootstrap methods [21,22].

Originally introduced by Efron [23], bootstrap has been widely applied in survey sampling for variance estimation and data imputation; overviews of bootstrap methods in survey sampling can be found in Shao [24] and Lahiri [25]. Regarding complex surveys, two important contributions are those of Rao and Wu [26] and Sitter [27], where the performance of bootstrap methods in survey sampling is compared to linearization and jackknife. Although most of the methods for variance estimation in survey sampling are based on the frequentist proposal, Bayesian bootstrap (BB) methods, originally introduced by Rubin [28], have also been developed: see Lo [29], who introduced them in finite populations and discussed the case of stratified samples, Aitkin [30] for an application in complex surveys, and Carota [31] for a discussion about the choice of priors.

This paper aims at exploring the potential of BB techniques in estimating the variance in multiple frame surveys. In particular, we develop a new BB-based algorithm, which allows the estimation of the variance of the multiplicity estimator. The main advantage of the BB algorithms is that they allow the estimated variance to be obtained without any evaluation of second-order inclusion probabilities. A related contribution can be found in Lohr [21], who proposes two frequentist bootstrap algorithms (named separate and



**Citation:** Cocchi, D.; Marchi, L.; Ievoli, R. Bayesian Bootstrap in Multiple Frames. *Stats* **2022**, *5*, 561–571. <https://doi.org/10.3390/stats5020034>

Academic Editor: Fulvia Mecatti

Received: 10 May 2022

Accepted: 13 June 2022

Published: 15 June 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



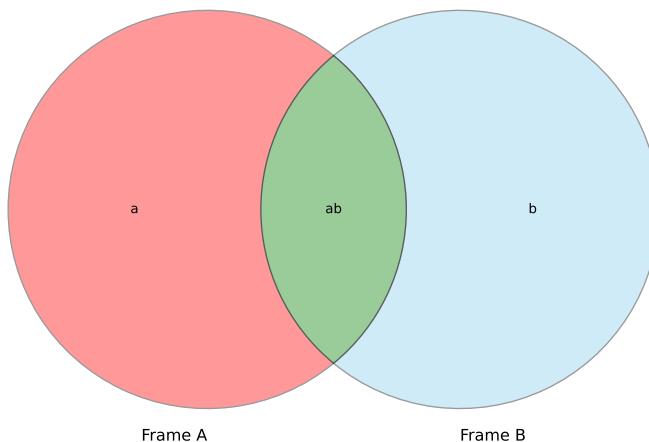
**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

combined, respectively) based on the Rao and Wu [26] rescaling technique. From a different perspective, Dong et al. [32] used BB with the aim of taking multiple complex surveys into account (without considering multiplicities), while Aidara [22] applied frequentist bootstrap (FB) in quasi-random sequences to estimate the variance of the multiplicity estimator [7].

The paper is organized as follows: Section 2 introduces the problem of variance estimation in multiple frames, while Section 3 illustrates the peculiarities of our non-parametric Bayesian proposal. A small-scale simulation study was performed in Section 4, while a case study appears in Section 5. Some concluding remarks are contained in the final section.

## 2. Multiple Frames and Variance Estimation

Multiple-frame sampling refers to surveys in which two or more frames are available and samples are drawn (usually independently) from each frame. This solution is preferred over the single sampling frame approach whenever a coverage improvement is needed, e.g., for dealing with elusive populations or for cost reduction purposes. The simplest case with two frames (A and B) is depicted in Figure 1.



**Figure 1.** Example of a dual-frame situation.

More generally, the situation can be expounded as follows: let  $U_1, \dots, U_q, \dots, U_{real-timeQ}$  be a collection of  $Q \geq 2$  frames. The sample data collected from a generic frame  $U_q$  can be classified into  $D_q$  disjoint domains  $U_{1(q)}, \dots, U_{d(q)}, \dots, U_{(D_q(q))}$ . The potential number of non-empty domains allowed for each frame is defined as  $D_q = 2^{(Q-1)}$ . Table 1 reports the case of  $Q = 3$  frames:  $\{A, B, C\}$  where four domains can be identified in each frame. For instance, for frame A, the domains  $D_A$  are as follows:  $\{a(A), ab(A), ac(A), abc(A)\}$ .

**Table 1.** Partition of multi-frame samples into frame-specific domains: example with  $Q = 3$ :  $\{A, B, C\}$ .

| Frame A | Frame B | Frame C |
|---------|---------|---------|
| a(A)    | b(B)    | c(C)    |
| ab(A)   | ab(B)   | ac(C)   |
| ac(A)   | bc(B)   | bc(C)   |
| abc(A)  | abc(B)  | abc(C)  |

The population is completely covered by two or more frames that may be overlapping, and this multiplicity should be taken into account when proposing an estimator. A relevant discussion can be found in Lohr and Rao [16]. A solution is proposed by Mecatti [7] and included in Singh and Mecatti [8] and Mecatti and Singh [33]: the multiplicity estimator for the total  $Y$  has the following form:

$$\hat{Y} = \sum_{q=1}^Q \sum_{k \in s_q} \frac{y_k}{m_k \pi_k}, \quad (1)$$

where  $s_q$  is a sample extracted from frame  $U_q$  under a given sampling design,  $y_k$  represents the individual target characteristic of interest,  $m_k$  is the multiplicity factor (i.e., the number of frames in which a given individual is sampled), and  $\pi_k$  is the first order inclusion probability for the  $k$ -th primary sampling unit (psu). Estimator (1) does not need to assign the  $k$ -th unit to any  $U_q$ . It belongs to the Horvitz–Thompson (HT) class, for which a closed-form for the variance can be computed, which depends on second-order inclusion probabilities. In fact, following Singh and Mecatti [8] and Mecatti and Singh [33], the variance of (1) is as follows:

$$V(\hat{Y}) = \sum_{q=1}^Q \left[ \sum_{k \in U_q} y_k^2 m_k^{-2} \frac{1 - \pi_{k(q)}}{\pi_{k(q)}} + \sum_{k \neq k' \in U_q} \frac{y_k m_k^{-1} y_{k'} m_{k'}^{-1}}{\pi_{k(q)} \pi_{k'(q)}} (\pi_{kk'(q)} - \pi_{k(q)} \pi_{k'(q)}) \right], \quad (2)$$

where, for each frame, all first-order ( $\pi_{k(q)}$ ) and second-order ( $\pi_{kk'(q)}$ ) inclusion probabilities must be specified. In the case of simple random sampling from each frame, the variance of (1) reduces to the following:

$$V(\hat{Y}) = \sum_{q=1}^Q \frac{N_q - n_q}{n_q(N_q - 1)} \left[ N_q \sum_{k \in U_q} y_k^2 m_k^{-2} - \left( \sum_{k \in U_q} y_k m_k^{-1} \right)^2 \right], \quad (3)$$

where  $N_q$  and  $n_q$  are the population and sample size of frame  $q$ , respectively.

## 2.1. Variance Estimation for the Multiplicity Estimator

The Sen–Yates–Grundy estimator of (2) is as follows. [33]:

$$\hat{V}(\hat{Y}) = \frac{1}{2} \sum_{q=1}^Q \sum_{k \neq k'} \sum_{k' \in s_q} \frac{\pi_{k'(q)} \pi_{k'(q)} - \pi_{kk'(q)}}{\pi_{kk'(q)}} \left( \frac{y_k m_k^{-1}}{\pi_{k(q)}} - \frac{y_{k'} m_{k'}^{-1}}{\pi_{k'(q)}} \right)^2. \quad (4)$$

In cases of simple random sampling of each frame, the estimator of (3) is [7].

$$\hat{V}(\hat{Y}) = \sum_{q=1}^Q \frac{N_q(N_q - n_q)}{n_q^2(N_q - 1)} \left[ N_q \sum_{k \in s_q} y_k^2 m_k^{-2} - \frac{N_q}{n_q} \left( \sum_{k \in s_q} y_k m_k^{-1} \right)^2 \right]. \quad (5)$$

Estimator (4) needs the first- and second-order inclusion probabilities of the sampled units. The last quantities are unknown and not trivial to estimate, especially in complex surveys.

When second-order inclusion probabilities are not available, and non-linear methods to estimate the variance are required, resampling techniques can be used to obtain an estimate of (2). The most used resampling methods for variance estimation in survey sampling [34] are the balanced repeated replications [35], the jackknife [36], and the bootstrap [25]. The jackknife has been introduced by Lohr and Rao [17,19] in dual frames, and it is further developed by Lohr and Rao [16] in multiple frames. In addition, jackknife has been recently proposed to estimate variance in the case of ordinal data in multiple frames [13].

## 2.2. Frequentist Bootstrap for Variance Estimation

The frequentist bootstrap (FB) is currently applied to variance estimation in (complex) survey sampling [26,27,37]. See Mashreghi et al. [38] for a general overview. In multiple frames, the technique is introduced by Lohr [21] using the rescaling bootstrap of Rao and Wu [26] and developed by [39] to obtain confidence intervals in the case of pseudo-

empirical likelihood-based estimator. Two possible procedures can be carried out in this regard. The former jointly resamples psus from all available frames. As per the latter, an algorithm is implemented to resample from each frame separately. In this case, a different number of iterations may be set in each frame. Regarding the multiplicity estimator (1), Aidara [22] applies the algorithm of [21] in a three-frame context using quasi Monte Carlo methods to improve bootstrap convergence.

In agreement with the mentioned literature, Algorithm 1 below summarizes the general procedure, sketched, in the two-frame case, by Lohr [21]. In a given frame  $q$ , for each bootstrap iteration  $(n_{h(q)} - 1)$ , psus are sampled from stratum  $h$  in frame  $q$ . Defining  $x_{h(q)k}(b)$  as how many times the psu  $k$  in the stratum  $h$  is drawn at the  $b$ -th bootstrap iteration, the sampling weight  $w_{h(q)k}$  is scaled according to the following scheme.

$$w_{h(q)k}(b) = w_{h(q)k} \frac{n_{h(q)}}{n_{h(q)} - 1} x_{h(q)k}(b). \quad (6)$$

---

**Algorithm 1** Frequentist bootstrap
 

---

```

for each frame  $q$  do
  for each bootstrap iteration  $b$  do
    for each stratum  $h(q)$  do
      (a) generate a synthetic sample  $s_{h(q)}^*$  of size  $n_{h(q)} - 1$  using SRSWR
      (b) adjust unit-specific sampling weights using Equation (6)
    end for
    estimate population total using the  $q$ -th row of Equation (7)
  end for
  estimate bootstrap variance of the frame using (8)
end for
aggregate frame-specific variances (9)
  
```

---

Similarly to the jackknife technique, the variance estimator can be expressed as a function of weights in (6) as follows. Indeed, for a given  $\hat{\tau}$  estimator of interest, for each iteration  $b$ , the following  $Q$ -elements vector is constructed:

$$\begin{aligned}
 \hat{\tau}_{(1)}^*(b) &= g(w_{(1)}(b), w_{(2)}, \dots, w_{(Q)}) \\
 &\dots \\
 \hat{\tau}_{(q)}^*(b) &= g(w_{(1)}, \dots, w_{(q)}(b), \dots, w_{(Q)}) \\
 &\dots \\
 \hat{\tau}_{(Q)}^*(b) &= g(w_{(1)}, w_{(2)}, \dots, w_{(Q)}(b))
 \end{aligned} \quad (7)$$

where  $g(\cdot)$  is a duly specified function.

Assuming that  $g(\cdot)$  has the functional form (1), for each frame, a bootstrap-based variance estimator can be computed as follows:

$$\hat{V}_{(q)}^*(\hat{Y}) = \frac{1}{B_q} \sum_{b=1}^{B_q} (\hat{\tau}_{(q)}^*(b) - \hat{Y})^2. \quad (8)$$

where  $B_q$  is the total number of bootstrap iterations, which can be different for each frame. Finally, the variance estimator of (1) is obtained by aggregating separate estimators (8):

$$\hat{V}^*(\hat{Y}) = \sum_{q=1}^Q \hat{V}_{(q)}^*(\hat{Y}) \quad (9)$$

### 3. Bayesian Bootstrap in Multiple Frames

In what follows, we propose a non-parametric Bayesian approach for variance estimation in multi-frame sampling designs. Bayesian bootstrap (BB) constitutes an additional method to approach variance estimation based on resampling and is a little explored opportunity in the case of multiple frame surveys. While the classical bootstrap resamples from the observed sampled values (in the “naive” case), BB starts from the posterior distribution of the sampled units [28]. It was introduced by Lo [29] for survey sampling in finite populations and, more recently, discussed by Aitkin [30] and Carota [31]. In case of multiple surveys, contributions can be found in Dong et al. [32,40].

Broadly speaking, BB can be defined in terms of the Dirichlet–Multinomial compound model. Let  $\{y\}_1^N$  be the values of a characteristic attributed to an exchangeable population, where  $J \leq N$  includes the values that can be sampled according to a vector of probabilities  $\theta = (\theta_1, \theta_2, \dots, \theta_J)$ . In the simplest case, the prior distributions for the parameters  $\theta$  are assumed to be uniformly distributed, i.e., flat priors [41]. We assume that  $\theta \sim F(\alpha)$ , where  $F(\alpha)$  is the Dirichlet prior distribution over the parameter of the population generating process, as follows.

$$\begin{cases} y_i | \theta \sim \text{Multinomial}(\theta_1, \theta_2, \dots, \theta_J) \\ \theta | \alpha \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_J) \end{cases} \quad (10)$$

Thus, according to the Bayes theorem, the posterior predictive distribution can be derived [32] as follows:

$$\begin{aligned} p(Y|y) &= \int \frac{p(Y, \theta, y)}{p(y)} d\theta \\ &= \frac{\int_0^1 \dots \int_0^1 p(Y|y, \theta) p(y|\theta) p(\theta) d\theta_1 \dots d\theta_J}{\int_0^1 \dots \int_0^1 p(y|\theta) p(\theta) d\theta_1 \dots d\theta_J} \\ &= \frac{\int_0^1 \dots \int_0^1 p(Y|y, \theta) p(y|\theta) p(\theta) d\theta_1 \dots d\theta_J}{\int_0^1 \dots \int_0^1 p(y|\theta) p(\theta) d\theta_1 \dots d\theta_J} \\ &= \frac{\int_0^1 \dots \int_0^1 \prod_{j=1}^J \theta_j^{N_j - n_j} \prod_{j=1}^J \theta_j^{n_j} \prod_{j=1}^J \theta_j^{\alpha_j - 1} \theta_1 \dots d\theta_J}{\int_0^1 \dots \int_0^1 \prod_{j=1}^J \theta_j^{n_j} \prod_{j=1}^J \theta_j^{\alpha_j - 1} d\theta_1 \dots d\theta_J} \\ &= \frac{\prod_{j=1}^J \Gamma(N_j + \alpha_j) / \Gamma(\alpha_j)}{\Gamma(N + \alpha_0) / \Gamma(\alpha_0)} \left( \frac{\prod_{j=1}^J \Gamma(n_j + \alpha_j)}{\Gamma(n + \alpha_0)} \right)^{-1}, \end{aligned} \quad (11)$$

where  $N = \sum_{j=1}^J N_j$  and  $n = \sum_{j=1}^J n_j$  represent the total number of elements in the population and in the sample, respectively. In addition,  $\alpha_0 = \sum_{j=1}^J \alpha_j$  and  $\Gamma(\cdot)$  stands for the Gamma function.

Resampling from the posterior predictive (12) is not trivial at all. Consequently, the suggestion for practical implementation is to leverage Pólya’s urn scheme to simulate such a distribution [29]. In a nutshell, the Pólya’s urn scheme contains  $J$  values  $\alpha_j$  with  $j = 1, \dots, J$ . Initially, a value  $j$  is randomly sampled from the urn and reinserted together with another value of the same type. After a sufficiently large number of iterations, the distribution converges to  $F(\alpha) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_J)$  [42].

#### The Proposed Algorithm

As mentioned above, BB relies on generating a Dirichlet-based posterior predictive distribution via Pólya’s urn scheme. For a single frame and non-complex sampling, the application of the BB is straightforward, but the presence of multiple frames and multi-strata sampling designs [29] induces a further degree of complexity that must be considered.

Our proposal in this regard is summarized in Algorithm 2. The generation of a synthetic population represents its core, starting from generating synthetic samples. A number of  $(N - n)$  elements is generated using Pólya's urn scheme. In the generic synthetic sample, we can identify  $(N_j - n_j)$  as the number of draws of units belonging to the same group. This corresponds to the draw from the Dirichlet-based posterior predictive distribution  $p(Y|y)$  of Equation (12).

---

**Algorithm 2** Bayesian bootstrap

---

```

for each frame  $q$  do
    for each bootstrap iteration  $b(q)$  do
        for each stratum  $h(q)$  do
            (a) generate a synthetic sample  $s_{h(q)}^*$  of size  $(N_{h(q)} - n_{h(q)})$  using the
                Pólya Urn model on the original sample  $s_{h(q)}$ 
            (b) construct  $C_{h(q)}$  by concatenating the original sample  $s_{h(q)}$  with  $s_{h(q)}^*$  (12)
            (c)  $n_{h(q)}$ -sized sampled is drawn from  $C_{h(q)}$ 
            (d) adjust unit-specific sampling weights using Equation (13)
        end for
        estimate population total using the  $q$ -th row of Equation (14)
    end for
    estimate bootstrap variance of the frame using Equation (15)
end for
aggregate frame-specific variances (16)

```

---

Therefore, in any frame  $q$ , for the variable of interest  $y$  in any stratum  $h(q)$ , the final bootstrapped population is obtained by concatenating the  $n_{h(q)}$  units of the original sample and the  $N_{h(q)} - n_{h(q)}$  bootstrapped units. Then, the BB-based population values are obtained for each  $h(q)$  as follows:

$$C_{h(q)} = \{y_1, \dots, y_{n_{h(q)}}\} \cup \{y_1^*, \dots, y_{N_{h(q)} - n_{h(q)}}^*\} \quad (12)$$

where  $y^*$ 's represent values sampled from the Dirichlet-based posterior predictive distribution. The population in (12) is then used to resample  $n_{h(q)}$  units, which constitute the frame and stratum-specific bootstrap sample. Distinct from the FB-based weights in (6), in BB, the weights are obtained as follows.

$$w_{h(q)k}^{BB}(b) = w_{h(q)k} x_{h(q)k}(b). \quad (13)$$

Then, for a given  $\hat{\tau}$  estimator, for each iteration  $b$ , the following  $Q$ -elements vector is constructed using weights: (13):

$$\begin{aligned} \hat{\tau}_{(1)}^{*BB}(b) &= g(w_{(1)}^{BB}(b), w_{(2)}, \dots, w_{(Q)}) \\ &\dots \\ \hat{\tau}_{(q)}^{*BB}(b) &= g(w_{(1)}, \dots, w_{(q)}^{BB}(b), \dots, w_{(Q)}) \\ &\dots \\ \hat{\tau}_{(Q)}^{*BB}(b) &= g(w_{(1)}, w_{(2)}, \dots, w_{(Q)}^{BB}(b)) \end{aligned} \quad (14)$$

where  $g(\cdot)$  is the function of the previous paragraph in the frequentist case.

Assuming that  $g(\cdot)$  has the functional form (1), for each frame, a BB-based variance estimator can be computed as follows.

$$\hat{V}_{(q)}^{*BB}(\hat{Y}) = \frac{1}{B_q} \sum_{b=1}^{B_q} (\hat{\tau}_{(q)}^{*BB}(b) - \hat{Y})^2. \quad (15)$$

Similarly to the FB case, the variance estimator can be obtained as follows.

$$\hat{V}^{*BB}(\hat{Y}) = \sum_{q=1}^Q \hat{V}_{(q)}^{*BB}(\hat{Y}) \quad (16)$$

#### 4. Simulation Study

In this section, a small-scale simulation study was performed to assess the proposed methodology via a comparison between BB and FB.

##### 4.1. Set-Up

As a first Data Generating Process (DGP), we consider a three-frame design with simple random sampling in each frame ( $DGP_1$ ). Following Mecatti, we generate  $M$  Monte Carlo pseudo-populations of  $N = 2400$  elements from a Gamma distribution with parameters  $(1.5, 2)$  such that the population total is  $Y = 7200$ . Each element of the population is then randomly assigned to a frame via Bernoulli trials. We state two alternative expected values for each frame,  $p = 0.4$  and  $p = 0.6$ , and ensure overlapping between frames and non-empty frames. Three possible sampling fractions ( $f_q = n_q/N_q$ ) are considered: 0.05, 0.15, and 0.40.

Secondly, we consider a more complex sample design with stratification ( $DGP_2$ ). Two strata are generated according to  $N_1 = N_2 = 1200$ ; the individual values of the characteristic under study are sampled from the following.

- A Gamma distribution with parameters  $(1.5, 2)$ ;
- A Gamma distribution with parameters  $(2, 4)$ .

The population total is now equal to  $Y = 13,200$ , while the same Bernoulli trials (to construct frames) and sampling fractions of  $DGP_1$  are used.

The number of Monte Carlo simulations is set equal to  $M = 500$ , and the number of bootstrap replications is  $B = 399$ . In agreement with previous studies [7,16,22], two performance indicators were used: Relative Bias (RB) and the Coefficient of Variation (CV). The RB is computed as follows:

$$RB = \frac{1}{M} \sum_{m=1}^M \frac{(\hat{V}_m^* - MSE)}{MSE} \cdot 100 \quad (17)$$

where the  $\hat{V}_m^*$  is either the BB- or the FB-based variance estimate for the  $m$ -th sample, according to the theoretical choice performed. The expression for the CV is as follows.

$$CV = \frac{\sqrt{\frac{1}{M} \sum_{m=1}^M (\hat{V}_m^* - MSE)^2}}{MSE} \quad (18)$$

The reference MSE in (17) and (18) is computed via 10,000 Monte Carlo simulations as follows:

$$MSE = \frac{1}{10,000} \sum_{m=1}^{10,000} (\hat{Y}_m - Y)^2$$

where  $\hat{Y}_m$  is the estimate computed for the  $m$ -th synthetic sample.

##### 4.2. Main Results

For  $DGP_1$ , Tables 2 and 3 illustrate the results about performance indicators, i.e., the RB (17) and the CV (18), with  $p = 0.4$  and  $p = 0.6$ , respectively.

**Table 2.** Performance indicators for DGP<sub>1</sub> considering  $p = 0.4$ .

| $f_q$ | RB      |         | CV    |       |
|-------|---------|---------|-------|-------|
|       | BB      | FB      | BB    | FB    |
| 0.05  | −8.328  | −6.198  | 0.288 | 0.279 |
| 0.15  | −7.575  | −6.459  | 0.173 | 0.168 |
| 0.40  | −13.828 | −11.546 | 0.166 | 0.150 |

**Table 3.** Performance indicators for DGP<sub>1</sub> considering  $p = 0.6$ .

| $f_q$ | RB     |        | CV    |       |
|-------|--------|--------|-------|-------|
|       | BB     | FB     | BB    | FB    |
| 0.05  | −1.216 | −3.541 | 0.316 | 0.312 |
| 0.15  | −0.217 | −0.095 | 0.196 | 0.181 |
| 0.40  | −7.141 | −5.935 | 0.137 | 0.130 |

In terms of RB, both tables show how the BB and the FB have a similar counterintuitive behaviour, achieving the poorest results for the highest sampling fraction (0.4). Furthermore, in the case of  $p = 0.6$  (Table 3), a lower value of the indicator is associated with the BB at the lowest sampling fraction (0.05), while it performs poorly when the sampling fraction is higher. With an intermediate value of sampling fraction (0.15), still considering  $p = 0.6$ , the performances in terms of RB are very satisfactory for both methods: the values are very close to zero.

As per the CV, the two methods show comparable results in both Tables 2 and 3, even if FB slightly outperforms BB, with a decreasing behaviour of the indicator as the sampling fraction increases, either for  $p = 0.4$  or  $p = 0.6$ .

The results for the more complex DGP<sub>2</sub> witness the reliability of the BB with respect to the FB, as shown in Tables 4 and 5. In particular, FB severely overestimates the variability for each  $p$  and  $f_q$ , as denoted by the high positive values of its RB. Conversely, BB, which slightly underestimates the variability of the estimator, performs well, especially with  $p = 0.6$  (Table 5) and when the sampling fraction is not too small ( $f_q = 0.15$  and  $f_q = 0.40$ ). In terms of CV, even if both methods show a decreasing trend according to the increase in sampling fraction, BB definitively outperforms FB.

**Table 4.** Performance indicators for DGP<sub>2</sub> with  $p = 0.4$ .

| $f_q$ | RB      |        | CV    |       |
|-------|---------|--------|-------|-------|
|       | BB      | FB     | BB    | FB    |
| 0.05  | −21.996 | 62.258 | 0.375 | 1.088 |
| 0.15  | −12.827 | 56.689 | 0.236 | 0.755 |
| 0.40  | −12.849 | 49.886 | 0.174 | 0.605 |

**Table 5.** Performance indicators for DGP<sub>2</sub> with  $p = 0.6$ .

| $f_q$ | RB      |        | CV    |       |
|-------|---------|--------|-------|-------|
|       | BB      | FB     | BB    | FB    |
| 0.05  | −14.944 | 74.108 | 0.432 | 1.211 |
| 0.15  | −2.479  | 68.636 | 0.266 | 0.848 |
| 0.40  | −2.010  | 58.881 | 0.159 | 0.688 |

## 5. Case Study

To stress the advantages of using the BB in multiple frame surveys, the proposed algorithm is applied to the two-frame dataset included as a running example in the R package Frames2 [43]. Two households populations are considered with  $N_A = 1735$  and

$N_B = 1191$  and with an intersection of  $N_{ab} = 601$  such that  $N = 2325$ . The first population is organized in  $H = 6$  strata with the following sizes:  $N_{h(A)} = \{727, 375, 113, 186, 115, 219\}$ . Two samples are selected without replacement in the following manner:

- $n_A = 105$  by simple random sampling in each stratum:

$$n_{h(A)} = \{15, 20, 15, 20, 15, 20\}$$

;

- $n_B = 135$  by simple random sampling.

The available variables are three types of expenditures: Feeding, Clothing, and Leisure (in Euros). The number of bootstrap replications is set equal to  $B = 999$ . Table 6 summarizes the main results where the estimated variances are divided by  $10^6$ .

**Table 6.** Resampling-based variance estimates (divided by  $10^6$ ) for the case study

| Variable | FB     | BB     | (FB – BB)/FB |
|----------|--------|--------|--------------|
| Feeding  | 348.24 | 280.12 | 19.56%       |
| Clothing | 6.33   | 5.40   | 14.74%       |
| Leisure  | 2.50   | 1.88   | 24.52%       |

Empirical results confirms that the BB-based variance estimates are lower than the competing ones computed with FB under stratified sampling in multiple frames. In particular, BB exhibits a relative percentage difference (with respect to FB) that ranges (approximately) between 15% and 25%.

## 6. Discussion and Conclusions

The novelty of the present paper is the proposal of a Bayesian non-parametric technique (BB) to estimate the variance of a multiple frame estimator for a parameter of interest, namely the population total. The BB is proposed to construct the first-order inclusion probabilities in a non-frequentist manner and without modifying design-based properties. BB is also compared with frequentist bootstrap (FB), suggested by [21] and applied by [22]. The motivation for using resampling methods in multiple frames is due to the fact that they do not require the estimation of second-order inclusion probabilities.

Results of a small-scale simulation study show that the BB and FB perform similarly under simple random sampling in each frame, with a slight advantage in favor of the FB except when the sampling fraction is very low. However, this result should be considered as a benchmark, since under simple random sampling in each frame, a closed-form for the variance estimator is currently available [8,33].

Under a more complex sampling design like stratification, FB becomes practically unusable, severely overestimating the variability of the estimator in the context of multiple frames. Few previous experiments of FB in multiple frames have been performed [21, 22], but they are not directly comparable with our findings due to different DGPs as a starting point for the simulation studies. Possible issues related to the application of FB to stratified samples in multiple frames should be investigated in further studies even from a theoretical perspective. Conversely, BB exhibits satisfactory performance, especially when the sampling fraction is not very low. A case study also reveals the suitability of BB in the context of dual frames with stratification.

The results here presented need future investigation with more intensive Monte Carlo simulations, alternative Data Generating Processes, different population parameters, further estimators in the context of multiple frames in addition to (1), and other complex sampling design, e.g., cluster sampling. In addition, BB is insensitive to the choice of prior for totals or means; its sensitivity for non-linear estimators should be deepened using tools similar to those provided by Aitkin [30] and Carota [31]. Finally, a relevant advantage for

scholars and practitioners would be the implementation of BB (and the FB) in (possibly open source) statistical software.

**Author Contributions:** Conceptualization, D.C., L.M., and R.I.; methodology, D.C., L.M., and R.I.; software, L.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Data Availability Statement:** The codes to generate simulated Monte Carlo populations for the simulations can be obtained on request. The data of the case study are taken from Frames2 package in R software [43].

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Hartley, H.O. Multiple frame surveys. In Proceedings of the social statistics section, American Statistical Association, Washington, DC, USA, 7–10 September 1962; Volume 19, pp. 203–206.
2. Hartley, H.O. Multiple frame methodology and selected applications. *Sankhya* **1974**, *36*, 118.
3. Fuller, W.A.; Burmeister, L.F. Estimators for samples selected from two overlapping frames. In *Proceedings of the Social Statistics Section*; American Statistical Association: Boston, MA, USA, 1972; Volume 245249.
4. Bankier, M.D. Estimators based on several stratified samples with applications to multiple frame surveys. *J. Am. Stat. Assoc.* **1986**, *81*, 1074–1079.
5. Kalton, G.; Anderson, D.W. Sampling rare populations. *J. R. Stat. Soc. Ser. A (General)* **1986**, *149*, 65–82.
6. Skinner, C.J. On the efficiency of raking ratio estimation for multiple frame surveys. *J. Am. Stat. Assoc.* **1991**, *86*, 779–784.
7. Mecatti, F. A single frame multiplicity estimator for multiple frame surveys. *Surv. Methodol.* **2007**, *33*, 151–157.
8. Singh, A.C.; Mecatti, F. Generalized multiplicity-adjusted Horvitz-Thompson estimation as a unified approach to multiple frame surveys. *J. Off. Stat.* **2011**, *27*, 633.
9. Lohr, S.L.; Raghunathan, T.E. Combining survey data with other data sources. *Stat. Sci.* **2017**, *32*, 293–312.
10. Wu, C.; Thompson, M.E. Dual Frame and Multiple Frame Surveys. In *Sampling Theory and Practice*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 305–317.
11. Lohr, S. Multiple-frame surveys for a multiple-data-source world. *Surv. Methodol.* **2021**, *47*, 229–264.
12. Ranalli, M.G.; Arcos, A.; del Mar Rueda, M.; Teodoro, A. Calibration estimation in dual-frame surveys. *Stat. Methods Appl.* **2016**, *25*, 321–349.
13. Rueda, M.d.M.; Arcos, A.; Molina, D.; Ranalli, M.G. Estimation techniques for ordinal data in multiple frame surveys with complex sampling designs. *Int. Stat. Rev.* **2018**, *86*, 51–67.
14. Sánchez-Borrego, I.; Arcos, A.; Rueda, M. Kernel-based methods for combining information of several frame surveys. *Metrika* **2019**, *82*, 71–86.
15. del Mar Rueda, M.; Ranalli, M.G.; Arcos, A.; Molina, D. Population empirical likelihood estimation in dual frame surveys. *Stat. Pap.* **2021**, *62*, 2473–2490.
16. Lohr, S.; Rao, J.K. Estimation in multiple-frame surveys. *J. Am. Stat. Assoc.* **2006**, *101*, 1019–1030.
17. Lohr, S.L. Multiple-frame surveys. In *Handbook of statistics*; Elsevier: Amsterdam, The Netherlands, 2009; Volume 29, pp. 71–88.
18. Skinner, C.J.; Rao, J.N. Estimation in dual frame surveys with complex designs. *J. Am. Stat. Assoc.* **1996**, *91*, 349–356.
19. Lohr, S.L.; Rao, J. Inference from dual frame surveys. *J. Am. Stat. Assoc.* **2000**, *95*, 271–280.
20. Demnati, A.; Rao, J.N.; Hidiroglou, M.A.; Tambay, J.L. On the allocation and estimation for dual frame survey data. In *Proceedings of the Survey Research Methods Section*; American Statistical Association: Boston, MA, USA, 2007; pp. 2938–2945.
21. Lohr, S. Recent developments in multiple frame surveys. *Cell* **2007**, *46*, 6.
22. Aidara, C.A.T. Quasi Random Resampling Designs for Multiple Frame Surveys. *Statistica* **2019**, *79*, 321–338.
23. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26.
24. Shao, J. Impact of the bootstrap on sample surveys. *Stat. Sci.* **2003**, *18*, 191–198.
25. Lahiri, P. On the impact of bootstrap in survey sampling and small-area estimation. *Stat. Sci.* **2003**, *18*, 199–210.
26. Rao, J.N.; Wu, C. Resampling inference with complex survey data. *J. Am. Stat. Assoc.* **1988**, *83*, 231–241.
27. Sitter, R.R. A resampling procedure for complex survey data. *J. Am. Stat. Assoc.* **1992**, *87*, 755–765.
28. Rubin, D.B. The bayesian bootstrap. *Ann. Stat.* **1981**, *9*, 130–134.
29. Lo, A.Y. A Bayesian bootstrap for a finite population. *Ann. Stat.* **1988**, *16*, 1684–1695.
30. Aitkin, M. Applications of the Bayesian bootstrap in finite population inference. *J. Off. Stat.* **2008**, *24*, 21.
31. Carota, C. Beyond objective priors for the Bayesian bootstrap analysis of survey data. *J. Off. Stat.* **2009**, *25*, 405.
32. Dong, Q.; Elliott, M.R.; Raghunathan, T.E. Combining information from multiple complex surveys. *Surv. Methodol.* **2014**, *40*, 347.

33. Mecatti, F.; Singh, A.C. Estimation in multiple frame surveys: a simplified and unified review using the multiplicity approach. *J. Soc. Fr. Stat.* **2014**, *155*, 51–69.
34. Cocchi, D.; Ievoli, R. Resampling Procedures for Sample Surveys. In *Wiley StatsRef: Statistics Reference Online*; Wiley: Hoboken, NJ, USA, 2020; pp. 1–8.
35. McCarthy, P.J. Pseudo-replication: Half samples. *Rev. Inst. Int. Stat.* **1969**, *37*, 239–264.
36. Miller, R.G. The jackknife—a review. *Biometrika* **1974**, *61*, 1–15.
37. Sitter, R.R. Comparing three bootstrap methods for survey data. *Can. J. Stat.* **1992**, *20*, 135–154.
38. Mashreghi, Z.; Haziza, D.; Léger, C. A survey of bootstrap methods in finite population sampling. *Stat. Surv.* **2016**, *10*, 1–52.
39. Rao, J.; Wu, C. Pseudo-empirical likelihood inference for multiple frame surveys. *J. Am. Stat. Assoc.* **2010**, *105*, 1494–1503.
40. Dong, Q.; Elliott, M.R.; Raghunathan, T.E. A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Surv. Methodol.* **2014**, *40*, 29.
41. Lo, A.Y. Bayesian statistical inference for sampling a finite population. *Ann. Stat.* **1986**, *14*, 1226–1233.
42. Frigyik, B.A.; Kapila, A.; Gupta, M.R. *Introduction to the Dirichlet Distribution and Related Processes*; Technical Report, UWEETR-2010-0006; Department of Electrical Engineering, University of Washington: Washington, DC, USA, 2010.
43. Arcos, A.; Molina, D.; Ranalli, M.G.; del Mar Rueda, M. Frames2: A Package for Estimation in Dual Frame Surveys. *R J.* **2015**, *7*, 52–72. <https://doi.org/10.32614/RJ-2015-005>.