mathematics

MDPI

Article

Single Imputation Methods and Confidence Intervals for the Gini Index

Encarnación Álvarez-Verdejo, Pablo José Moya-Fernández and Juan Francisco Muñoz-Rosas *

Department of Quantitative Methods in Economics and Business, University of Granada, 18011 Granada, Spain; encarniav@ugr.es (E.Á.-V.); pjmoyafernandez@ugr.es (P.J.M.-F.)
* Correspondence: jfmunoz@ugr.es

Abstract: The problem of missing data is a common feature in any study, and a single imputation method is often applied to deal with this problem. The first contribution of this paper is to analyse the empirical performance of some traditional single imputation methods when they are applied to the estimation of the Gini index, a popular measure of inequality used in many studies. Various methods for constructing confidence intervals for the Gini index are also empirically evaluated. We consider several empirical measures to analyse the performance of estimators and confidence intervals, allowing us to quantify the magnitude of the non-response bias problem. We find extremely large biases under certain non-response mechanisms, and this problem gets noticeably worse as the proportion of missing data increases. For a large correlation coefficient between the target and auxiliary variables, the regression imputation method may notably mitigate this bias problem, yielding appropriate mean square errors. We also find that confidence intervals have poor coverage rates when the probability of data being missing is not uniform, and that the regression imputation method substantially improves the handling of this problem as the correlation coefficient increases.

Keywords: missing data; variance estimation; coverage; inequality; non-response mechanism

Citation: Álvarez-Verdejo, E.; Moya-Fernández, P.J.; Muñoz-Rosas, J.F Single Imputation Methods and Confidence Intervals for the Gini Index. *Mathematics* **2021**, *9*, 3252. https://doi.org/10.3390/math9243252

Academic Editors: Antonio Di Crescenzo and Bahram Adrangi

Received: 24 September 2021
Accepted: 14 December 2021
Published: 15 December 2021

1. Introduction

Most surveys suffer from the problem of missing data, and this issue may have an important impact on results and conclusions. Missing data may appear for many reasons, and cases of both unit and item non-response can be observed. Unit non-response indicates that data for certain units are missing, i.e., there is no information at all for such units. On the other hand, item non-response arises when only some variables of the study have missing values. Note that it is quite common for individuals to choose not to answer sensitive questions, such as those related to income, wealth, drugs use, etc. This distinction between unit and item non-response is important when it comes to handling the problem of missing data. Thus, weighting adjustment procedures ([1]) are commonly used in the presence of unit non-response, whereas imputation methods ([2]) are usually considered for item non-response.

The consequences of missing data may be serious. Non-response bias is possibly the most critical issue; it is the bias of a given estimate that appears when respondents and non-respondents have, in general, different values for the target variables. A second common consequence is the fact that the variance of estimators may increase, which implies, for example, that the precision of the study decreases, and confidence intervals will be wider. Variance estimates may also have a bias, and this issue has an impact on confidence intervals, hypothesis testing, etc. Finally, missing data may mean obtaining smaller sample sizes, with valuable information potentially being removed from the study.

Rubin [3] proposed the classification *MCAR*, *MAR* and *MNAR* for surveys with missing data. According to Rubin's theory, non-response is viewed as a random process where each unit has a certain probability of being missing. This process is termed a non-response mechanism, and is unknown in real applications. *MCAR* (Missing Completely

At Random) applies when the probability of being missing is constant for all units, and does not depend on either the observed or missing data. The non-response bias is not a problem when the $MCAR$ assumption holds, since the missing data can be considered as a random sample taken from the original sample. The use of imputation classes is a common practice when dealing with missing data (see [4]). Imputation classes are homogeneous groups of respondents created with the aim of minimizing the bias. Note that the $MCAR$ assumption is often unrealistic, although it is quite common to assume $MCAR$ inside imputation classes. A MAR (Missing At Random) mechanism arises when the probability of missing data depends only on the observed data. The MAR assumption is more common than the $MCAR$ assumption, and the non-response bias is usually small under a MAR mechanism. Finally, the $MNAR$ (Missing Not At Random) mechanism applies otherwise, i.e., when neither the $MCAR$ nor the MAR assumption holds. For the $MNAR$ assumption, the probability of missing data depends on both observed and missing data, and the non-response bias is a serious issue under this non-response mechanism. Additional information on non-response mechanisms can be found in [5].

Many statistical and machine-learning techniques can be used in the presence of missing data. The simplest solution is to do nothing, i.e., remove the units containing missing values from the study and analyse only units without any missing data. This method is commonly referred to as Complete Case Analysis (CCA) or Listwise Deletion, and it has some desirable properties, such as the fact that it provides unbiased estimators under an $MCAR$ mechanism. However, this assumption is not very common, as has been previously discussed. Note that CCA suffers from some serious disadvantages. For instance, it is obvious that the sample size may decrease considerably, and the main effect is that the efficiency of estimators decreases under the various non-response mechanisms, including $MCAR$. In addition, valuable information is removed, and there is a high probability of non-response bias.

Weighting adjustment procedures can also be used when dealing with missing data. The idea of weighting is similar to the survey sampling theory ([6]), i.e., parameter estimates are obtained using a set of weights that are calculated to compensate for non-response. While it may reduce the non-response bias, there is no simple application of this method to item non-response. In addition, weighting adjustment may produce unstable estimators when large weights are obtained.

Finally, the use of a single imputation method is a common solution to the problem of missing data. Single imputation consists of replacing each missing value with a plausible value in order to obtain accurate parameter estimates. In general, imputation is used for item non-response and has the advantage of ensuring all observed data are used. Imputation also suffers from some disadvantages, such as the fact that this method may modify the relationship between variables. The variance of estimators may also be underestimated, but this problem can be solved by using multiple imputation (see [7,8]). Multiple imputation consists of replacing each missing value with M plausible values, where $M > 1$, and M different datasets are thus obtained. Each completed dataset is then analysed using the statistical analysis of interest, and the M results are combined using Rubin's rules (see [5]).

Measuring inequality lies within the scope of numerous fields. For instance, refs. [9–11] analyse inequality in health, environmental and educational studies, respectively. However, this topic has a special relevance in economic studies, where income inequality has been extensively investigated (see, for example, refs. [12,13]). The most popular statistic used to measure inequality is the Gini index. This indicator was originally proposed by [14], since when it has attracted a great deal of attention. For instance, additional formulations of the Gini index have been suggested by [15,16], among others. The use of a bias correction technique for the Gini index is discussed by [17,18]. Variance estimation of the Gini index has been investigated, for example, by [19,20]. An exhaustive review of the problem of estimating the variance in the Gini index estimation can be seen in [21]. Some confidence intervals for the Gini index have been proposed by [22–24]. An excellent review of the Gini index can be found in [25]. It is important to note that the Gini index and related mea-

sures have also been adopted in other contexts, such as for the construction of topological indices for trees and graphs (see [26,27]), for the analysis of reliability systems ([28]) and for constructing decision-making methods ([29]). A key advantage of the Gini index is its ease of interpretation, as it takes values between 0 and 1, where 0 indicates perfect equality and 1 the opposite. In addition, this simplicity facilitates cross-country comparisons, since the Gini index does not depend on the size of the population. Furthermore, obtaining Gini index estimates is very straightforward as, they are regularly reported by countries and international organizations such as the Word Bank and Eurostat. A limitation of the Gini index is that it is less sensitive to changes at the top and the bottom of the income distribution than it is to changes in the middle of the income distribution (see [30]). Some recent references that use the Gini index to measure income inequality are [31–33]. The quintile share ratio (see [34,35]) is another indicator commonly used to measure inequality. For instance, income inequality in European Union member states is described using the Gini Index and the quintile share ratio.

The main limitation of the quintile share ratio is the fact that it ignores inequality in the middle of the income distribution, but it provides a good measure of the income inequality between the top and the bottom of the income distribution. Note that other decile ratios for measuring income inequality can also be found in the literature, but the quintile share ratio is usually preferred over other decile ratios because it is less sensitive to extreme values.

The first contribution of this paper is to analyse the empirical performance (in terms of bias and efficiency) of some traditional single imputation methods when they are applied to the estimation of the Gini index. Note that the empirical bias and the empirical efficiency are measured, respectively, in terms of Relative Bias (RB) and Relative Root Mean Square Error ($RRMSE$). The RB is of special relevance, since this measure tells us the magnitude of the non-response bias. First, we empirically quantify the biases of the usual estimator of the Gini index for the various non-response mechanisms, which allows us to easily compare the impact of the non-response mechanism on the bias of this estimator of the Gini index. Similarly, we investigate the loss of empirical efficiency of the customary estimator of the Gini index for the various non-response mechanisms, and the results can be compared with the $RRMSE$ value of this estimator of the Gini index based on the original sample without missing data. Second, we analyse the evolution of both the RB and $RRMSE$ values when the proportion of missing data increases, which allows us to identify the situations where the loss of efficiency is non-negligible in comparison to results from the original sample without missing data. It is worth noting that most surveys contain auxiliary variables related to the variable of interest, and they can be used at the estimation stage to improve the estimation of a given statistic. Single imputation methods can also be based on auxiliary variables, and this approach may improve the estimation of the Gini index. Third, we also analyse the evolution of both the RB and $RRMSE$ values when the linear correlation coefficient between the target and auxiliary variables increases. Finally, the bias and the efficiency of the various procedures are investigated for small and large Gini coefficients.

The second contribution is to analyse the empirical performance, in terms of empirical coverage rate (CR) and empirical width (W), of the aforementioned basic single imputation methods when they are applied to the construction of confidence intervals for the Gini index. In this case, the same scenarios are investigated, i.e., we evaluate confidence intervals for different non-response mechanisms, proportions of missing data, correlation coefficients and Gini coefficients. For both these contributions, results are obtained using Monte Carlo simulation studies with a total of 48 different scenarios.

We use single imputation methods for various reasons. For instance, single imputation is more frequently used than multiple imputation in National Statistical Institutes, besides the fact that single imputation is simpler and less computationally intensive than multiple imputation. In addition, as discussed in [36], a small value of M may provide a poor estimation of the between imputation variance, and it may have an important effect on the precision of the variance estimator obtained from the multiple imputation. Finally, there is

no simple application of multiple imputation to some issues related to survey sampling, such as clustering, stratification or weighting to compensate for the selection of units with unequal probabilities. Obviously, multiple imputation has various advantages over single imputation. For example, multiple imputation takes into account the uncertainty in the imputation process and may considerably improve the estimation of the variance of estimators. For this reason, as discussed in Section 5, the analysis of the empirical performance of multiple imputation methods when they are applied to the estimation of the Gini index, and the comparison with results derived from single imputation methods, are suggested as avenues for further research in the near future.

Ref. [37] analyses the impact of missing data on the estimation of a measure of inequality that is similar to the Gini index and is commonly used to study health variables. Ref. [37] conducts a simulation study to compare CCA and a multiple imputation procedure. Only four scenarios were investigated, all of which involve the MAR non-response mechanism. In addition, this study analyses the bias, but not the efficiency or the impact on confidence intervals. Assuming a single case study based on a Health and Nutrition Survey, ref. [38] compares estimates of the Gini index based on the CCA approach and a multiple imputation method. Similarly, results from [39,40] are based on a case study. As discussed in [37], results from case studies may be less suitable to generalise the findings than Monte Carlo simulation studies based on a large number of replications.

The purpose of Section 2 is to provide researchers with a comprehensive view of two relevant topics: the Gini index and some basic single imputation methods to deal with the problem of missing data. First, the formal definition of the Gini index in continuous distributions is described in Section 2. Then, we present the most common estimators of the Gini index in discrete distributions. The variance estimation and the construction of confidence intervals are also discussed. Finally, some common single imputation methods are introduced in Section 2. The main contribution of this paper is to empirically compare, in Section 3, the various single imputation methods when they are applied to the estimation of the Gini index, and to analyse their effect on the accuracy of confidence intervals. The conclusions are detailed in Section 4, and a brief discussion is presented in Section 5.

2. Methods

2.1. The Gini Index

Let Y be a non-negative continuous random variable that represents the incomes of a given population. The distribution function of Y is denoted as $F_Y(y) = P(Y \leq y)$, and $f(y)$ is the corresponding probability density function. Finally, Y^+ also denotes a random variable with the same distribution $F_Y(y)$, and it is assumed that Y^+ and Y are independent. The Gini index can be defined as (see [15]):

$$G = \frac{1}{2\mu_Y} \int_0^{+\infty} \int_0^{+\infty} |y^+ - y| dF_Y(y^+) dF_Y(y), \tag{1}$$

where

$$\mu_Y = E[Y] = \int_0^{+\infty} y f(y) dy = \int_0^{+\infty} y dF_Y(y)$$

is the mean of income. Additional formulations of the Gini index can be found in [16,22,41].

Equation (1) is valid for continuous distributions. However, in practice, it is quite common to analyse income inequality in the context of a sample survey, i.e., samples are derived from a finite population, which is denoted as U, and it is assumed that it has size N (see [6]). Let $\{Y_1, \ldots, Y_N\}$ be N copies of Y, and $\{y_1, \ldots, y_N\}$ a realisation of these copies, i.e., they represent the observed incomes of individuals included in the finite population. For discrete distributions, G is generally replaced by a specific approach. For instance, the classical approach of the Gini index based on population values is given by (see [42]):

$$G_N^B = \frac{1}{2N^2\overline{Y}} \sum_{i \in U} \sum_{j \in U} |y_i - y_j|, \tag{2}$$

where the population of income is defined as $\overline{Y} = N^{-1}\sum_{i\in U} y_i$. Note that Equation (2) is the plug-in expression of Equation (1). As with the case of continuous distributions, many formulations of the Gini index have been suggested for discrete distributions (see, among others [16,43]). An exhaustive review of formulations of the Gini index for both discrete and continuous distributions can be seen in [25]. An interesting discussion in the literature concerns the use of the bias correction approach

$$G_N = \frac{N}{N-1}G_N^B = \frac{1}{2N(N-1)\overline{Y}}\sum_{i\in U}\sum_{j\in U}|y_i - y_j|,$$

For instance, [17,18] explain that the bias corrected approach may minimize the bias of G_N^B. In addition, the bias of G_N^B may have an impact on the coverage of confidence intervals for the Gini index. For these reasons, G_N is used throughout this paper.

In survey sampling, the population values are unknown, which implies that a random sample S, with size n, must be selected from U under a given sampling design. The idea of this paper is to empirically compare various common statistical procedures, some which are designed for samples derived under simple random sampling without replacement ($SRSWOR$); hence, this is the sampling design considered in this paper. A discussion on the extension to a general sampling design can be seen in Section 5. The usual estimator of G_N is defined as

$$\widehat{G} = \frac{n}{n-1}\widehat{G}^B = \frac{1}{2n(n-1)\overline{y}}\sum_{i\in S}\sum_{j\in S}|y_i - y_j|,$$

where $\overline{y} = n^1\sum_{i\in S} y_i$ is the sample mean and

$$\widehat{G}^B = \frac{1}{2n^2\overline{y}}\sum_{i\in S}\sum_{j\in S}|y_i - y_j|$$

is the estimator of G_N^B.

The variance estimator or standard error of a given statistic plays an important role at the estimation stage, since such measures give an idea of the accuracy of the point estimate, as well as allowing the construction of confidence intervals. The variance estimation of the Gini index has been extensively investigated, with an excellent review on this topic provided by [21], who also analyses and compares various variance estimators in the literature. Results from [21] indicate that both jackknife and linearization approaches have desirable properties in comparison to alternatives. Accordingly, we use these methods for variance estimation and the construction of confidence intervals for the Gini index. An extensive description of the linearization approach can be found in [20,44]. Relevant references that describe the jackknife method for the Gini index are [45,46]. Some alternative methods for variance estimation and/or construction of confidence intervals for the Gini index that can be found in the literature are the bootstrap ([47,48]) and empirical likelihood ([22,23]).

The variance estimator for the Gini index based on the linearization approach is defined as (see [19,21]):

$$\widehat{V}_L(\widehat{G}) = \widehat{V}_L\left(\frac{n}{n-1}\widehat{G}^B\right) = \frac{n^2}{(n-1)^2}\widehat{V}_L(\widehat{G}^B), \tag{3}$$

where

$$\widehat{V}_L(\widehat{G}^B) = N^2\frac{1-f}{n(n-1)}\sum_{i\in S}(l_i - \bar{l})^2,$$

$f = n/N$ is the sampling fraction, $\bar{l} = n^{-1}\sum_{i\in S} l_i$ and

$$l_i = \frac{1}{N\overline{y}}\left[2y_i\widehat{F}(y_i) - \left(\widehat{G}^B + 1\right)(y_i + \overline{y}) + \frac{2}{n}\sum_{j\in S} y_j\delta(y_i \le y_j)\right]$$

are the pseudo-values derived from the linearization approach (see [19]). Finally,

$$\widehat{F}(y) = \frac{1}{n}\sum_{i\in S}\delta(y_i \leq y)$$

is the empirical distribution function based on the sample S, and $\delta(\cdot)$ is the indicator variable that takes the value 1 if its argument is true and 0 otherwise.

The variance estimator for the Gini index based on Ogwang's jackknife is defined as (see [21,43]):

$$\widehat{V}_O(\widehat{G}) = \widehat{V}_O\left(\frac{n}{n-1}\widehat{G}^B\right) = \frac{n^2}{(n-1)^2}\widehat{V}_O(\widehat{G}^B),$$

where

$$\widehat{V}_O(\widehat{G}^B) = \frac{n-1}{n}\sum_{i\in S}\left(\widehat{G}^B(i) - \overline{G}^B\right)^2,$$

$\overline{G}^B = n^{-1}\sum_{i\in S}\widehat{G}^B(i)$, the jackknife estimates are given by

$$\widehat{G}^B(i) = \widehat{G}^B + \frac{2}{n\overline{y} - y_{(i)}}\left\{\frac{y_{(i)}\widehat{\beta}}{n} + \frac{\sum_{j=1}^{n} j y_{(j)}}{n(n-1)} - \frac{n\overline{y} - \sum_{j=1}^{i} y_{(j)} - i y_{(i)}}{n-1}\right\} - \frac{1}{n(n-1)},$$

$y_{(i)}$ are the values y_i sorted in increasing order, and

$$\widehat{\beta} = \frac{\sum_{i\in S} i y_{(i)}}{\sum_{i\in S} y_{(i)}}.$$

Different methods can be applied to construct confidence intervals for the Gini index. For instance, normal approximation confidence intervals for the Gini index have been examined by [18,22], among others. Assuming that the asymptotic normality assumption holds, the $(1-\alpha)$-level normal approximation confidence interval based on the linearization variance estimator is given by

$$\left(\widehat{G} - z_{1-\alpha/2}\sqrt{\widehat{V}_L(\widehat{G})},\quad \widehat{G} + z_{1-\alpha/2}\sqrt{\widehat{V}_L(\widehat{G})}\right),$$

where z_a denotes the ath quantile of the standard normal distribution. Similarly, the corresponding confidence interval based on Ogwang's jackknife is given by

$$\left(\widehat{G} - z_{1-\alpha/2}\sqrt{\widehat{V}_O(\widehat{G})},\quad \widehat{G} + z_{1-\alpha/2}\sqrt{\widehat{V}_O(\widehat{G})}\right). \tag{4}$$

As noted by [22], confidence intervals based on the asymptotic normality assumption may have issues with undercoverage probabilities when samples are small. Alternatively, bootstrap procedures may be used for the construction of confidence intervals, some of which may depend on a given variance estimator of the Gini index. For this purpose, we consider Ogwang's jackknife variance estimator because the results from Section 3 indicate that the jackknife approach provides confidence intervals with better empirical coverage rates than the linearization approach. However, bootstrap confidence intervals based on the linearization variance estimator can be similarly defined. Let $\{y_1^*(b), \ldots, y_n^*(b)\}$ be the bth bootstrap sample selected from the artificial bootstrap population U^* by *SRSWOR*, and $b = \{1, \ldots, B\}$, where B is the total number of bootstrap samples. Let $\widehat{G}^*(b)$ and

$\widehat{V}_O(\widehat{G}^*(b))$ be, respectively, the estimates $\widehat{G}$ and $\widehat{V}_O(\widehat{G})$ based on the bth bootstrap sample. A bootstrap-t confidence interval is defined as (see [22]):

$$\left(\widehat{G} + t^*_{\alpha/2}\sqrt{\widehat{V}_O(\widehat{G})}, \quad \widehat{G} + t^*_{1-\alpha/2}\sqrt{\widehat{V}_O(\widehat{G})}\right), \tag{5}$$

where t^*_a denotes the ath quantile of the values

$$t^*(b) = \frac{\widehat{G}^*(b) - \widehat{G}}{\sqrt{\widehat{V}_O(\widehat{G}^*(b))}}.$$

Finally, the confidence interval based on the percentile bootstrap is defined as

$$\left(\widehat{G}^*_{\alpha/2}, \quad \widehat{G}^*_{1-\alpha/2}\right),$$

where $\widehat{G}^*_a$ is the ath quantile of the bootstrapped values $\widehat{G}^*(b)$.

2.2. Some Single Imputation Methods

We now describe the single imputation methods considered in this paper. As discussed in Section 1, the use of auxiliary variables may considerably improve the performance of imputation methods. For simplicity, we consider a single auxiliary variable X associated with the variable of interest Y. In addition, we assume that missing values only appear in the sample values of the variable Y, i.e., all the sample values of the auxiliary variable X are observed. Note that this scheme is usually required by imputation methods based on auxiliary variables (see [2,49]). Therefore, we consider that r of the n sample values of the variable Y are observed (respondents), and this subset is denoted as $S_r = \{i \in S \mid y_i \textit{ is observed}\}$. The $m = n - r$ remaining values are considered as missing data (non-respondents), i.e., we may define the subset $S_m = \{i \in S \mid y_i \textit{ is missing}\}$. The proportion of missing values in the variable of interest is thus defined as $p = m/n$.

The popular Random Hot Deck (RHD) imputation method (see [50]) consists of replacing each of the m missing values with a random value selected from the r available values of the variable Y, i.e., the missing value y_i, with $i \in S_m$, is substituted by y^*_i, which is randomly selected from S_r. Although this imputation method is widely used, it has some limitations. For example, RHD can be easily used when the sample S is selected under $SRSWOR$, but a modification is required to accommodate this method to a general sampling design with unequal inclusion probabilities. In addition, it should be noted that this stochastic imputation method may perform better if imputation classes or adjustment cells are created.

The regression method (see [51]) is an imputation method based on auxiliary variables. For a single auxiliary variable and assuming the usual regression model

$$y_i = a + bx_i + u_i,$$

where u_i are independent and identically distributed random variables with zero mean, this method consists of replacing the missing value y_i, with $i \in S_m$, by

$$y^*_i = \widehat{y}_i + \epsilon_i,$$

where

$$\widehat{y}_i = \bar{y}_r + \hat{b}(x_i - \bar{x}_r)$$

is the predicted value obtained from the regression model, $\bar{x}_r = r^{-1}\sum_{i\in S_r} x_i$ and $\bar{y}_r = r^{-1}\sum_{i\in S_r} y_i$ are the sample means of X and Y, respectively, and based on the sample S_r, and

$$\hat{b} = \frac{\sum_{i\in S_r}(x_i - \bar{x}_r)(y_i - \bar{y}_r)}{\sum_{i\in S_r}(x_i - \bar{x}_r)^2}.$$

Predicted values can be used to replace the missing data, but this imputation method may underestimate the true variance of the variable of interest. For this reason, random disturbances are usually added to the predicted values to increase variability. For instance, ϵ_i can be randomly selected from the residuals of the regression model and associated with the respondents, i.e., ϵ_i is a random residual taken from the set of residuals $e_j = y_j - \widehat{y}_j$, with $j \in S_r$. Alternatively, the random disturbances can be generated from a parametric distribution, such as the normal distribution.

Finally, the Nearest Neighbour Imputation (NNI) method (see [52]) is a popular imputation method that has also been used in many applications. The NNI method consists of replacing each missing value with the value of the nearest observation for one or more auxiliary variables. For a single auxiliary variable, the NNI method substitutes the missing value y_i, with $i \in S_m$, by y_{min}, where x_{min} is the value of the auxiliary variable that minimizes the absolute distance

$$\delta_{i,j} = |x_i - x_j|,$$

with $j \in S_r$. For the case of categorical or dichotomous variables, this distance between neighbours is calculated as

$$\delta_{i,j} = \begin{cases} 0 & if \quad x_i = x_j, \\ 1 & if \quad x_i \neq x_j. \end{cases}$$

A review of candidate distances that may be used by the NNI method can be seen in [53]. Note that various solutions can be obtained in this minimizing problem. If this is the case, y_{min} is randomly selected from among the various values of the auxiliary variable that minimize the absolute distance.

3. Monte Carlo Simulation Studies

In this section, we empirically analyse the impact of the single imputation methods described in Section 2.2 on the estimator $\widehat{G}$ and the confidence intervals for the Gini index defined in Section 2.1. For this purpose, we carried out a set of Monte Carlo simulation studies based on different scenarios, which are described in Section 3.1. Results can be seen in Section 3.2.

3.1. Description of the Study

Monte Carlo simulation studies are based on $R = 1000$ replications. The methods described in Section 2 assume that survey samples (with size n) are selected from a finite population (with size N). The population size in this study is fixed at $N = 1000$. The N values of the variable Y are selected from the Lognormal distribution, which is quite common in the modelling of income distributions. Cases of both low and high income inequalities are considered: for this purpose we use the Gini coefficients $G = \{0.2, 0.6\}$, which are obtained when the standard deviation of the Lognormal distribution takes, respectively, the values $\sigma = \{0.36, 1.19\}$. In addition, we consider the mean $\mu = 5$ for this distribution. The auxiliary variable is generated using the expression $X = Y + \epsilon$, where ϵ is a random variable with a normal distribution. The standard deviation of ϵ is selected such that the correlation coefficient between Y and X takes the values $\rho = \{0.5, 0.95\}$, meaning cases of both weak and strong correlations are analysed. Applying this method, an additional auxiliary variable Z is also generated, and where the correlation coefficient between Y and Z is 0.7. Z is only used for the selection of missing units under the $MNAR$ mechanism, i.e., Z is not used for estimation purposes. For each replication, the sample S with size $n = 100$ is selected from the aforementioned finite population under $SRSWOR$, yielding the sample observations $\{y_1, \ldots, y_n\}$ and $\{x_1, \ldots, x_n\}$. Then, missing units in the variable of interest are randomly selected using the $MCAR$, MAR and $MNAR$ mechanisms, by means of the function *ampute* (package *mice*) of the statistical software *R*. Different proportions of missing data are considered; specifically, $p = \{0.1, 0.2, 0.3, 0.4\}$. In summary, we analyse 4 values of p, 2 values of both G and ρ and the 3 non-response

mechanisms, which means that a total of 48 different scenarios are investigated. Confidence intervals based on the bootstrap method are constructed using $B = 1000$ bootstrap samples. Estimators and confidence intervals for the Gini index are calculated using: (1) all units in the sample S ($All - S$); (2) Complete Case Analysis (CCA); (3) imputation and the Random Hot Deck method (RHD); (4) imputation and the Regression imputation method (Reg); and (5) imputation and the Nearest Neighbour Imputation method (NNI).

The various statistical methods are compared in terms of different empirical measures. The Relative Bias

$$RB = \frac{E[\widehat{G}] - G}{G}$$

and the Relative Root Mean Square Error

$$RRMSE = \frac{\left(MSE[\widehat{G}]\right)^{1/2}}{G}$$

are used to compare the performance of the various estimates of the true Gini index G, where the empirical expectation is defined as

$$E[\widehat{G}] = \frac{1}{R}\sum_{r=1}^{R}\widehat{G}(r),$$

the empirical mean square error is defined as

$$MSE[\widehat{G}] = \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{G}(r) - G\right)^2,$$

and $\widehat{G}(r)$ denotes the estimator $\widehat{G}$ when it is calculated at the rth replication. On the other hand, confidence intervals are compared in terms of empirical Coverage Rate

$$CR = \frac{1}{R}\sum_{r=1}^{R}\delta(L(r) \leq G \leq U(r))$$

and empirical Width

$$W = \frac{1}{R}\sum_{r=1}^{R}(U(r) - L(r)),$$

where $L(r)$ and $U(r)$ denote, respectively, the lower and upper limits of a given confidence interval obtained at the rth replication. The confidence level is fixed at 95%.

3.2. Results

Results from the various Monte Carlo simulation studies can be seen in Figures 1–8. Values of RB and $RRMSE$ can be seen in Figures 1 and 2 when the true Gini index is given by $G = \{0.2, 0.6\}$, respectively.

For a low Gini index (Figure 1), we observe that CCA and Reg yield satisfactory values for RB under an $MCAR$ mechanism and for the various values of p. However, for MAR and $MNAR$ mechanisms, values of RB for the CCA approach decrease as the values of p increase. RHD and NNI methods produce serious biases for the various non-response mechanisms. However, the empirical performance of NNI improves as the value of ρ increases. As expected, Reg and NNI perform better than RHD and CCA in terms of RB when ρ is large, and the Reg method provides smaller values of RB, in absolute terms, than the NNI method. The various methods show poor empirical performance when ρ is small and under an $MNAR$ mechanism. In summary and as expected, the non-response bias is not a serious issue under an $MCAR$ mechanism, although RHD and NNI methods may provide slightly biased estimates. For the MAR mechanism, the various imputation

methods give RB values close to -2% when the proportion of missing data is $p = 0.1$, and empirical biases increase, in absolute terms, as p increases, so the non-response bias may be non-negligible for large proportions of missing data. Finally, biases based on the $MNAR$ mechanism are larger, in absolute terms, that those obtained from an MAR mechanism, so the non-response bias may be a serious issue in this situation.

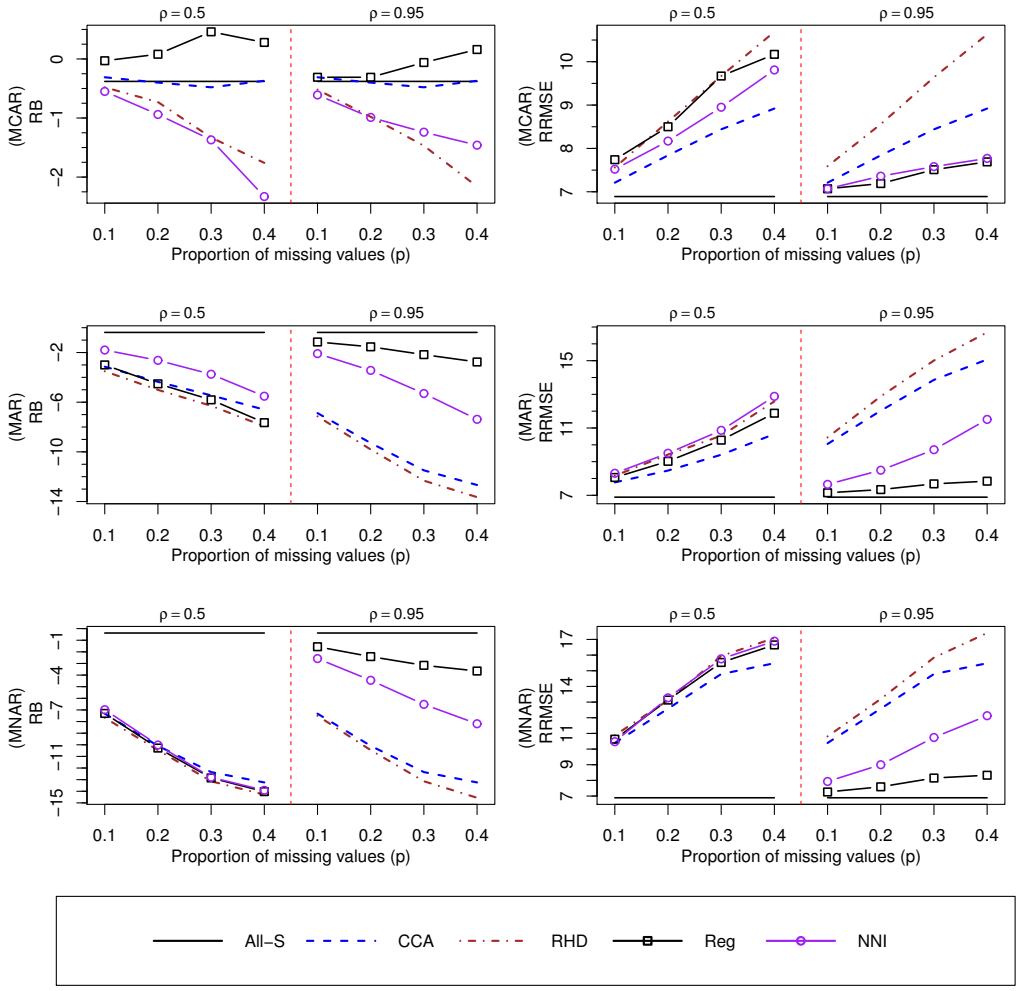

Figure 1. Values of RB and $RRMSE$ for $\widehat{G}$ when estimating $G = 0.2$.

We may reach some different conclusions, in terms of bias, when G is large (see Figure 2). For instance, the biases of CCA and Reg seem to be affected by p when the $MCAR$ assumption holds, since the values of RB decrease substantially when $p = 0.4$. In addition, the Reg method shows the worst empirical performance in comparison to alternative approaches when $\rho = 0.5$ and under a MAR mechanism. Finally, note that the RB values when $G = 0.2$ are slightly smaller, in absolute terms, than those recorded when $G = 0.6$. In summary, our results indicate that the non-response bias problem may get worse as the Gini index increases.

As far as the empirical efficiency is concerned, for a low Gini index (see Figure 1), we observe that the various approaches give similar values of $RRMSE$ when $\rho = 0.5$, although the CCA approach is slightly better than alternative methods. However, Reg and NNI provide more efficient results than RHD and CCA when ρ is large, and the Reg method is better than NNI, especially under MAR and $MNAR$ mechanisms. Similar conclusions, in terms of $RRMSE$, are reached when $G = 0.6$ (see Figure 2).

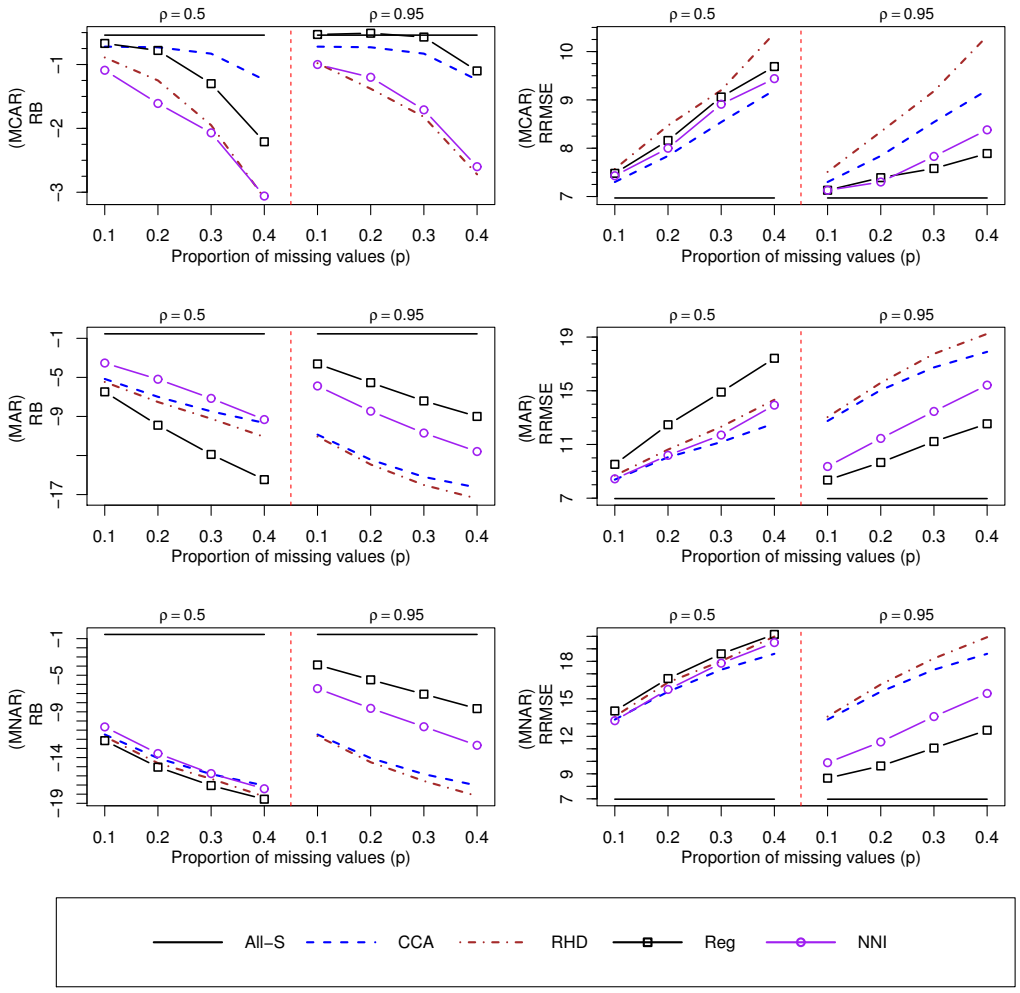

Figure 2. Values of RB and $RRMSE$ for $\widehat{G}$ when estimating $G = 0.6$.

Confidence intervals for $G = 0.2$ are empirically investigated in Figures 3–5, which consider the $MCAR$, MAR and $MNAR$ mechanisms, respectively. First, we observe that the jackknife variance estimator performs slightly better (in terms of CR) than the linearization variance estimator (see, in Figure 3, confidence intervals based on the normal approximation), and for this reason, the various confidence intervals in this study are based on the jackknife variance estimator.

For the $MCAR$ mechanism (Figure 3), CCA provides satisfactory empirical coverage rates, but the confidence intervals widen considerably as the proportion of missing data increases. Alternative methods perform poorly in terms of CR when $\rho = 0.5$, although the Reg and NNI imputation methods also give reasonable coverage rates when $\rho = 0.95$, and satisfactory values of W for the various values of p. The various methods for the construction of confidence intervals (normal approximation, studentized bootstrap and percentile bootstrap) give similar results. However, confidence intervals based on the studentized bootstrap are slightly wider than confidence intervals based on alternative methodologies (normal approximation and percentile bootstrap).

For the MAR mechanism (Figure 4), CCA also provides unsatisfactory coverage rates as p increases. When ρ is large, the best results, in terms of CR, are obtained using the Reg imputation method, while the RHD imputation method shows the worst performance. As expected, a strong correlation provides better coverage rates with imputation methods based on auxiliary variables (Reg and NNI). Note that the bias observed for the $MNAR$ mechanism has an impact on the coverage rates of confidence intervals. In particular,

values of CR under the $MNAR$ mechanism (Figure 5) are smaller than the corresponding coverage rates under the MAR mechanism (Figure 4).

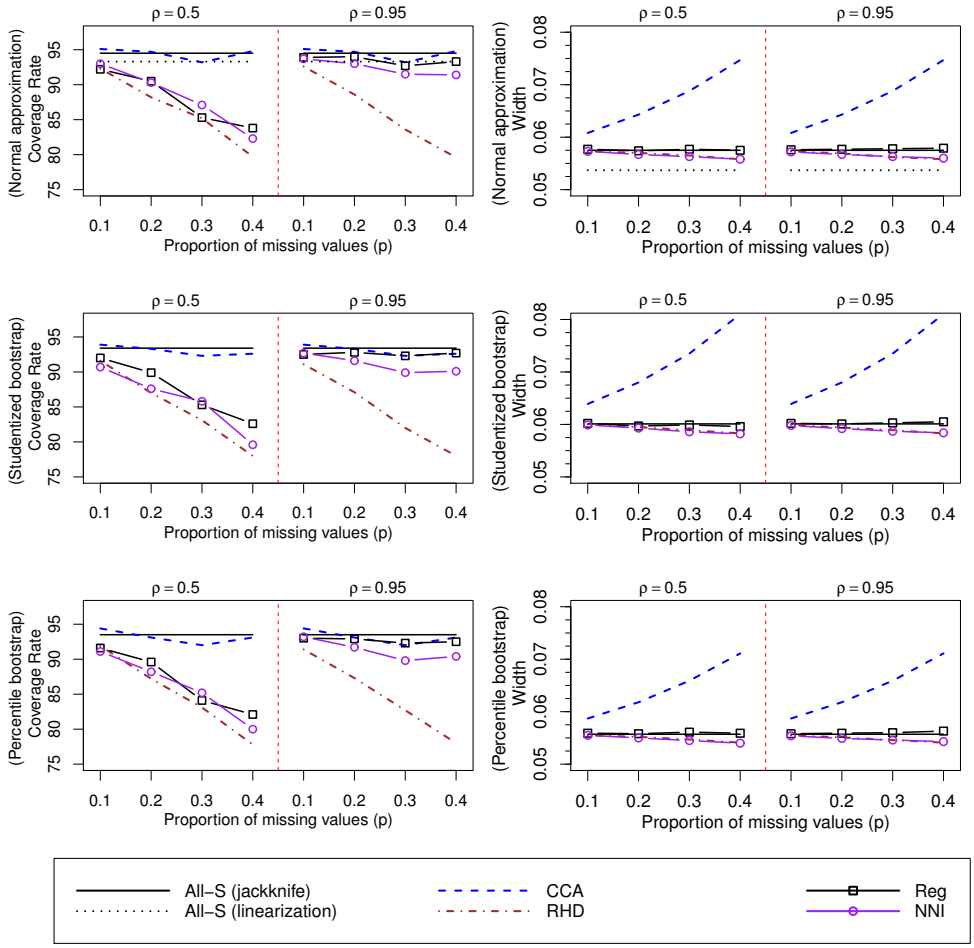

Figure 3. Values of CR and W associated with 95% confidence intervals for $G = 0.2$, and based on the jackknife variance estimator. The $MCAR$ mechanism is considered. Linearization and jackknife variances are compared using the normal approximation and the $All - S$ approach.

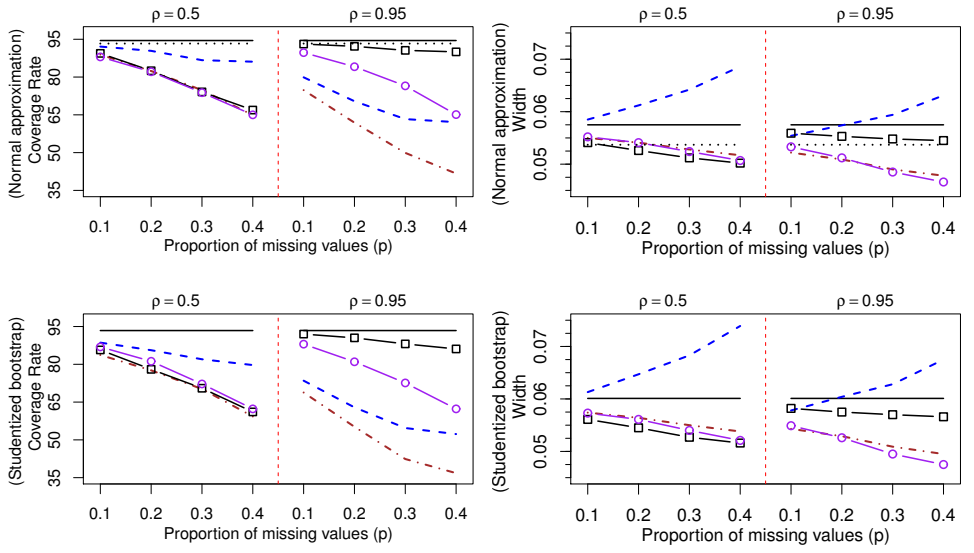

Figure 4. *Cont.*

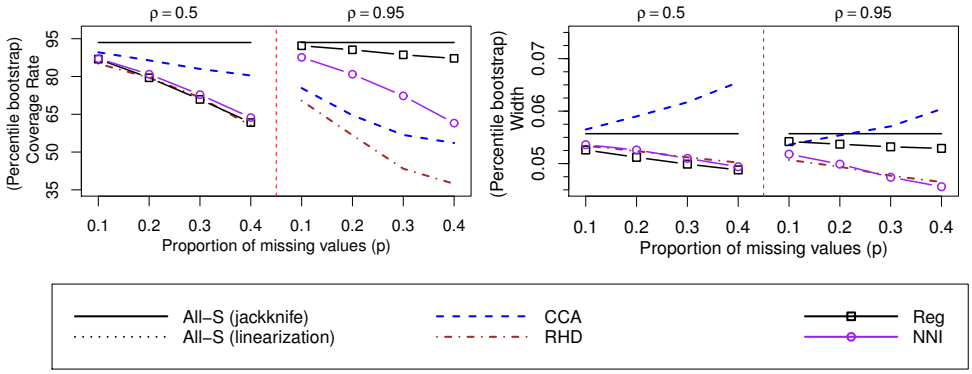

Figure 4. Values of CR and W associated with 95% confidence intervals for $G = 0.2$, and based on the jackknife variance estimator. The MAR mechanism is considered. Linearization and jackknife variances are compared using the normal approximation and the $All - S$ approach.

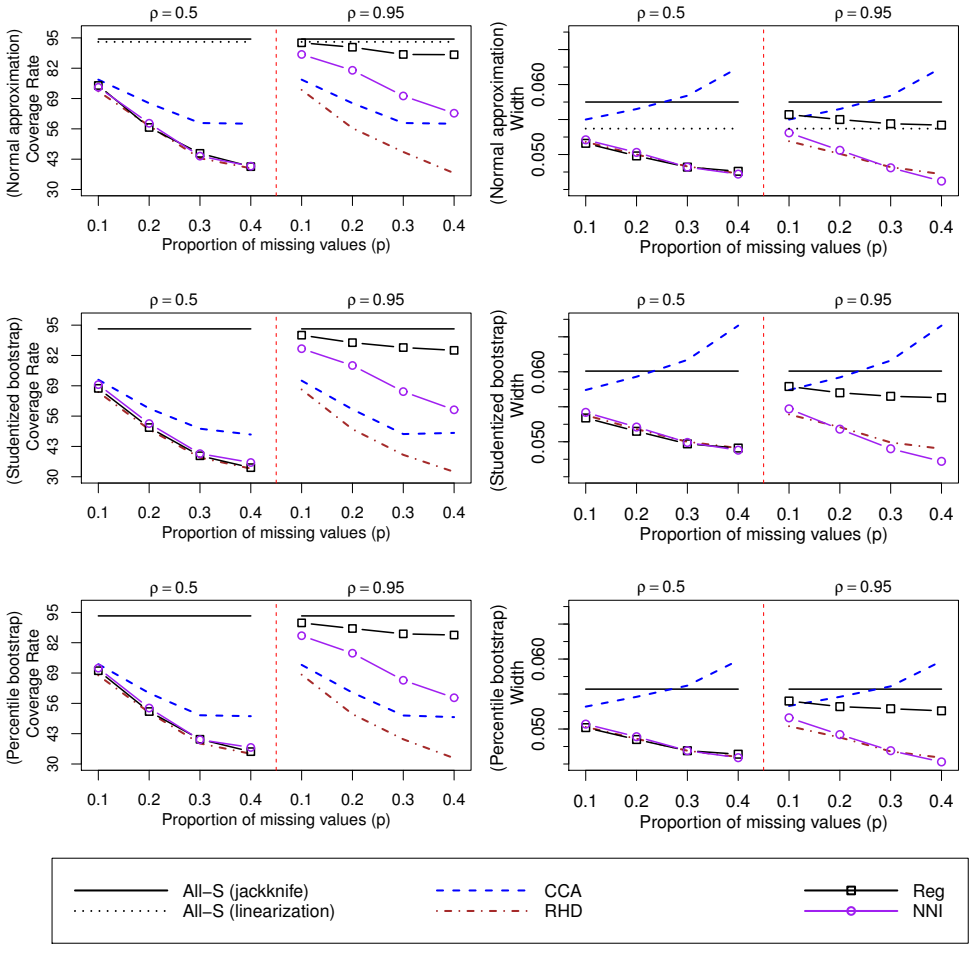

Figure 5. Values of CR and W associated to 95% confidence intervals for $G = 0.2$, and based on the jackknife variance estimator. The $MNAR$ mechanism is considered. Linearization and jackknife variances are compared using the normal approximation and the $All - S$ approach.

Finally, results from confidence intervals for $G = 0.6$ can be seen in Figures 6–8. First, we observe that confidence intervals perform worse when $G = 0.6$, since the values of CR are closer to the required nominal level (95%) when $G = 0.2$. This is probably due to the fact that estimates of G are slightly more biased when $G = 0.6$. Again, the *Reg* imputation method has the best coverage rates when ρ is large, and the RHD method performs poorly for the various scenarios analysed. For the MAR and $MNAR$ mechanisms, the various imputation methods provide unsatisfactory coverage rates when p is large, i.e., the bias observed under such situations has a relevant impact on the coverage rates.

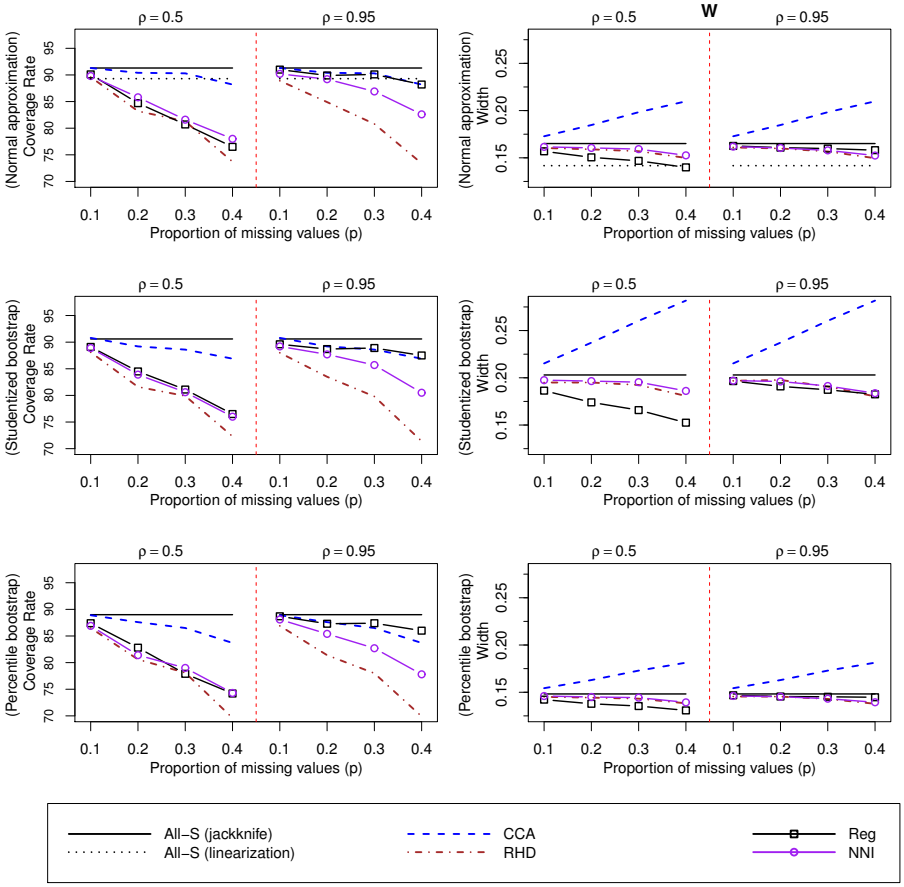

Figure 6. Values of CR and W associated to 95% confidence intervals for $G = 0.6$, and based on the jackknife variance estimator. The $MCAR$ mechanism is considered. Linearization and jackknife variances are compared using the normal approximation and the $All - S$ approach.

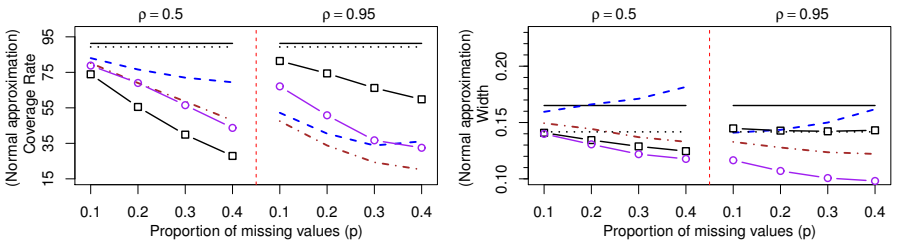

Figure 7. *Cont.*

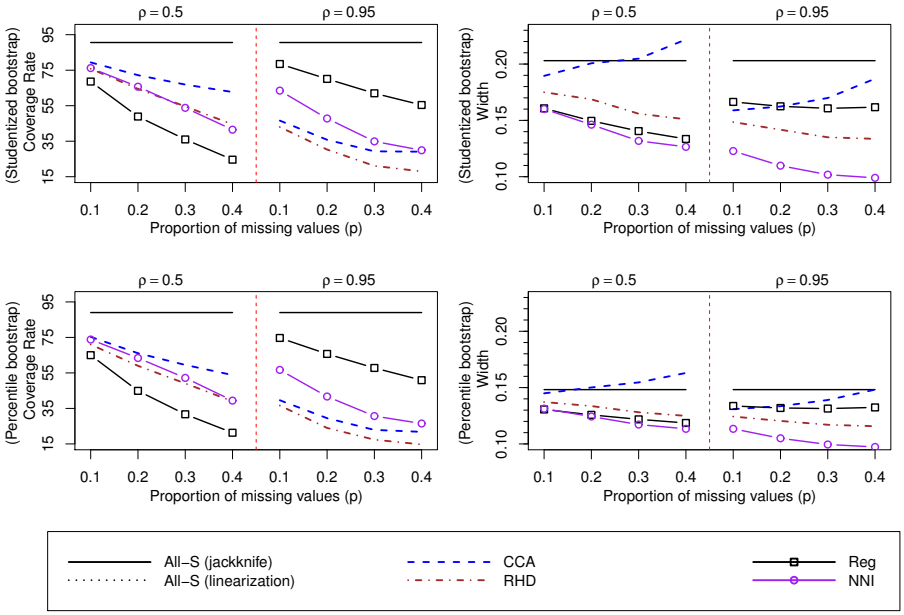

Figure 7. Values of CR and W associated to 95% confidence intervals for $G = 0.6$, and based on the jackknife variance estimator. The MAR mechanism is considered. Linearization and jackknife variances are compared using the normal approximation and the $All - S$ approach.

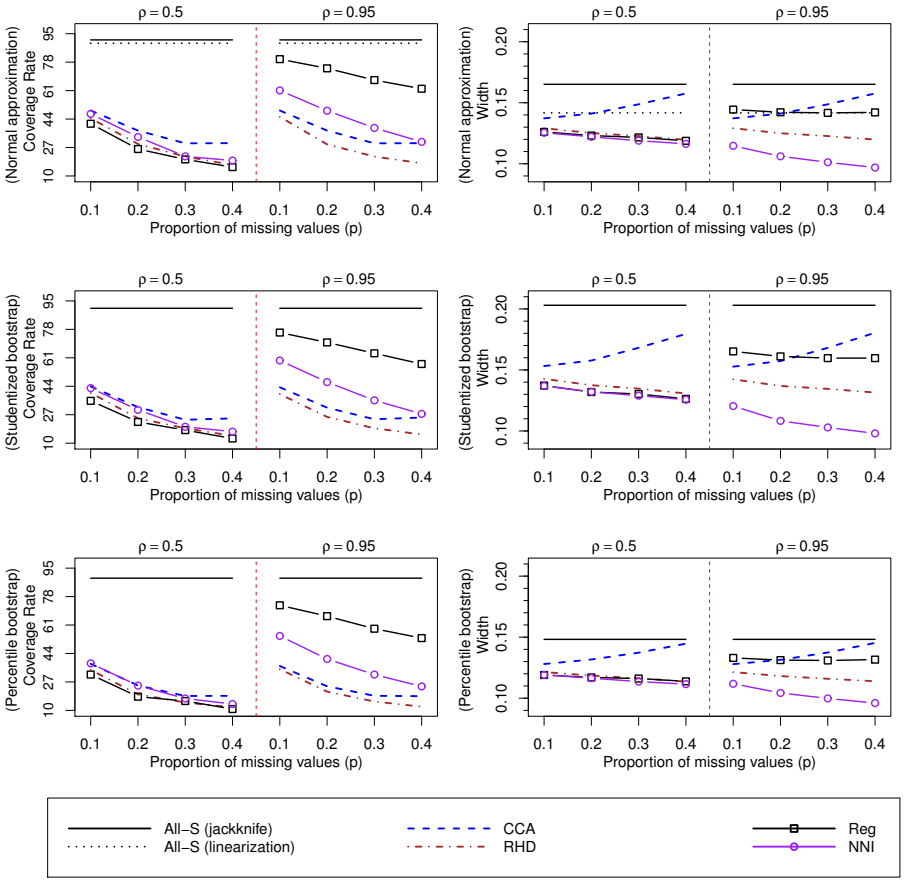

Figure 8. Values of CR and W associated to 95% confidence intervals for $G = 0.6$, and based on the jackknife variance estimator. The $MNAR$ mechanism is considered. Linearization and jackknife variances are compared using the normal approximation and the $All - S$ approach.

4. Conclusions

The problem of missing data may appear in many real-world applications, and various solutions can be applied to handle this problem. The solution adopted in this paper is to use traditional single imputation methods, since they are simple techniques widely used in many National Statistical Institutes, among other official organisms. The non-response bias is an important issue when dealing with missing data, which requires particular attention, especially when the $MNAR$ assumption holds. On the other hand, income inequality is a topic of interest in many economic studies, and the Gini index is probably the most commonly-used indicator to measure this phenomenon. In this paper, we empirically evaluate various traditional single imputation methods when applied to the estimation of G, analysing them for multiple interesting scenarios that may arise in practice. In particular, the empirical performance of the customary estimator of G is analysed, and different methods for the construction of confidence intervals are compared. Low and high income inequalities ($G = \{0.2, 0.6\}$), and weak and strong correlation coefficients ($\rho = \{0.5, 0.95\}$) are analysed. Finally, results are also presented for the various non-response mechanisms ($MCAR$, MAR and $MNAR$).

First, we analyse the various non-response mechanisms. For an $MCAR$ mechanism, CCA and Reg provide appropriate biases. RHD and NNI may yield slightly biased estimates, but they lie within a reasonable range. As expected, the non-response bias is not a problem in this case. In terms of efficiency, the various approaches give similar results for small proportions of missing data, but the CCA and RHD methods show poor values of $RRMSE$ when the proportion of missing data is large. The various methods give appropriate coverage rates for a small proportion of missing data. CCA provides satisfactory coverage rates for the various values of p, but the confidence intervals based on CCA widen considerably as the value of p increases. Reg and NNI also yield reasonable values of CR when ρ is large, while poor coverage rates are provided by the RHD method as the value of p increases. For an MAR mechanism, negligible biases are obtained when p is small, but the non-response bias can be a problem if p is large. The Reg method provides the best results, in terms of both RB and $RRMSE$, when ρ is large. The various methods only give reasonable coverage rates when p is small. The Reg method yields good coverage rates for the various values of p when ρ is large. For an $MNAR$ mechanism, the non-response bias is a problem for the various methods and the various values of p. However, the Reg method may produce reasonable biases when ρ is large, with values of RB that can be smaller than 5%, in absolute terms, when $p = 0.4$. Reasonable coverage rates are only obtained using the Reg method when ρ is large and p is smaller than 0.2, approximately.

Second, we analyse conclusions in terms of the Gini index G. We find that biases increase slightly, in absolute terms, as the income inequality increases. Consequently, coverage rates of confidence intervals are closer to the required confidence level (95%) as the Gini index decreases. As expected, the confidence intervals also widen as the value of G increases.

Third, we analyse the empirical performance of the various imputation methods according to the various proportions of missing data p. The biases of the CCA and Reg methods are not affected by p when the non-response mechanism is $MCAR$ and for low income inequalities. Otherwise, the empirical biases increase, in absolute terms, as the proportion of missing data increases. Similar conclusions are reached in terms of CR, i.e., the values of p do not have an impact on the coverage rates for the $MCAR$ mechanism when G is small and ρ is large. As expected, estimators are less efficient as the values of p increase. For an $MCAR$ mechanism, the width of confidence intervals based on the various imputation methods is not affected by the value of p, but the width of confidence intervals based on the CCA method increases considerably as the value of p increases. For the MAR and $MNAR$ mechanisms, the width of the various confidence intervals is affected by the value of p, although the effect is not relevant for the Reg method when ρ is large.

Fourth, we analyse conclusions in terms of correlation coefficient ρ. As expected, a larger ρ improves the estimation of the *Reg* and NNI imputation methods, as they make use of the auxiliary variable at the estimation stage. The *Reg* method clearly outperforms the NNI method when ρ is large. For a large value of ρ, the *Reg* method can provide empirical biases within a reasonable range for the various non-response mechanisms. However, with a low value of ρ, the non-response bias is a serious problem because the various imputation methods perform poorly in the presence of an $MNAR$ mechanism. In addition, the non-response bias is a problem when p is large and the non-response mechanism is MAR. The conclusions are similar in relation of CR, i.e., poor coverage rates are observed for a low value of ρ and for the MAR and $MNAR$ mechanisms, but the *Reg* method can provide appropriate values of CR when ρ is large.

Finally, we briefly describe and compare the empirical performance of the various methods investigated in this paper. CCA can be a solution when the non-response mechanism is $MCAR$ and ρ is small, but alternative approaches are preferred otherwise. This finding implies that CCA should rarely be used in practice, since the $MCAR$ assumption is often unrealistic.

The traditional RHD method provide poor estimates of the Gini index, even for the $MCAR$ mechanism when p is large. Note that alternative and more complex techniques can be used in the imputation process and for the various imputation methods, and may yield better results. For instance, the use of imputation classes is a well-known technique that may improve the accuracy of imputation methods.

The NNI method is a good solution when using auxiliary variables and may mitigate the non-response bias problem better than the *Reg* method when ρ is not extremely large.

The *Reg* method outperforms its competitors when ρ is large, registering good results in terms of the various empirical measures analysed in this paper and for the various non-response mechanisms. In particular, with a large value of ρ, the *Reg* method outperforms its competitors when p is large and for the MAR and $MNAR$ mechanisms.

As far as the construction of confidence intervals is concerned, we first find that confidence intervals based on the jackknife variance estimator provide coverage rates that are slightly better than those obtained using the linearization variance estimator. The normal approximation and the percentile bootstrap provide confidence intervals with similar empirical properties, while confidence intervals based on the studentized bootstrap are slightly wider than confidence intervals based on the normal approximation and the percentile bootstrap.

5. Discussion

This paper points to various potential areas for future research. First, serious biases have been detected in this study, and they have an important impact on the coverage of confidence intervals. Therefore, the question of how to reduce these biases is an interesting direction for future research. In particular, the bias corrected estimator $\widehat{G}$ is considered, but large biases, in absolute terms, are observed when the Gini index is large. The use of additional bias correction procedures has the potential to be a fruitful contribution that may improve the estimation of the Gini index and the corresponding properties of confidence intervals.

We consider single imputation methods, but multiple imputation is also a popular approach that may offer desirable features when it comes to the estimation of the Gini index. Additional single imputation methods can also be investigated, such as the $kNNI$ imputation method (see [54,55]), the EM algorithm (see [56,57]), and the Forest imputation method (see [58,59]), etc.

Recently, the empirical likelihood approach has been used for the construction of confidence intervals for the Gini index (see [22–24]). The analysis of the empirical likelihood methodology when dealing with missing data is also an interesting topic for future research.

This study could also be extended to unequal sampling designs and/or multiple auxiliary variables. In particular, the traditional jackknife technique requires an adjustment

for samples with unequal inclusion probabilities, and Campbell's jackknife (see [19,60]) can be a solution when samples selected under a general sampling design suffer from the problem of missing data.

Note that imputation methods have been evaluated here without using imputation classes, and more efficient results are expected for the various imputation methods when using said technique. Finally, we focus exclusively on the Gini index as the indicator to measure inequality. However, the quintile share ratio is another statistic commonly used to measure inequality. Thus, an interesting avenue for future research would be to analyse the performance of the quintile share ratio when single imputation methods are used and compare it with the results obtained in this paper.

Author Contributions: J.F.M.-R., P.J.M.-F. and E.Á.-V. have collaborated equally in the realization of this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been partially supported by the Ministry of Economy, Industry and Competitiveness, the Spanish State Research Agency (SRA) and European Regional Development Fund (ERDF) (project reference ECO2017-86822-R). This research has been partially supported by the Ministry of Economy, Industry and Competitiveness, the Spanish State Research Agency (SRA) and European Regional Development Fund (ERDF) (project reference ECO2017-84138-P).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MCAR	Missing Completely At Random
MAR	Missing At Random
MNAR	Missing Not At Random
SRSWOR	Simple Random Sampling Without Replacement
All-S	All units in the sample S
CCA	Complete Case Analysis
RHD	Random Hot Desk imputation method
Reg	Regression imputation method
NNI	Nearest Neighbour Imputation method
RB	Relative Bias
RRMSE	Relative Root Mean Square Error
CR	Coverage Rate
W	Width

References

1. Haziza, D.; Lesage, É. A discussion of weighting procedures for unit nonresponse. *J. Off. Stat.* **2016**, *32*, 129–145. [CrossRef]
2. Van Buuren, S. *Flexible Imputation of Missing Data*; CRC Press: Boca Raton, FL, USA, 2018.
3. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
4. Haziza, D.; Beaumont, J.F. On the construction of imputation classes in surveys. *Int. Stat. Rev.* **2007**, *75*, 25–43. [CrossRef]
5. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*, 3rd ed.; John Wiley & Sons: New York, NY, USA, 2019.
6. Särndal, C.E.; Swensson, B.; Wretman, J. *Model Assisted Survey Sampling*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2003.
7. Rubin, D.B. Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* **1996**, *91*, 473–489. [CrossRef]
8. Carpenter, J.; Kenward, M. *Multiple Imputation and Its Application*; John Wiley & Sons: Chichester, UK, 2012.
9. Allison, R.A.; Foster, J.E. Measuring health inequality using qualitative data. *J. Health Econ.* **2004**, *6*, 505–524. [CrossRef] [PubMed]
10. Boyce, J.K.; Zwickl, K.; Ash, M. Measuring environmental inequality. *Ecol Econ.* **2016**, *124*, 114–123. [CrossRef]
11. Ferreira, F.H.; Gignoux, J. The measurement of educational inequality: Achievement and opportunity. *World Bank Econ. Rev.* **2014**, *28*, 210–246. [CrossRef]

12. Solt, F. Measuring income inequality across countries and over time: The standardized world income inequality database. *Soc. Sci. Q.* **2020**, *101*, 1183–1199. [CrossRef]
13. Ravallion, M. Income inequality in the developing world. *Science* **2014**, *344*, 851–855. [CrossRef]
14. Gini, C. *Variabilità e mutabilità. Reprinted in Memorie di Metodologica Statistica*; Pizetti, E., Ed.; Libreria Eredi Virgilio Veschi: Rome, Italy,1912.
15. Kendall, M.; Stuart, A. *The Advanced Theory of Statistics: Vol. 1. Distribution Theory*, 4th ed.; Charles Griffin: London, UK, 1977.
16. Lerman, R.I.; Yitzhaki, S. A note on the calculation and interpretation of the Gini index. *Econ. Lett.* **1984**, *15*, 363–368. [CrossRef]
17. Deltas, G. The small-sample bias of the Gini coefficient: Results and implications for empirical research. *Rev. Econ. Stat.* **1979**, *44*, 870–872.
18. Davidson, R. Reliable inference for the Gini index. *J. Econom.* **2009**, *150*, 30–40. [CrossRef]
19. Berger, Y.G. A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient. *J. Off. Stat.* **2008**, *24*, 541–555.
20. Deville, J.C. Variance estimation for complex statistics and estimators: linearization and residual techniques. *Surv. Methodol.* **1999**, *25*, 193–204.
21. Langel, M.; Tillé, Y. Variance estimation of the Gini index: Revisiting a result several times published. *J. R. Stat. Soc. A Stat. Soc.* **2013**, *176*, 521–540. [CrossRef]
22. Qin, Y.; Rao, J.; Wu, C. Empirical likelihood confidence intervals for the gini measure of income inequality. *Econ. Modllng.* **2010**, *27*, 1429–1435. [CrossRef]
23. Wang, D.; Zhao, Y.; Gilmore, D.W. Jackknife empirical likelihood confidence interval for the Gini index. *Stat. Probab. Lett.* **2016**, *110*, 289–295. [CrossRef]
24. Berger, Y.; Gedik Balay, İ. Confidence intervals of Gini coefficient under unequal probability sampling. *J. Off. Stat.* **2020**, *36*, 237–249. [CrossRef]
25. Giorgi, G.M.; Gigliarano, C. The Gini concentration index: a review of the inference literature. *J. Econ. Surv.* **2017**, *31*, 1130–1148. [CrossRef]
26. Balaji, H.; Mahmoud, H. The Gini index of random trees with an application to caterpillars. *J. Appl. Probab.* **2017**, *54*, 701–709. [CrossRef]
27. Ren, Y.; Zhang, P.; Dey, D.K. Investigating Several Fundamental Properties of Random Lobster Trees and Random Spider Trees. *Methodol. Comput. Appl. Probab.* **2021**, 1–17. DOI: 10.1007/s11009-021-09863-9 [CrossRef]
28. Parsa, M.; Di Crescenzo, A.; Jabbari, H. Analysis of reliability systems via Gini-type index. *Eur. J. Oper. Res.* **2018**, *264*, 340–353. [CrossRef]
29. Ma, J. Generalised grey target decision method for mixed attributes based on the improved Gini–Simpson index. *Soft Comput.* **2018**, *23*, 13449–13458. [CrossRef]
30. Atkinson, A.B. On the measurement of inequality. *J. Econ. Theory* **1970**, *2*, 244–263. [CrossRef]
31. Evans, M.D.; Kelley, J.; Kelley, S.M.; Kelley, C.G. Rising Income Inequality During the Great Recession Had No Impact on Subjective Wellbeing in Europe, 2003–2012. *J. Happiness Stud.* **2019**, *20*, 203–228. [CrossRef]
32. Detollenaere, J.; Desmarest, A.S.; Boeckxstaens, P.; Willems, S. The link between income inequality and health in Europe, adding strength dimensions of primary care to the equation. *Soc. Sci. Med.* **2018**, *201*, 103–110. [CrossRef] [PubMed]
33. Zagorski, K.; Evans, M.D.; Kelley, J.; Piotrowska, K. Does national income inequality affect individuals' quality of life in Europe? Inequality, happiness, finances, and health. *Soc. Indic. Res.* **2014**, *117*, 1089–1110. [CrossRef]
34. Rueda, M.M.; Muñoz, J.F. Estimation of poverty measures with auxiliary information in sample surveys. *Qual. Quant.* **2011**, *45*, 687–700. [CrossRef]
35. Langel, M.; Tillé, Y. Statistical inference for the quintile share ratio. *J. Stat. Plan. Inference* **2011**, *141*, 2976–2985. [CrossRef]
36. Rao, J.N.K. On variance estimation with imputed survey data. *J. Am. Stat. Assoc.* **1996**, *91*, 499-–506. [CrossRef]
37. Zhong, H. The impact of missing data in the estimation of concentration index: A potential source of bias. *Eur. Health Econ.* **2010**, *11*, 255–266. [CrossRef] [PubMed]
38. Chen, Y.; Fu, D. Measuring income inequality using survey data: the case of China. *J. Econ. Inequal.* **2015**, *13*, 299–307. [CrossRef]
39. Ardington, C.; Lam, D.; Leibbrandt, M.; Welch, M. The sensitivity to key data imputations of recent estimates of income poverty and inequality in South Africa. *Econ. Model.* **2005**, *23*, 822–835. [CrossRef]
40. Jenkins, S.P. World income inequality databases: an assessment of WIID and SWIID. *J. Econ. Inequal.* **2015**, *13*, 629–671. [CrossRef]
41. Yitzhaki, S. More than a dozen alternative ways of spelling Gini. *Res. Econ. Inequal.* **1998**, *8*, 13–30.
42. David, H.A. *Order Statistics*; Wiley: NewYork, NY, USA, 1970.
43. Ogwang, T. A convenient method of computing the Gini index and its standard error. *Oxf. Bull. Econ. Stat.* **2000**, *62*, 123–129. [CrossRef]
44. Demnati, A.; Rao, J.N.K. Linearization variance estimators for survey data. *Surv. Methodol.* **2004**, *30*, 17–26.
45. Yitzhaki, S. Calculating jackknife variance estimators for parameters of the Gini method. *Surv. Methodol.* **1991**, *9*, 235–239.
46. Karagiannis, E.; Kovačević , M. A method to calculate the jackknife variance estimator for the Gini coefficient. *Oxf. Bull. Econ. Stat.* **2000**, *62*, 119–122. [CrossRef]
47. Kuan, X. Inference for generalized Gini indices using the iterated bootstrap method. *J. Bus. Econ. Statist.* **2000**, *18*, 223–227.

48. Giorgi, G.M.; Palmitesta, P.; Provasi, C. Asymptotic and bootstrap inference for the generalized gini indices. *Metron* **2006**, *64*, 107–124.
49. Muñoz, J.F.; Rueda, M. New imputation methods for missing data using quantiles. *J. Comput. Appl. Math.* **2009**, *232*, 305–317. [CrossRef]
50. Andridge, R.R.; Little, R.J. A review of hot deck imputation for survey non-response. *Int. Stat. Rev.* **2010**, *78*, 40–64. [CrossRef]
51. Healy, M.; Westmacott, M. Missing values in experiments analysed on automatic computers. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1956**, *5*, 203–206. [CrossRef]
52. Chen, J.; Shao, J. Nearest neighbor imputation for survey data. *J. Off. Stat.* **2000**, *16*, 113–131.
53. Gower, J.C. A general coefficient of similarity and some of its properties. *Biometrics* **1971**, *27*, 857–871. [CrossRef]
54. Kim, K.Y.; Kim, B.J.; Yi, G.S. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinform.* **2004**, *5*, 1–9. [CrossRef]
55. Moya-Fernández, P.J.; López-Ruiz, S.; Guardiola, J.; González-Gómez, F. Determinants of the acceptance of domestic use of recycled water by use type. *Sustain. Prod. Consum.* **2021**, *27*, 575–586. [CrossRef]
56. McLachlan, G.J.; Krishnan, T. *The EM Algorithm and Extensions*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 2007.
57. Lange, K. A gradient algorithm locally equivalent to the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1995**, *57*, 425–437. [CrossRef]
58. Pantanowitz, A.; Marwala, T. Missing data imputation through the use of the random forest algorithm. In *Advances in Computational Intelligence*; Springer: Berlin, Germany, 2009; pp. 53–62.
59. Tang, F.; Ishwaran, H. Random forest missing data algorithms. *Stat. Anal. Data. Min.* **2007**, *10*, 363–377. [CrossRef]
60. Campbell, N.A. Robust procedures in multivariate analysis I: Robust covariance estimation. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1980**, *29*, 231–237. [CrossRef]