MISCLASSIFICATION IN AGRICULTURAL SURVEYS BASED ON IMAGE REMOTE SENSING

RAYDONAL OSPINA[1,2], CRISTIANO FERRAZ [1], ANDRÉ LEITE [1] AND HEMÍLIO COELHO [1,3]

[1]*Department of Statistics, CASTLab, Universidade Federal de Pernambuco, Recife, Brazil*

[2]*Department of Statistics, Universidade Federal da Bahia, Salvador, Brazil*

[3]*Department of Statistics, Universidade Federal da Paraíba, Paraíba, Brazil*

ABSTRACT. Stratification is a sampling technique widely used in surveys to improve efficiency of estimates. Strata building processes are subject to different types of errors that can lead to misclassification of sampled units, a discrepancy between the stratum from which a unit was selected, and the stratum the unit belongs to in reality. This paper investigates the misclassification problem motivated by surveys using remote sensing stratified area frames of square segments to generate agricultural statistics. Estimators coping with the problem are introduced and their statistical performance is investigated using a Monte Carlo simulation experiment. The study rely on a real-case motivated scenario in which area frames of square segments were applied to surveys carried out in two Brazilian municipalities, aiming at comparing different sampling design strategies to generate efficient agricultural statistics. Simulation results indicate that the adoption of a naive unweighted estimator can introduce considerable bias. It also indicates that in the absence of the needed auxiliary information to use a post-stratified estimator, the best choice is to keep the design-based original sample stratification estimator.

KEY WORDS: errors-in-stratification, agricultural surveys, master frame, area sampling, area frame of square segments, crowd-sourcing, post-stratification

1. INTRODUCTION

Misclassification problems related to survey stratification occur when field collected data differs from auxiliary information used to stratify the population. In practice this means sampled units present characteristics that classify them in strata that do not correspond to the ones from which they were originally selected. When coping with strata misclassification in practice, options considered include keeping the misclassified sampled units in the original stratum or move them to the actual one. In addition, the choice of an appropriate estimator for the parameter of interest needs to take into account the type of information available. Several situations can lead to misclassification problems in practice. (Mulrow & Woodburn 1990) describe a study of the effect of errors in stratification in a problem of business taxation. In their case, a small simulation study was carried out investigating bias effects on estimates, but not addressing variance concerns. (Lamas et al. 2010) and (Abreu et al. 2010) report an example of misclassification of tracts as agricultural or non-agricultural in the June Area Survey, carried out by NASS, the US National Agricultural Statistics Service, leading to discrepancies between the estimated number of agricultural farms using the survey data, and the agricultural census at that time. A generalized linear model was used to model under-counting and to provide corrections for estimates. (Jang et al. 2009) describe misclassification in stratification exemplifying on the discrepancies observed at the National Survey of Recent

College Graduates (NSRCG), between the race/ethnicity registered in administrative records, and the self reported one. Another general situation leading to strata misclassification problems happens when there is a large time lag between the period the sample was selected, and the period of the field work for data collection. Sometimes this time lag is related to the stratification process itself. Agricultural surveys using satellite imagery to foster stratification of segments of lands, for example, are subject to this source of error when the image used for strata classification is not updated. This can represent in practice a time lag discrepancy from the information used to build the strata and the reality found in the field (Boryan et al. 2014; Gao et al. 2017). In this paper, misclassification is studied in the context of agricultural surveys based on remote sensing stratified area samples of squared segments. The study rely on a real-case motivated scenario where the efficiency of area frames of square segments, as master sampling frames for agricultural statistics, was investigated by surveys carried out in two Brazilian municipalities: Goiana, and Santos Dumont. The surveys were part of the studies carried out by the Food and Agriculture Organization of the United Nations - FAO's Global Strategy to Improve Agricultural and Rural Statistics (GSARS), described in detail in (Ferraz et al. 2018) and (FAO 2018). The stratification process used supervised classification of satellite image points within squared segments, relying on the idea of crowd-sourcing to foster stratification. Discrepancies between the stratification generated by imagery analysis and the stratification based on field collected data define the misclassification problem that motivates this paper. Several estimators that could cope with the situation are introduced and their performances analysed via Monte Carlo simulation. The computational experiment replicates stratified samples drawn from an artificial population built to resemble major characteristics of Goiana, simulating a misclassification process with the same rates found in practice, by data collection on the field.

2. Misclassification in crowd-sourcing

One of many aspects investigated by the GSARS surveys carried out in Brazil was stratification efficiency of area frames of square segments based on free satellite imagery, and supervised classification using crowd-sourcing (Saralioglu & Gungor 2019; Laso Bayas et al. 2016; See et al. 2016; Howe 2006). Area frame stratification applied to Brazil's experiment used free imagery resources from Google EarUnfortunately, your manuscript is not a good fit for the IJRS. Whilst interesting, your paper lacks impact due to a lack of scope and novelty of the research. The direction of the paper is not clear. The literature review is rather weak and a number of papers do not concern remote sensing research. I am not convinced by the title as well. th and Open Street Map (powered by `esri.com`) to support photo interpretation of points in square segments, by volunteers. The major advantage of this method relies on its low cost of implementation in terms of budget and timing to achieve stratification of the full territory. According to IBGE (`cidades.ibge.gov.br`), Goiana has an area of 445.814 Km2, and Santos Dumont, 637.373 Km2.

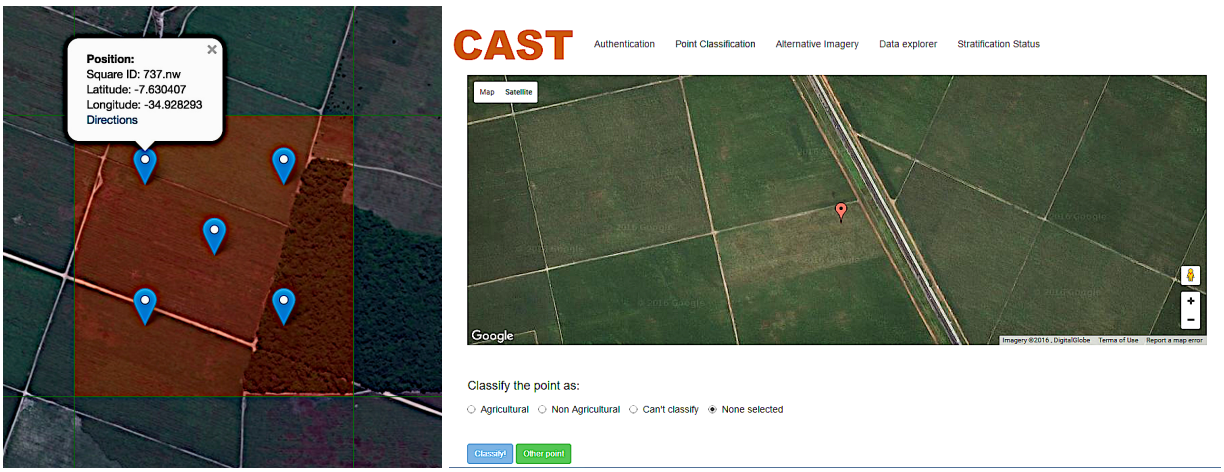
Area frames for Goiana and Santos Dumont used square segments of 49 hectares, and within each segment, a sample of 5 points was taken, following the pattern shown in Figure 1(a). The land cover in each point was assessed, by volunteers, and classified according to two classes: *"Cropland"* or *"Non-Cropland"*. Figure 1(b). shows an example of classification screen used in the experiment. After classifying all the points, square segments were grouped into strata according to Table I.

Using limited resources of six volunteers, stratification of Goiana was achieved in one week, while the same task for Santos Dumont, a larger territory, was completed in three weeks, using only three volunteers. Although the number of people involved in assessing images for the experiment is extremely low, the idea of crowd-sourcing, were a much larger number of people can be involved in the process, is the motivation

to make the method feasible for much larger territories, justifying the use of the term even for the small GSARS study.

TABLE I. Strata definitions

Strata	Description
1. Highly cropland	Segments with 4 or 5 points classified as "Cropland."
2. Cropland	Segments with 2 or 3 points classified as "Cropland."
3. Non-cropland	Segments with at most 1 point classified as "Cropland."



(a) Pattern of points in a segment return (b) Crowdsource screen for point interpretation

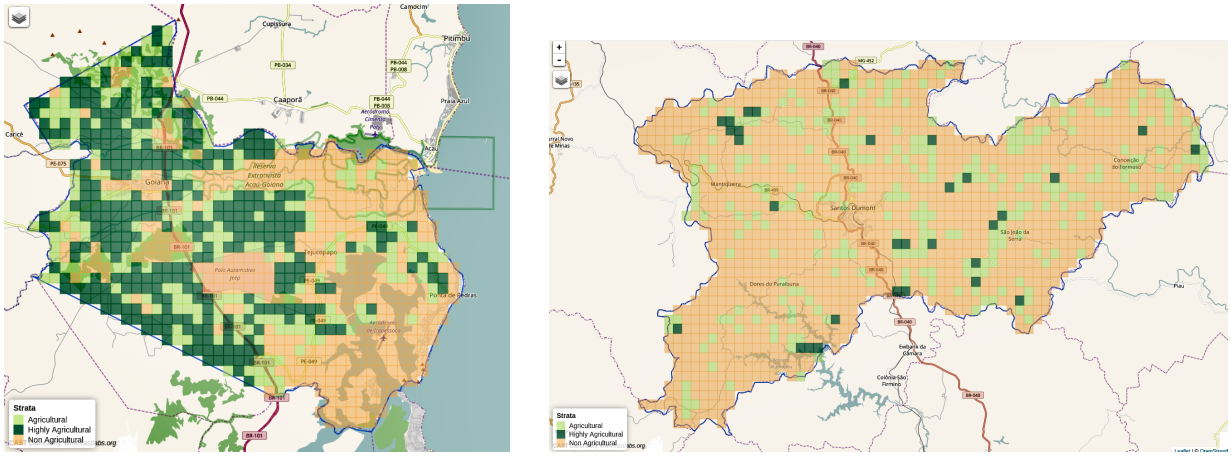
FIGURE 1. Classification of points in segments

The low cost process used to stratify the grid of square segments to at least three sources of errors: the degree of experience of the volunteer photo interpreter, the image quality and the time the image was acquired. The Photo interpretation was carried with a rough rate of 500 points per day approximately, 50,000 per week, with six volunteers. Images from 2010 and 2016 were used to stratify Goiana and Santos Dumont, respectively. Although it is possible to use even more up to date images than in Santos Dumont (based on European satellite Sentinel-2, for instance) this would not necessarily mean misclassification problems would be reduced to the point they would not occur. Therefore, to investigate the potential impact of strata misclassification on estimates is necessary and relevant.

Area frames of square segments on both municipalities were stratified according to the process described leading to the results shown in Figure 1(a).

A sample of 60 segments in each city was allocated to the strata according to Table II.

Tables III and IV introduce the sample rates of errors observed in the strata formation at the study. These numbers provide estimates of the proportions of segments misclassified in the whole grid of square segments. Based on them, approximately 58% of the segments in Goiana's area frame were classified in the right strata, while 79% of the segments in Santos Dumont's area frame were correctly classified. Using data information from both municipalities, the method classified 71% of the segments in the correct



(a) Stratification of Goiana's area frame. (b) Stratification of Santos Dumont's area frame

FIGURE 2. Stratified area frame

TABLE II. Original square segments classification in strata

	Municipality			
	Goiana		Santos Dumont	
Strata	Frame	Sample	Frame	Sample
1. Highly Cropland	314	42	31	29
2. Cropland	232	15	288	16
3. Non-Cropland	376	3	1078	15

strata. These data are used to study options to correct estimates taking into account the fact that the original stratification process was not 100% accurate.

Let p_{hj} be the sample proportion of segments classified in stratum j in the field, given all the segments classified in stratum h when using satellite imagery photo interpretation. These numbers correspond to the row percent values in Tables III and IV. In Goiana, for instance, $p_{12} = 0.119$ is the proportion of segments in the sample that were originally selected from stratum 1 (*Highly cropland*) but after data collection, were classified in stratum 2 (*Cropland*). This same proportion for Santos Dumont was 0.2759. It is noted that, in Goiana, $p_{31} = 0$, and in Santos Dumont, $p_{21} = p_{32} = p_{31} = 0$.

3. CORRECTING ESTIMATES

Let t_c be the total area cultivated with a given crop c, in a given municipality. t_c can be expressed, based on strata defined over the population, as

$$t_c = \sum_{h=1}^{H} \sum_{k \in U_h} y_k,$$

where h is the index identifying the original stratum (based on photo interpretation), U_h is the set of all segments of the population from stratum h and y_k is the area cultivated with crop c in segment $k \in U_h$. The general class of homogeneous linear estimators for this parameter is composed by estimators that can be written as

TABLE III. Goiana strata misclassification analysis

Satellite (lines) by Field (columns)				
Count Total % Col % Row %	**Highly Cropland (1)**	**Cropland (2)**	**Non-Cropland (3)**	**Total**
Highly Cropland (1)	**27** 45.00 87.10 64.29	**5** 8.33 45.45 11.90	**10** 16.67 55.56 23.81	**42** 70.00
Cropland (2)	**4** 6.67 12.90 26.67	**5** 8.33 45.45 33.33	**6** 10.00 33.33 40.00	**15** 25.00
Non-Cropland (3)	**0** 0.00 0.00 0.00	**1** 1.67 9.09 33.33	**2** 3.33 11.11 66.67	**3** 5.00
Total	31 51.67	11 18.33	18 30.00	**60**

TABLE IV. Santos Dumont strata misclassification analysis

Satellite (lines) by Field (columns)				
Count Total % Col % Row %	**Highly Cropland (1)**	**Cropland (2)**	**Non-Cropland (3)**	**Total**
Highly Cropland (1)	**7** 11.67 100.00 24.14	**8** 13.33 88.89 27.59	**14** 23.33 31.80 48.27	**29** 48.33
Cropland (2)	**0** 0.00 0.00 0.00	**1** 1.66 11.11 6.25	**15** 25.00 34.10 93.75	**16** 26,67
Non-Cropland (3)	**0** 0.00 0.00 0.00	**0** 0.00 0.00 0.00	**15** 25.00 34.10 100.00	**15** 25.00
Total	7 16.67	9 15.00	44 73.33	**60**

$$\hat{t}_c = \sum_{h=1}^{H} \sum_{k \in U_h} I_k w_k y_k$$

where I_k is the sample inclusion indicator for element k, and w_k does not contain any information concerning the response variable. In such conditions, sampling is non-informative and the mean square error

of $\hat{t}_c$ can be written as

$$\hat{t}_c = \sum_{h=1}^{H} \sum_{g=1}^{H} \sum_{k \in U_h} \sum_{l \in U_g} y_k y_l E_p (I_k w_k - 1)(I_l w_l - 1),$$

where $E_p(.)$ is the expectation with respect to the sample design $p(.)$. When $w_k = \pi_k^{-1}$, with π_k as the inclusion probability of element k, the estimator $\hat{t}_c$ corresponds to the traditional Horvitz-Thompson estimator.

Consider N_{hj} as the number of segments in the area frame that were originally classified in stratum h, based on image interpretation, and later classified in stratum j, based on field observation. Table V shows the population crosstabulation of N_{hj}, illustrating the relationship between the misclassified cells and the actual strata totals N_{+j}. Let n_{hj} be the number of segments in the area frame sample that were originally classified in stratum h, based on image interpretation, and later classified in stratum j, based on field observation. Table Vshows the sample crosstabulation of n_{jh}, illustrating the relationship between the misclassified sample cells and the actual marginal totals n_{+j}. If no misclassification occurs, then the design strata is the same as the actual strata, with $N_{h+} = N_{+j}$ and $n_{h+} = n_{+j}$.

TABLE V. Crosstabulation of stratum sizes

	Actual Strata							
Design	Population				Sample			
Strata	1	2	3	Total	1	2	3	Total
1	N_{11}	N_{12}	N_{13}	N_{1+}	n_{11}	n_{12}	n_{13}	n_{1+}
2	N_{21}	N_{22}	N_{23}	N_{2+}	n_{21}	n_{22}	n_{23}	n_{2+}
3	N_{31}	N_{32}	N_{33}	N_{3+}	n_{31}	n_{32}	n_{33}	n_{3+}
Total	N_{+1}	N_{+2}	N_{+3}	N_{++}	n_{+1}	n_{+2}	n_{+3}	n_{++}

In practice, sample selection is done within strata formed by the rows of Table V, and when misclassification is present, the values of N_{hj} for h differing to j, are not zero. Table III for example, shows that in Goiana, $n_{12} = 5$, meaning 5 squares originally selected from stratum 1 were actually found in stratum 2 in the field. In such cases the population and the sample can be partitioned in two groups. One based on the design strata, and another based on the field strata. Taking Goiana again as an example, Table III shows that according to the design strata, the sample of segments is partitioned in groups of size $n_{1+} = 42$, $n_{2+} = 15$ and $n_{3+} = 3$, while according to the field strata, the sample is partitioned in groups of size $n_{+1} = 31$, $n_2 = N_{+2} = 11$ and $n_{+3} = 18$. Let the set of sampled units classified by the field strata, of size n_{+j}, be denoted by A_j, for $j = 1, ..., H$ and keep the notation S_h for the set of sampled units of size n_{h+}, classified by the design strata, with $h = 1, ...H$. Estimators coping with errors in stratification consider either using A_j or S_h as the sample of units, and may modify the form of the sampling weight w_k as well. This paper considers the same set of estimators described in (Mulrow & Woodburn 1990), this time related to the use of a remote sensing stratified area sampling design with strata errors observed in crowd-sourcing based classification. The estimators' description, based on a stratified random sampling design (STSI), are presented next.

(1) **Basic estimator:** The basic estimator corresponds to ignore errors in stratification, keeping the design strata, and choosing to use $w_k = N_{h+}/n_{h+}$, for $k \in S_h$. Thus, under an STSI design, this estimator can be written as follows:

$$\hat{t}_c = \sum_{h=1}^{H} \sum_{k \in S_h} (N_{h+}/n_{h+})\, y_k = \sum_{h=1}^{H} N_{h+} \overline{y}_{h+},$$

where $\overline{y}_{h+}$ is the mean of all the segments in the sample originally selected from stratum h. In this case, no matter the sampling units actually belong to another stratum, they are kept in the design stratum anyway.

(2) **Unweighted estimator:** The unweighted estimator corresponds to proceed corrections to the basic estimator using only the field stratum observation in the sample. According to Table III, in Goiana, a sample of size $n_{1+} = 42$ was selected from stratum 1, and from these, only $n_{+1} = 31$ remained in stratum 1. Stratum 1, by design, was composed by $N_{1+} = 314$ segments (see Table II). Then for a segment k in stratum 1, $w_k = [314 - (42 - 31)]/31$. Thus, $w_k = \left[N_{j+} - \left(n_{j+} - n_{+j}\right)\right]/n_{+j}$, for $k \in A_j$. Note that corrections are made on both, the stratum population and the respective sample size. These corrections, however, take into account only those observed sampled units that changed strata. No tentative is made to proceed corrections to the non-sampled segment numbers. Thus, under STSI design, this estimator can be written as follows:

$$\hat{t}_c = \sum_{j=1}^{H} \sum_{k \in A_j} \left[N_{j+} - \left(n_{j+} - n_{+j}\right)\right] y_k/n_{+j}.$$

(3) **Weighted estimator:** The weighted estimator uses sample field data to correct information for the observed and the non-observed segments. Let $\hat{N}_{+j}$ be an estimator for the number of segments over the whole area frame that belong to stratum j. Let $w_k = \hat{N}_{+j}/n_{+j}$, for $k \in A_j$. Then,

$$\hat{t}_c = \sum_{j=1}^{H} \sum_{k \in A_j} \left(\hat{N}_{+j}/n_{+j}\right) y_k = \sum_{j=1}^{H} \hat{N}_{+j} \overline{y}_{+j}$$

where $\hat{N}_{+j} = \sum_{h=1}^{H} N_{h+} p_{hj}$ and p_{hj} are the proportions of segments in the sample that were originally classified in stratum h but actually belong to stratum j and $\overline{y}_{+j}$ is the mean of the segments in A_j.

(4) **Post-Stratified estimator:** The post-stratified estimator uses the most updated information for both, the sample and the population sizes. So, for a segment k in A_j, $w_k = N_{+j}/n_{+j}$. In practice, no information about the true values N_{+j} is available, but this estimator is important as a benchmark for the performances of the remaining ones.

4. Simulation

In order to evaluate the statistical performances of the considered estimators, an artificial population resembling the field characteristics of Goiana was built to provide support for a Monte Carlo simulation. The population was covered by an area frame of 922 square segments, keeping the strata population sizes presented in Table II: 314, 232 e 376, for strata 1, 2 and 3, respectively. Strata building observed the same rates of misclassification found in Goiana's experiment (see Table II). FAO's GSARS' experiments used a sample of size 60 due to budget constraints. In this study, a sample size of 120 with allocation proportional to stratum sizes was considered. The estimators described in Section 3 were applied to each one of 5,000 Monte Carlo sample replications, and their distribution, investigated. Population means and respective variances for the area cultivated in hectares with sugarcane in each stratum were obtained

empirically from Goiana's study and are $\mu_1 = 28.35$, $\mu_2 = 22.81$ and $\mu_3 = 1.42$, and $\sigma_1^2 = 2.5$, $\sigma_2^2 = 5.0$ and $\sigma_3^2 = 0.5$.

4.1. **Parameter's design.** Let y_k be the area cultivated with a crop, observed in segment k. Such crop is said to be sugarcane for the sake of matching Goiana's information. Define δ_{hj} as an indicator variable that segment k was sampled from the design stratum h, and later reclassified, in the field stratum j, so that δ_{hj}~Bernoulli (p_{hj}). The values of y_k were generated within each stratum h, according to the following models:

$$\xi_1 : y_k = \mu_1 + \varepsilon_{1k}, \text{ if } \delta_{h1} = 1;$$
$$\xi_2 : y_k = \mu_2 + \varepsilon_{2k}, \text{ if } \delta_{h2} = 1;$$
$$\xi_3 : y_k = \mu_3 + \varepsilon_{3k}, \text{ if } \delta_{h3} = 1;$$

5. Simulation results

Estimators were evaluated based on their estimates for the total area cultivated with sugarcane (total), bias, relative bias, root of the mean squared error ($\sqrt{\text{MSE}}$) and standard deviation (SD) over Monte Carlo sample replicates. Codes and data are available upon request, from the corresponding author.

The summary statistics for the results of the Monte Carlo experiment is presented in Tables VI. The numbers show that the unweighted estimator is the one with poorer performance, among all the ones investigated, showing a non-negligible bias and the largest mean square error. As expected, the post-stratified estimator is the one showing the best performance, with lower mean square error. Both, the basic and the weighted estimators show similar performances with no bias mean square error about the same magnitude, larger than the post-stratified, and much smaller than the unweighted estimator.

The performance of the estimators can be further visualized in Figure 3, showing their estimated sample distributions.

Table VI. Summary Statistics. Sample size: 120

Estimator	Total	Bias	Rel.bias	$\sqrt{\text{MSE}}$	SD
Basic	10407.28	-0.74	-0.00	736.07	736.15
PostStrata	10409.02	0.99	0.00	172.72	172.74
Unweighted	14120.02	3712.00	0.36	3724.64	306.61
Weighted	10405.22	-2.80	-0.00	736.30	736.37

6. Concluding remarks

The simulation investigated the effect of different estimators trying to make corrections to cope with strata misclassification of segments in area sampling. The post-stratified estimator shows better performance than the others, as expected. However, in practice, no information regarding the actual size of each stratum is likely to be available, making the use of such estimator not possible. The investigation has shown the performances of both, the basic and the weighted estimators, are similar for producing estimates to the whole population. They are both unbiased, and have shown mean square errors about the same magnitude. In this scenario, the simplest estimator, the basic one, should be preferable, as it implies a smaller cost. The investigation has also shown unweighted estimators should be avoided as they have a poor performance, introducing large bias. In summary, if auxiliary information is is available, one should

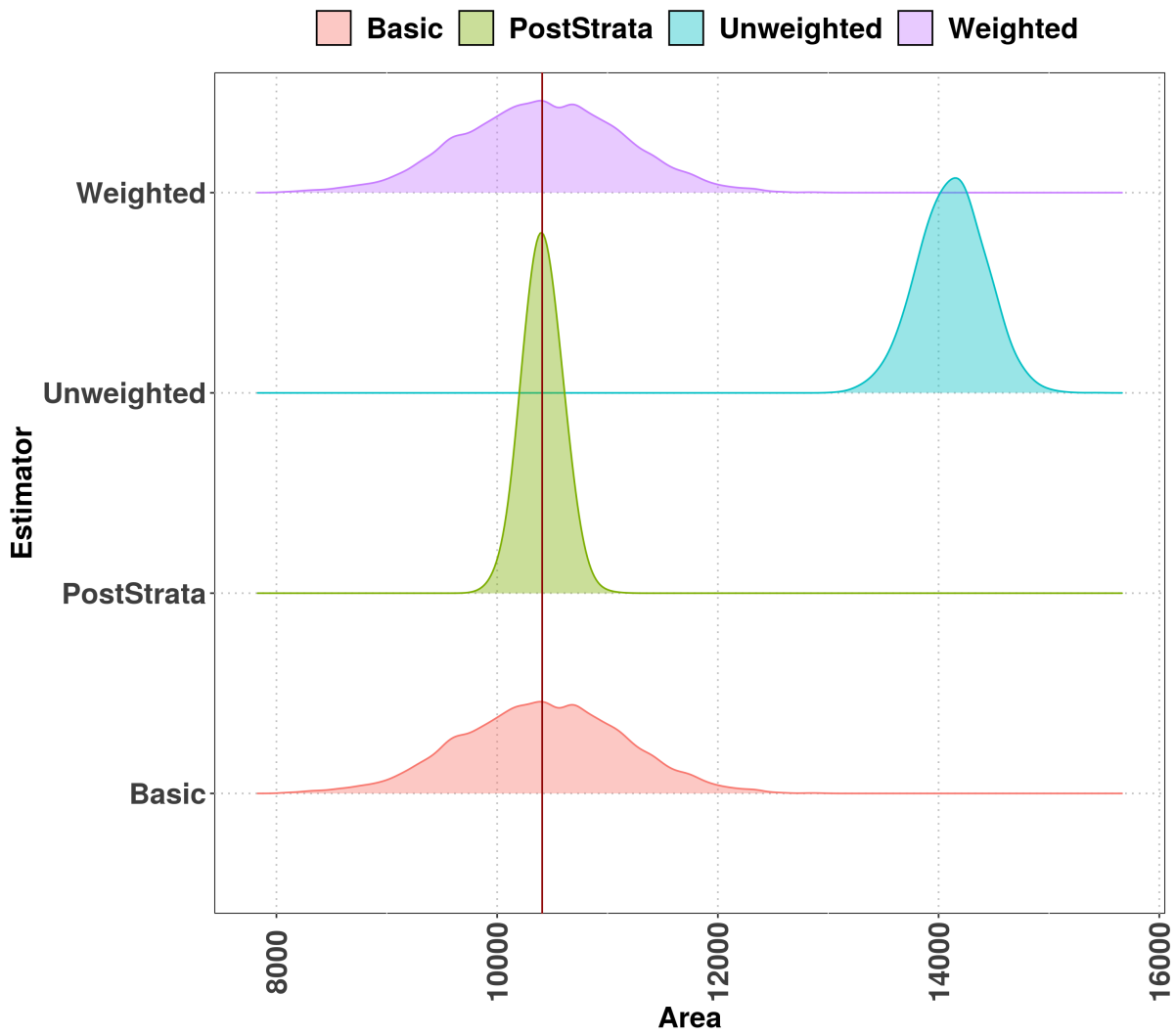


FIGURE 3. Empririal density plots of Estimators

use the post-stratified estimator to cope with misclassification of strata. However, if such information is not available, the best way to proceed is to choose the basic estimator. This means keeping the original strata design when producing estimates is the best way to prevent introducing bias, when there is error in strata formation.

REFERENCES

ABREU DA et al. (2010). Using the Census of Agriculture list frame to assess misclassification in the June Area Survey. *Proceedings of the Joint Statistical Meetings*, p. s/n.

BORYAN CG et al. (2014). A New Automatic Stratification Method for U.S. Agricultural Area Sampling Frame Construction Based on the Cropland Data Layer. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7:4317–4327.

FAO (2018). Handbook on remote sensing for agricultural statistics. Global Strategy to improve Agricultural and Rural Statistics (GSARS).

FERRAZ C, DELINCÉ J, & GALLEGO J (2018). Agricultural Master Sampling Frames: Lessons learned from international field experiments and case studies. Technical Report No. 39GSARS Technical Report: Rome. Tech. rep. FAO.

GAO B et al. (2017). Additional Sampling Layout Optimization Method for Environmental Quality Grade Classifications of Farmland Soil. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10:5350–5358.

HOWE J (2006). The rise of crowdsourcing. *Wired magazine* 14:1–4.

JANG D et al. (2009). Effects of misclassification of race/ethnicity categories in sampling stratification on survey estimates. *Proceedings of the American Statistical Association, Survey Methods Section*. American Statistical Association Alexandria, VA, p. 3414–28.

LAMAS AC et al. (2010). Modeling misclassification in the June Area Survey. *Proceedings of the section on survey research methods JSM*, p. s/n.

LASO BAYAS JC et al. (2016). Crowdsourcing in-situ data on land cover and land use using gamification and mobile technology. *Remote Sensing* 8:905.

MULROW J & WOODBURN L (1990). An Investigation of Stratification Errors. American Statistical Association 1990 Proceedings of the Section on Survey Research Methods: ASA.

SARALIOGLU E & GUNGOR O (2019). Use of crowdsourcing in evaluating post-classification accuracy. *European Journal of Remote Sensing* 52:137–147.

SEE L et al. (2016). Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information* 5:55.