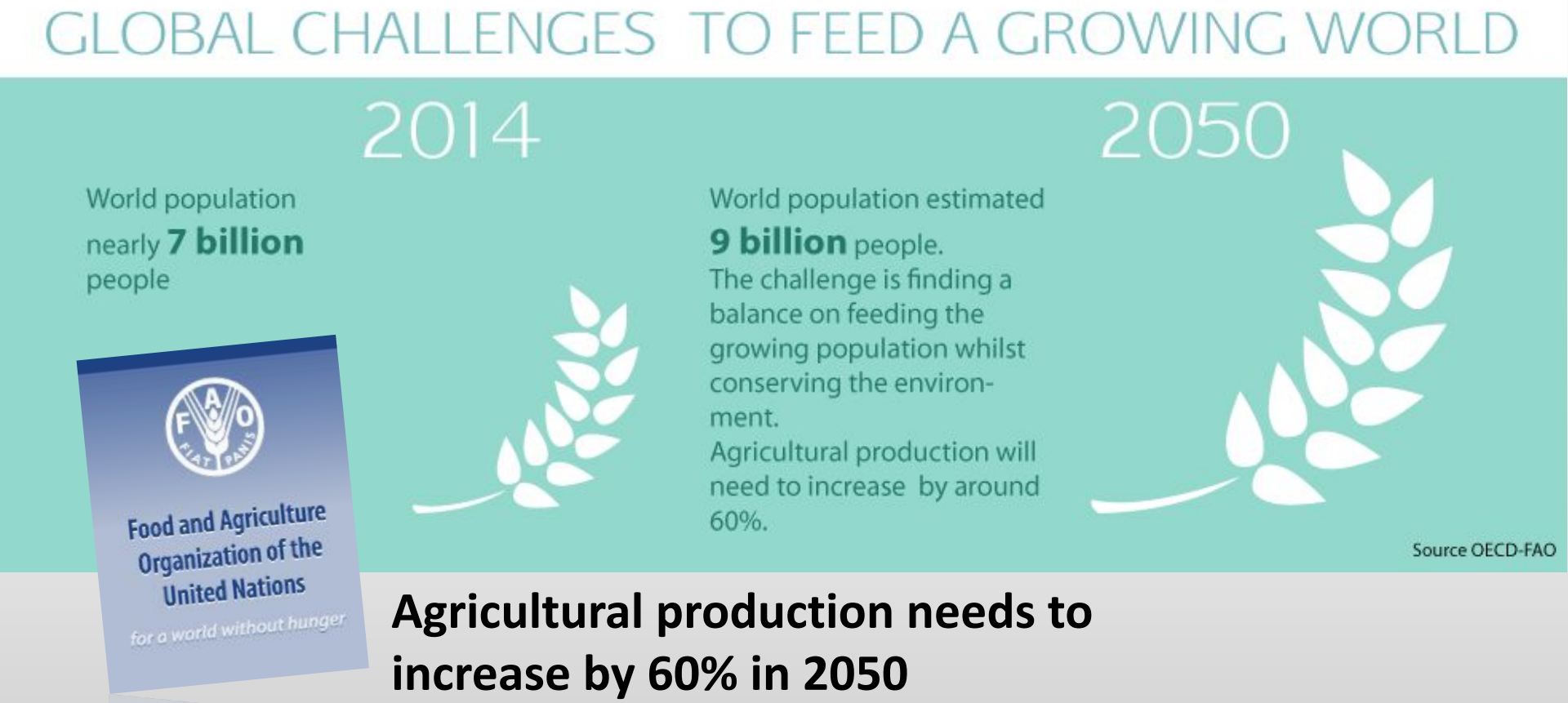
Strata Misclassification in Area Sampling of Square Segments

Raydonal Ospina

Why agricultural statistics?



Agricultural production needs to increase by 60% in 2050

IMPROVING AGRICULTURAL & RURAL STATISTICS Global Strategy

Food and Agriculture Organization of the United Nations

The Global Strategy for Improving Agricultural and Rural Statistics

Three pillars:

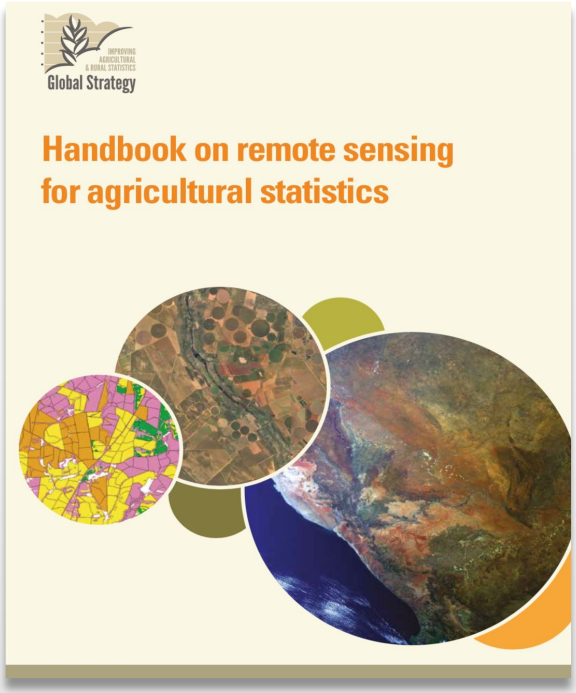
I. **Establishing a minimum set of core data;**
II. **Integrating agriculture into National Statistical Systems; and**
III. **Fostering the sustainability of the statistical system through governance and statistical capacity building.**

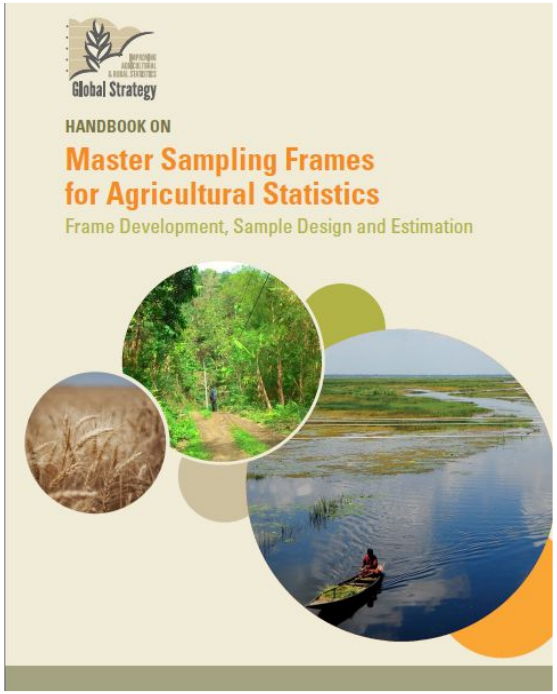
IMPROVING AGRICULTURAL & RURAL STATISTICS
Global Strategy

Food and Agriculture Organization of the United Nations

IMPROVING AGRICULTURAL & RURAL STATISTICS
Global Strategy

www.gsars.org





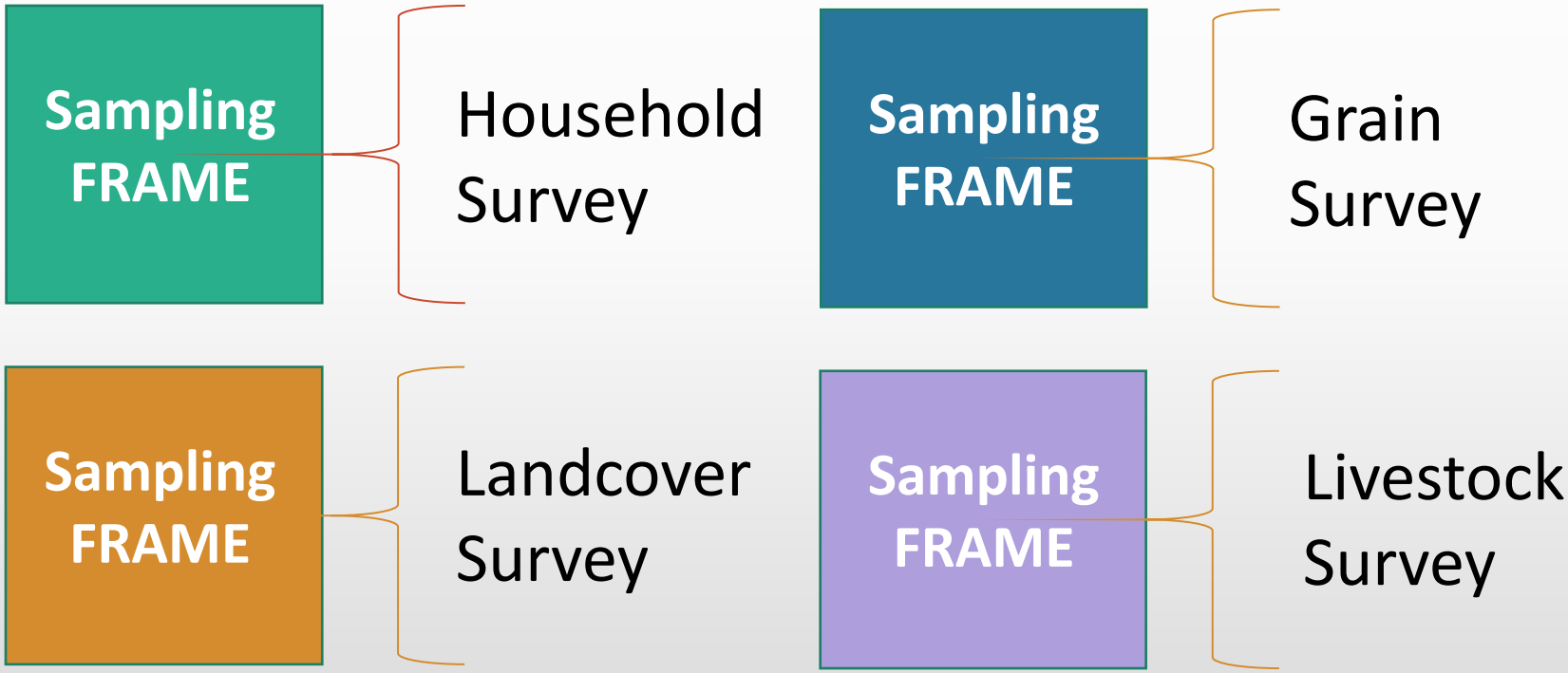
Download a free copy of the Handbooks at:

http://gsars.org/wp-content/uploads/2016/02/MSF-010216-web.pdf
http://gsars.org/en/handbook-on-remote-sensing-for-agricultural-statistics/

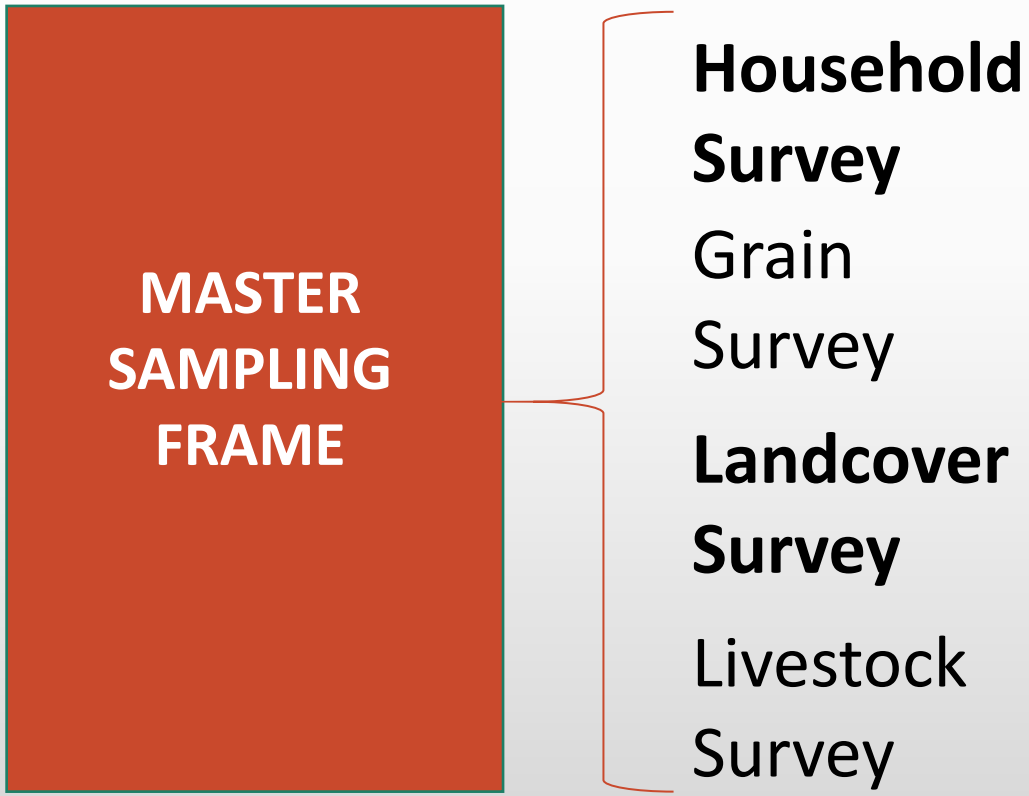
In this presentation

- Area frames in agriculture
- FAO experiments in Brazil and Strata Misclassification

Sampling Frames



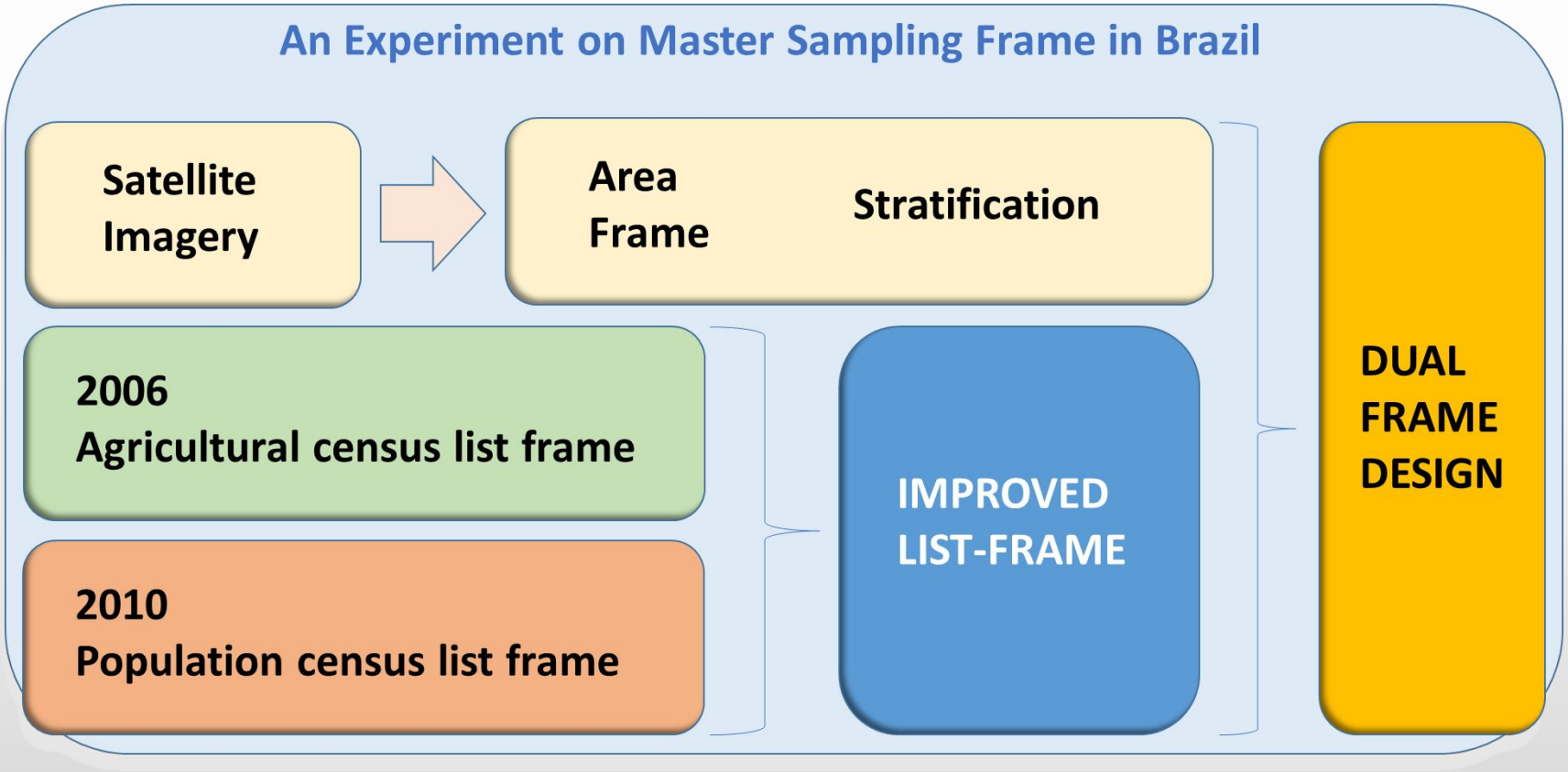
Master Sampling Frame



Agricultural Area Frames

- Sampling units can assume a variety of forms
- Built upon GPS/GIS/Remote sensing type of data
- Furnishes “complete” population coverage
- Usually provides indirect access to reporting units
- Keeps updated over time
- Needs maintenance for stratification purposes
- Have higher cost to reach a sampling unit

FAO Experiment in Brazil



Practical Methods for Area Surveys

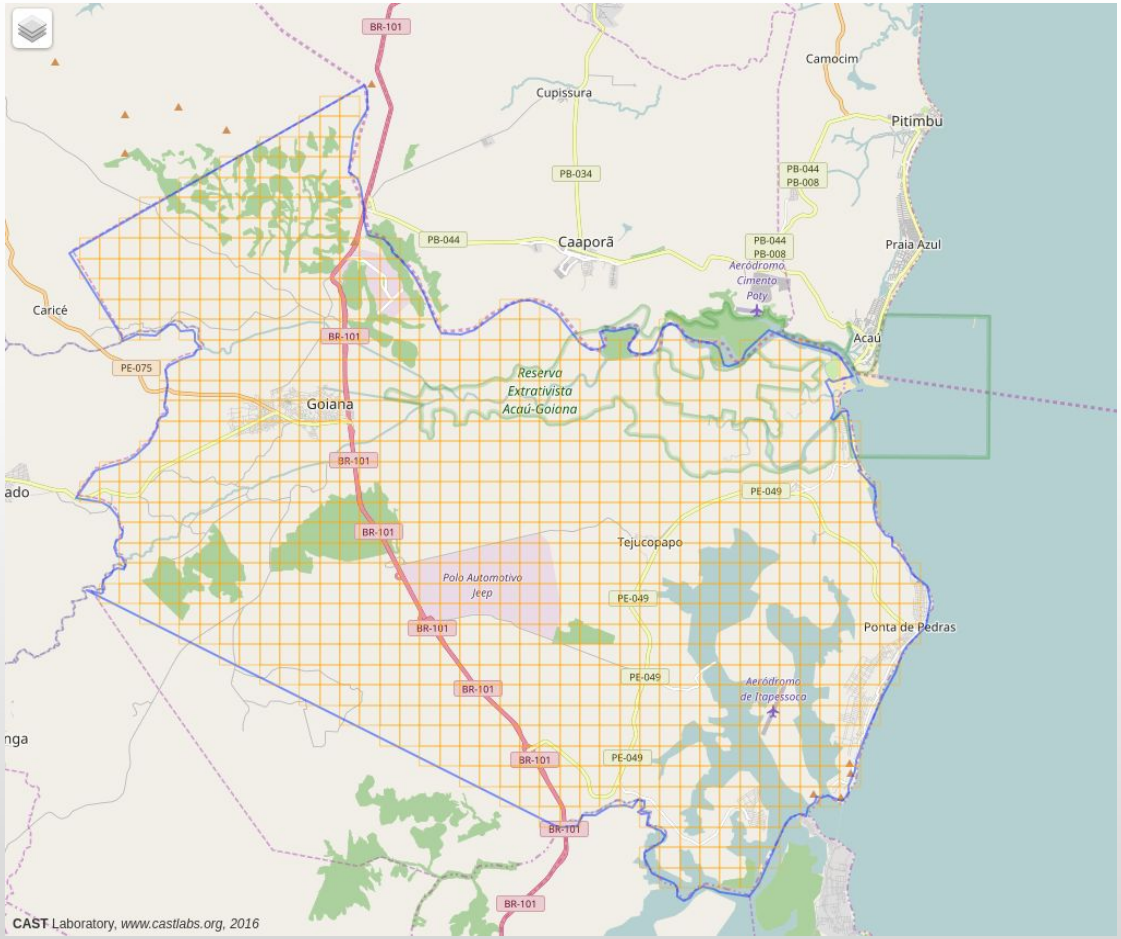
- Area frame of square segments
- Use points within squares to foster stratification
- Google imagery resources available for free
- The land cover in each point was assessed, using crowdsourced classification
- The major advantage of this method relies on its low cost of implementation in terms of budget and timing to achieve stratification of the full territory.
- From points, go to segment stratification

Practical Methods Area Survey

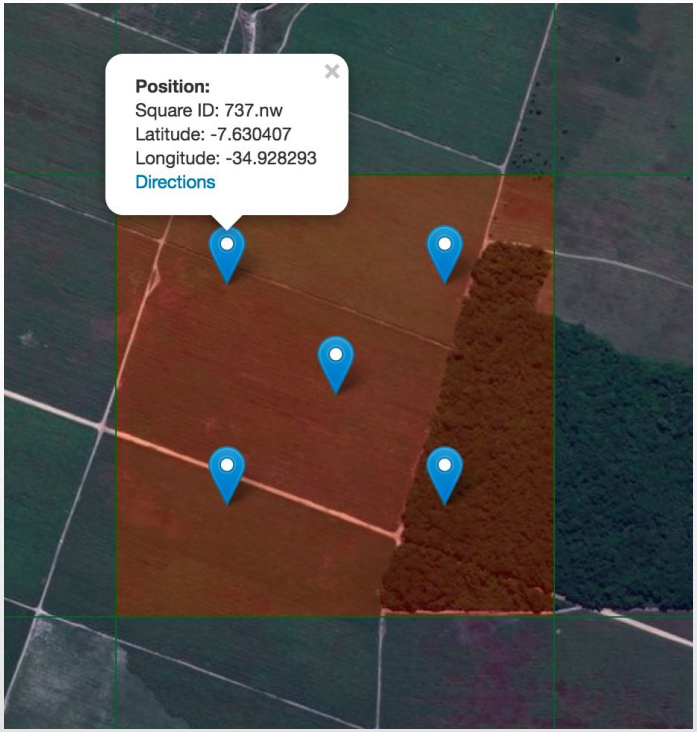
Goiana-PE

922

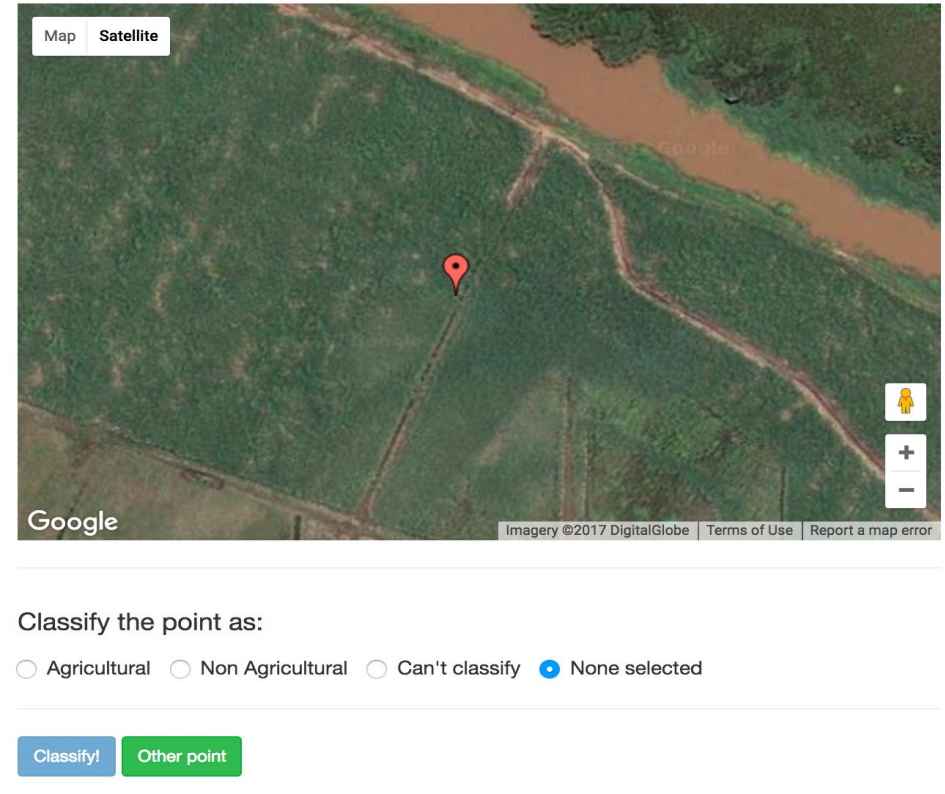
segments



Practical Methods Area Survey



Pattern of points in a segment



Crowdsource screen for Photo interpretation

Crowd-sourcing: from points to segment classification

Segment strata definition

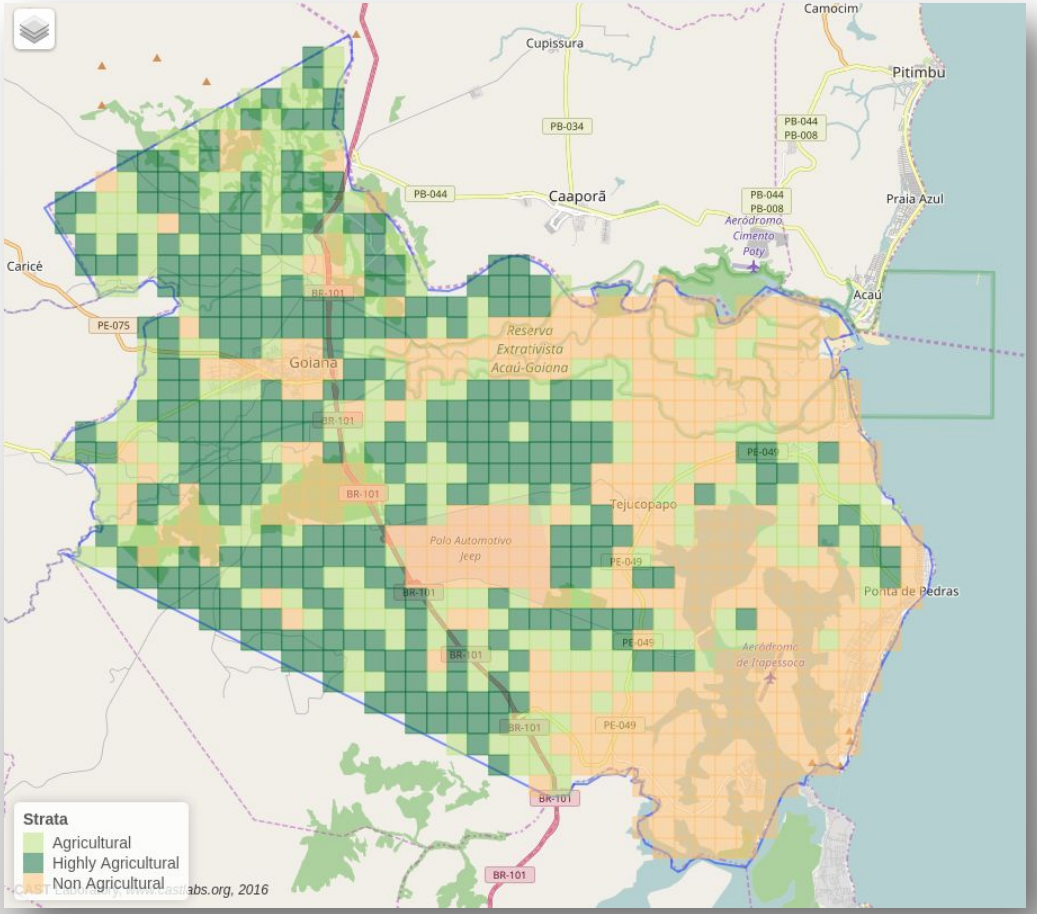
Strata	Description
1. Highly cropland	Segments with 4 or 5 points classified as “Cropland”
2. Cropland	Segments with 2 or 3 points classified as “Cropland”
3. Non-cropland	Segments with at most 1 point classified as “Cropland”

Area frame stratification

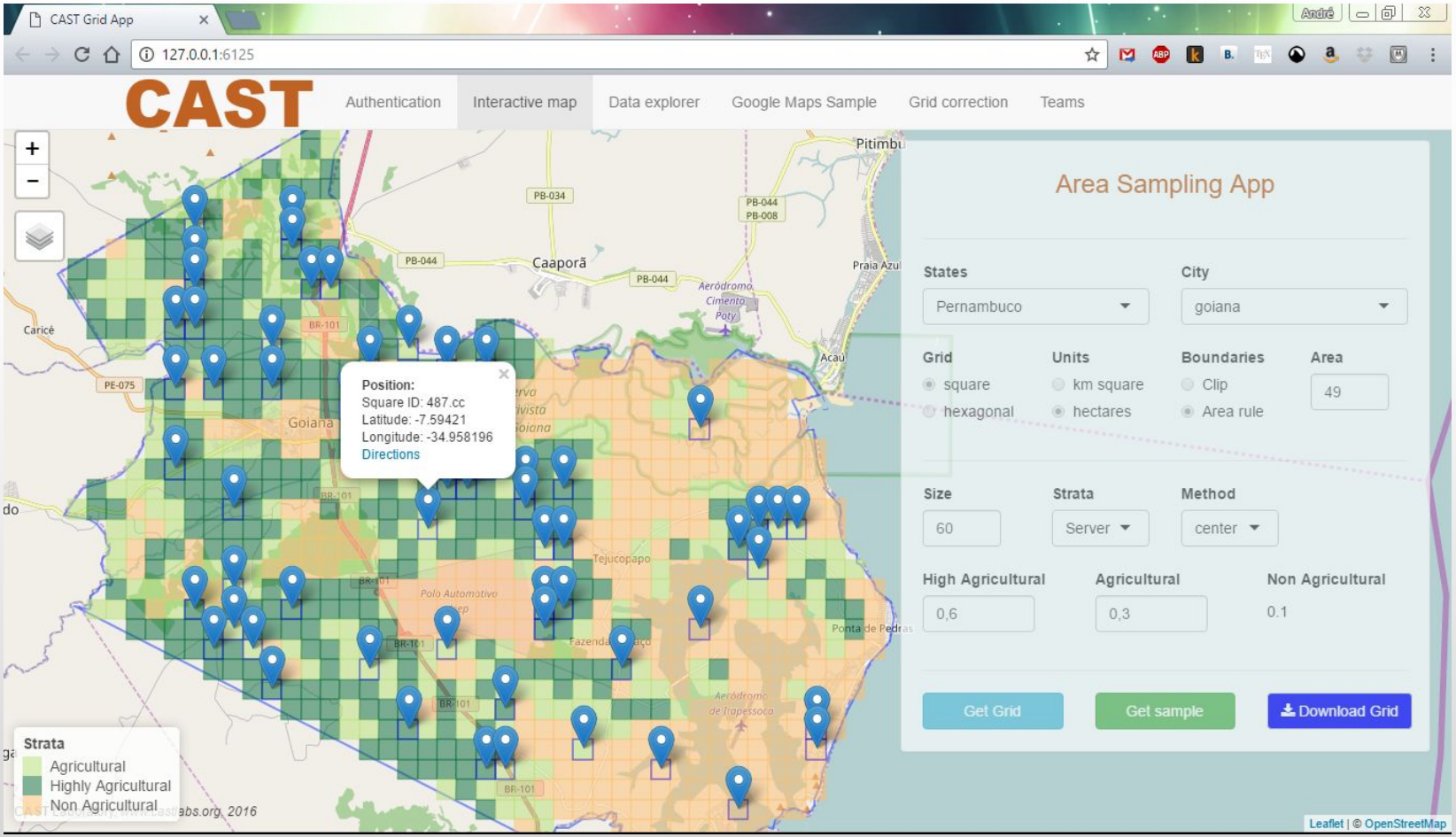
Goiana-PE

922

segments



Area frame sample Goiana



Area frame stratification in Goiana

Original square segments classification in strata by the time the sample was selected

Strata	Frame	Sample
1. Highly cropland	314	42
2. Cropland	232	15
3. Non-cropland	376	3

Goiana strata misclassification analysis

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

Correct estimators

The Horvitz-Thompson estimator for the total production with a specific crop c

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

- h is the index identifying the original stratum (based on photo interpretation);
- w_k is the sample weight for segment k;
- y_k is the production with crop c in segment k;
- S_h is the set of all segments originally selected from stratum h in the sample.

Illustrating the relationship between the misclassified sample cells

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.

If simple random sampling is used in each stratum, and no misclassification occur, then the design strata is the same as the actual strata. In such case, $N_{j+}=N_{+j}=N_{jj}$ and so, $w_k=\frac{N_{+j}}{n_{+j}}$, where:

- N_{+j} is the number of segments in the area frame that belong to stratum j;
- n_{+j} is the number of segments in the sample, selected from stratum j;

Design strata	Actual strata			Total
	1	2	3	
1	N_{11}	N_{12}	N_{13}	N_{1+}
2	N_{21}	N_{22}	N_{23}	N_{2+}
3	N_{31}	N_{32}	N_{33}	N_{3+}
Total	N_{+1}	N_{+2}	N_{+3}	N_{++}

population

Design strata	Actual strata			Total
	1	2	3	
1	n_{11}	n_{12}	n_{13}	n_{1+}
2	n_{21}	n_{22}	n_{23}	n_{2+}
3	n_{31}	n_{32}	n_{33}	n_{3+}
Total	n_{+1}	n_{+2}	n_{+3}	n_{++}

sample

Estimators

Weights

- **Basic Estimator:** Ignore the problem of errors in stratification

$$w_k = \frac{N_{h+}}{n_{h+}}.$$

- **Unweighted estimator:** The unweighted estimator corresponds to proceed corrections to the basic estimator using only the true observation in the sample.

$$w_k = \frac{\widehat{N_{+1}}}{n_{+1}},$$

- **Weighted estimator:** The weighted estimator use sample data to correct information regarding observed and non-observed segments. (example, stratum 1)

$$\widehat{N}_{+1} = N_{11}p_{11} + N_{21}p_{21} + N_{31}p_{31}$$

$$n_{+1} = n_{11} + n_{21} + n_{31}$$

$$p_{11} = n_{11}/n_{1+}; \quad p_{21} = n_{21}/n_{1+};$$

$$p_{31} = n_{31}/n_{1+}.$$

- **Post-Stratified estimator:** The post-stratified estimator uses the most updated information for both, the sample and the population sizes.

$$w_k = \frac{N_{+j}}{n_{+j}}.$$

- **Expansion estimator:** For a segment k in stratum 1, the expansion estimator uses

$$w_k = \frac{N}{n}\left(p_{11} + p_{21} + p_{31}\right).$$

Simulation

- We carried out a Monte Carlo simulation to evaluate the performance of the five different estimators for the total of sugarcane production (given in Tons) from Goaina.
- An artificial population of 922 segments was built, keeping the strata population sizes of Goiana 314, 232 e 376, for stratum 1, 2 and 3 respectively.
- The observations and respective misclassifications were generated according to the models described next for each stratum.

Stratum 1 (Highly-Cropland).

314 observations were generating according to model ξ_1

$$\xi_1 : y_k = \mu_1 - \mu_2 I_{12} - \mu_3 I_{13}(1 - I_{12}) + \varepsilon_k$$

Stratum 2 (Cropland).

232 observations were generating according to model ξ_2

$$\xi_2 : y_k = \mu_3 + (\mu_3 - \mu_1) I_{21} - (\mu_3 - \mu_2) I_{23}(1 - I_{21}) + \varepsilon_k$$

Stratum 3 (Non-Cropland).

376 observations were generating according to model ξ_3

$$\xi_3 : y_k = \mu_2 + (\mu_3 - \mu_2) I_{32} - (\mu_1 - \mu_2) I_{31}(1 - I_{32}) + \varepsilon_k$$

Simulation

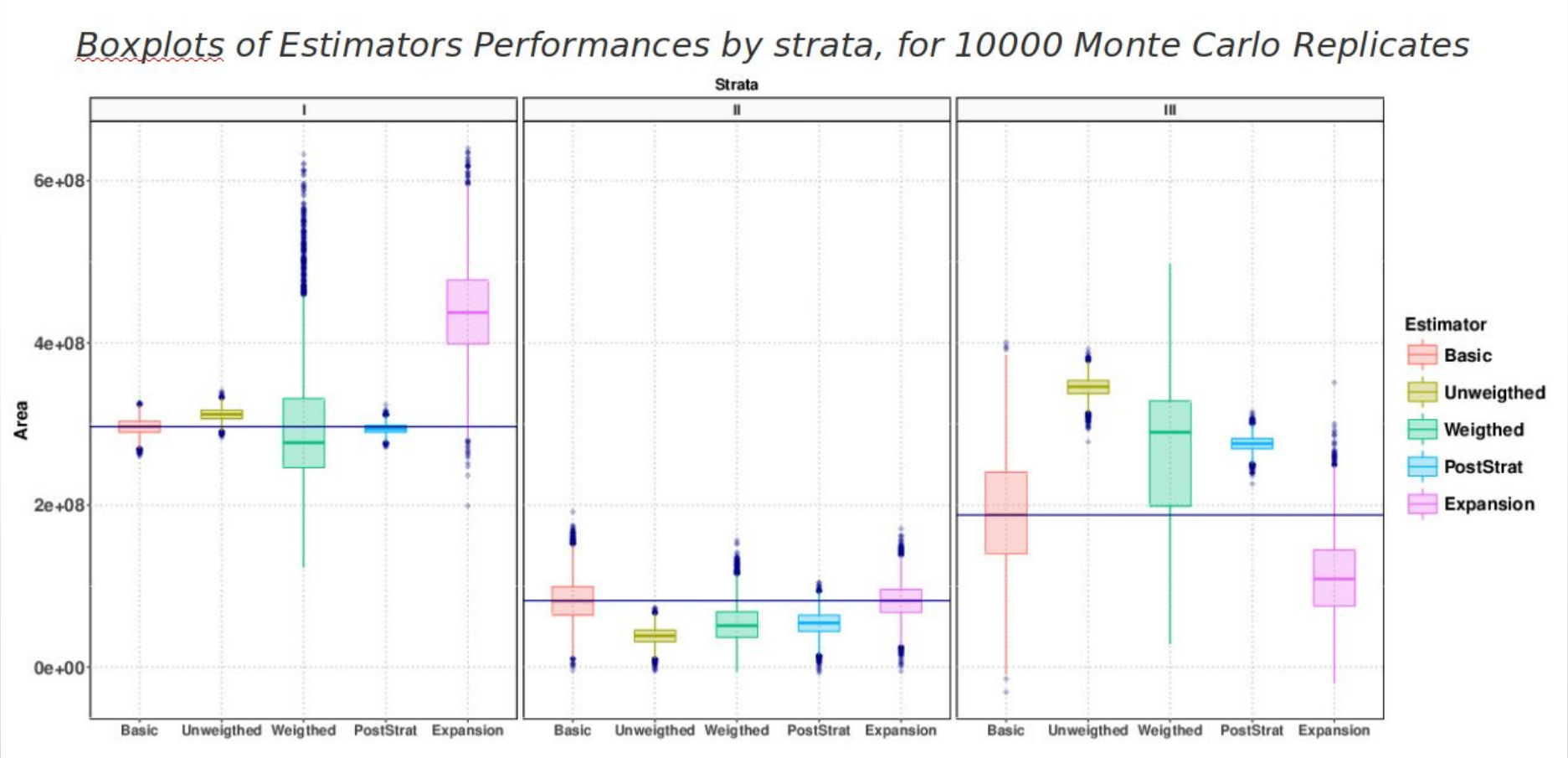
- I_{ij} is an indicator variable that segment k was originally sampled from stratum i, but reclassified, in the field, in stratum j ;
- $I_{ij} \sim Bernoulli(p_{ij});$
- $\varepsilon_k \sim N(0, \sigma_j^2)$;
- y_k is the production with a given crop, observed in segment k sampled from stratum j.

Under this scenario, the true production values in Tons. of the stratum population totals, $Y_h = \sum_{k \in U_h} y_k$, are: 296376416 (Stratum I), 81922366 (Stratum II), and 187677928 (Stratum III), leading to a population total of $Y = \sum_{h=1}^{3} Y_h = 565976710.$

Results

small sample size

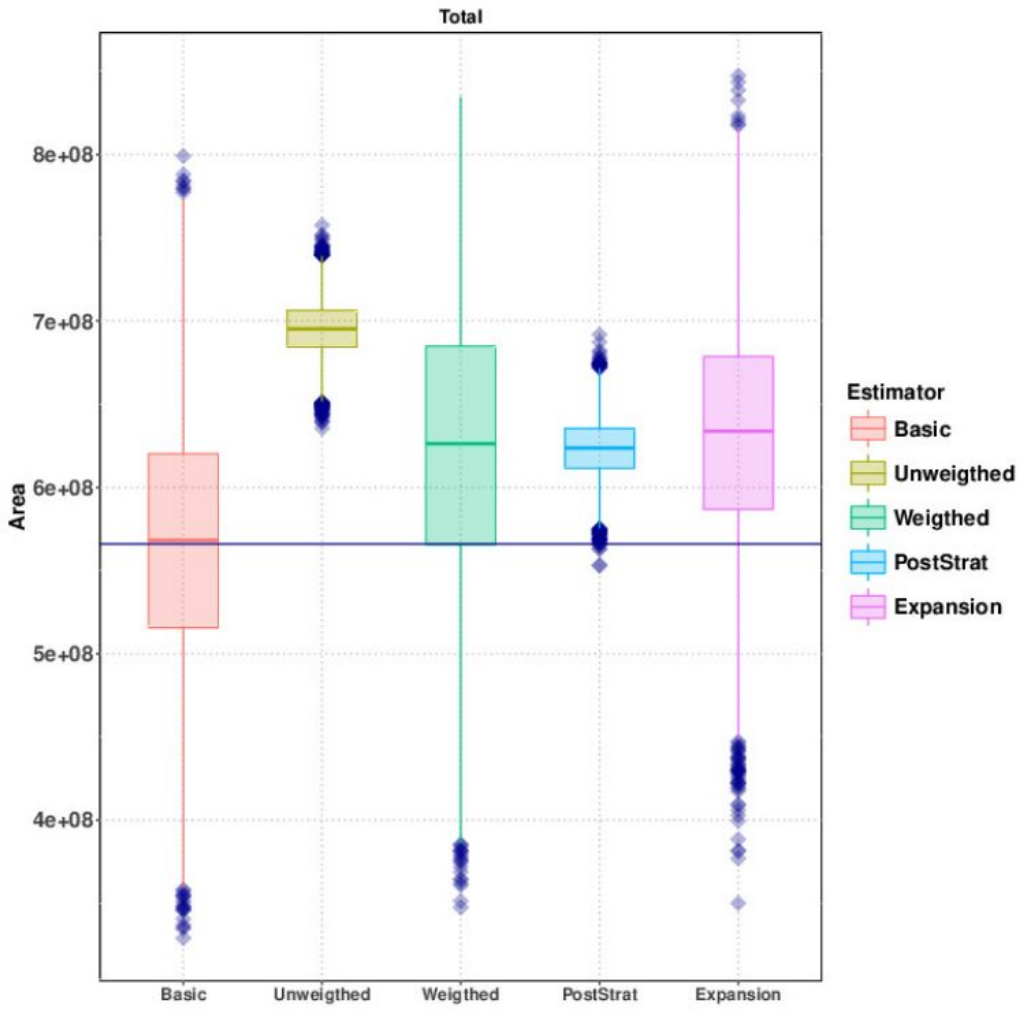
The strata sample sizes of Goiana are: 42, 15 and 3, for stratum 1, 2 and 3 respectively.



Results

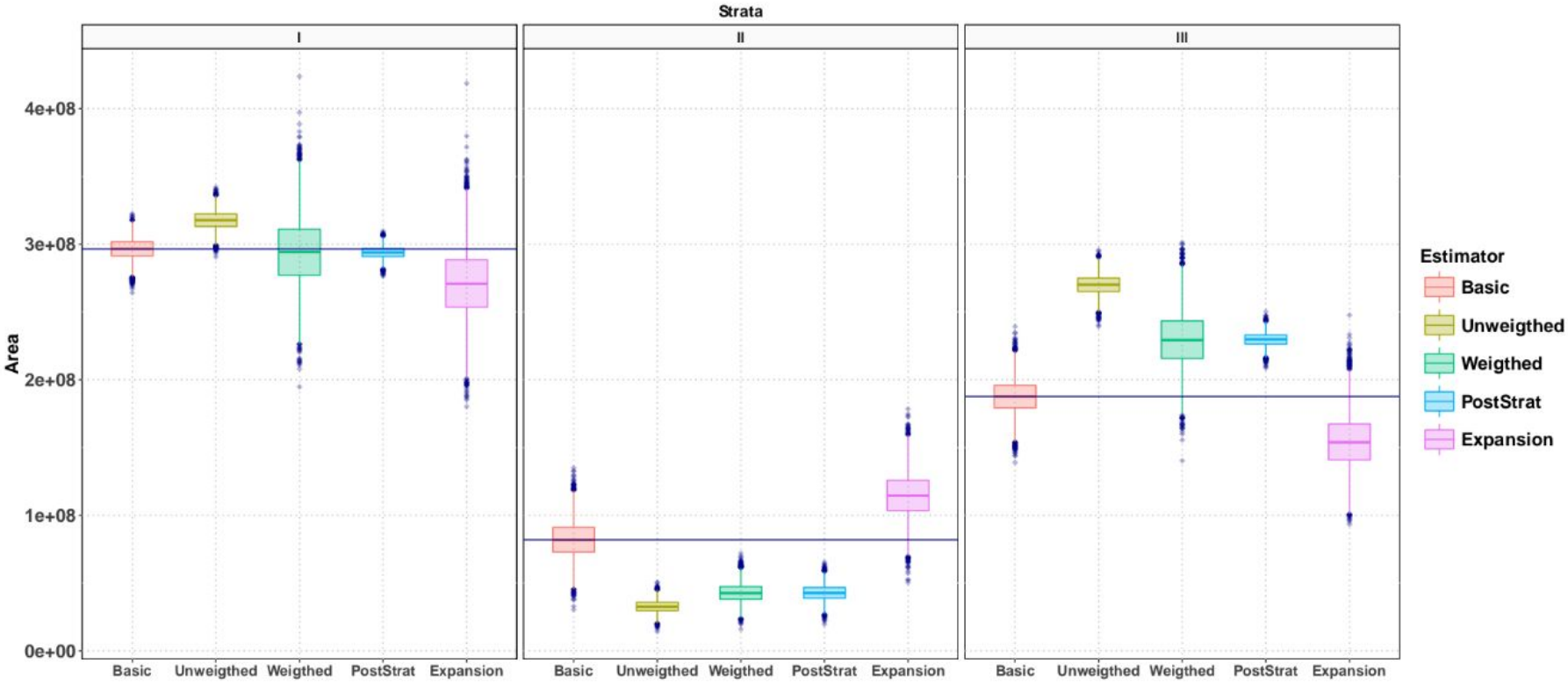
small sample size

Boxplot of Estimators Performances for 10000 Monte Carlo Replicates



Results

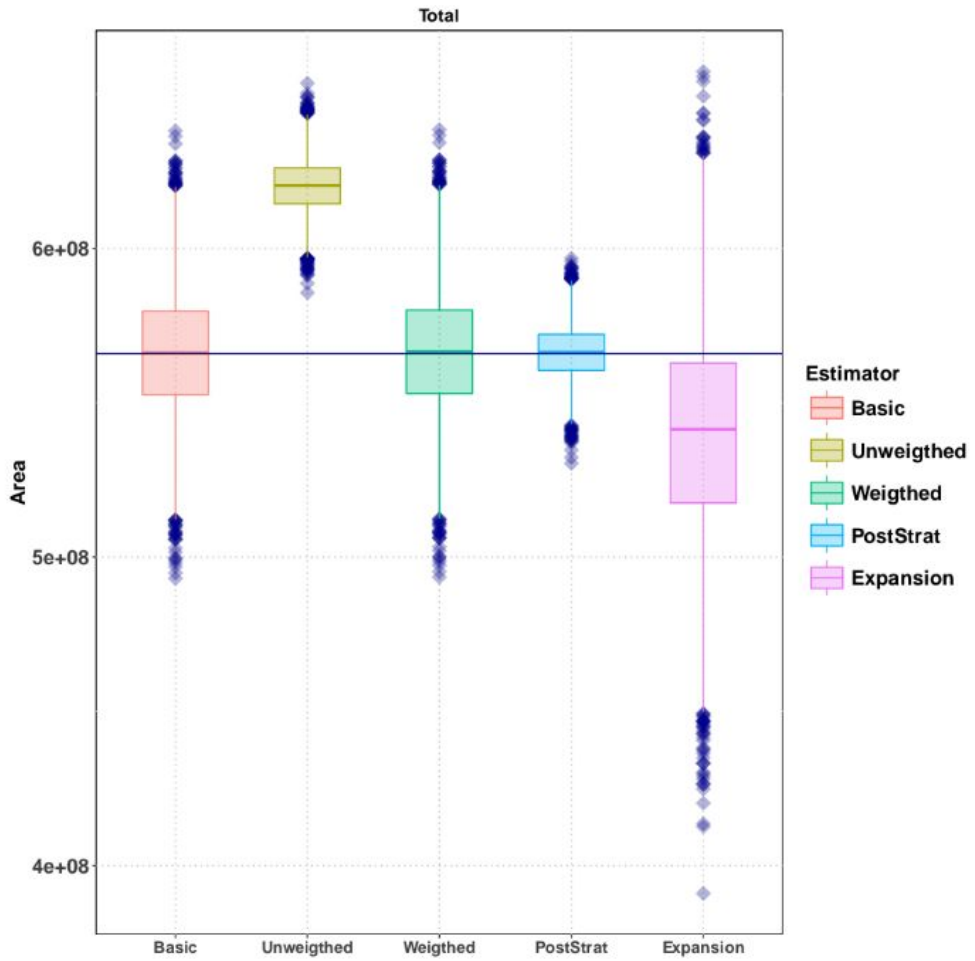
large sample size



Results

large sample size

Boxplots of Estimators Performances, for 10000 Monte Carlo Replicates
Sample size: 180



Concluding remarks

- The post-stratified estimator shows better performance than the others cases, as expected. However, no information regarding the actual size of each stratum is available in practice.

- The performances of the basic and the weighted estimators are similar for producing estimates to the whole population. However, inside each stratum, the weighted estimators is subjected to bias that can be non negligible.

- The unweighted estimators has shown a poor performance in all cases.

- The estimators behavior is affect by the percentage of misclassification.

- In all cases here, the percentages were kept the same as in the Goiana's experiment.

- Considering only the information collected and the experiments we would conclude that keeping the basic estimator as the choice in case of area sampling misclassification of square segments is a safe way to proceed.