Theoretical and practical aspects of using satellite observations for crop area estimation

S. Skakun

University of Maryland, College Park MD, USA

Problem statement: Area estimation

- Maps imperfect
  - **Wall-to-wall** coverage: information on the population (though inaccurate)
  - Omission and commission **errors**
  - Direct use: a ***pixel counting estimator*** is biased [*Gallego, 2004*; *Gallego et al., 2010*; *McRoberts, 2011*]

- Statistical frameworks
  - **Sampling-based** approach and a **design-based inference** framework [*Olofsson et al., 2014*]
    - Maps used for strata
    - Map **inaccuracies will not impact the bias**: depends on the sampling design and estimator
    - Map **inaccuracies will impact the efficiency of stratification**
    - **Model-assisted** estimation to further improve the precision of the estimator [*Carfagna & Gallego, 2005*; *Gallego et al., 2010*; *McRoberts et al., 2024*]

  - Alternative to the design-based inference framework
    - **Model-based** inference [*Ståhl et al., 2016*; *McRoberts et al., 2022*]
    - Requires specification of an adequate model that will impact the bias and precision of the estimator

Efficiency of stratification

- **Assumptions**
  - Sampling design: **stratified random sampling**
  - Map's **classes** used as **strata**
  - Two maps: *A* and *B* used for stratification when estimating an area proportion $\hat{y}$

- **Efficiency** of stratification [*Gallego, 2007*]

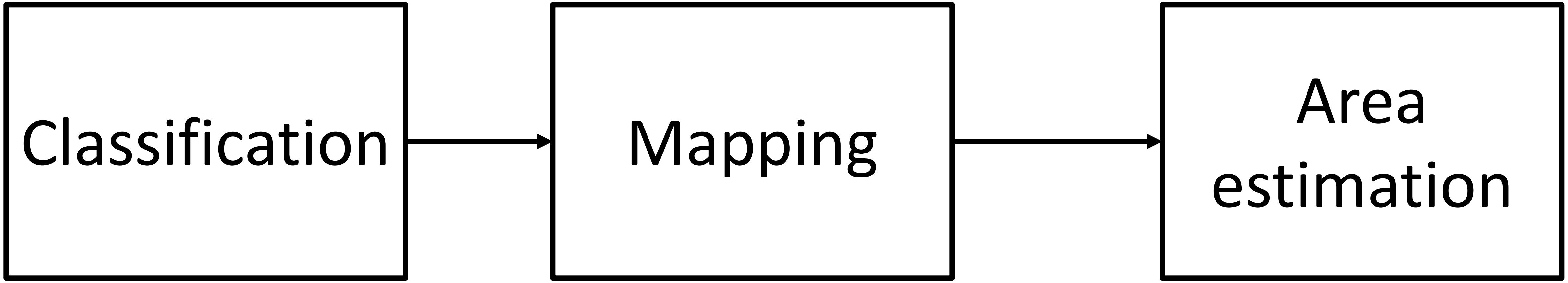
$$\eta_{A,B} = \frac{var(\hat{y}_B) \times n_B}{var(\hat{y}_A) \times n_A}$$

n_B and n_A are sample size, and $var(\hat{y}_B)$ and $var(\hat{y}_A)$ are estimated variances of $\hat{y}$

- **Hypothesis**: a more accurate map will be more efficient

- **Question**: how much?

Land cover / land use mapping and area estimation



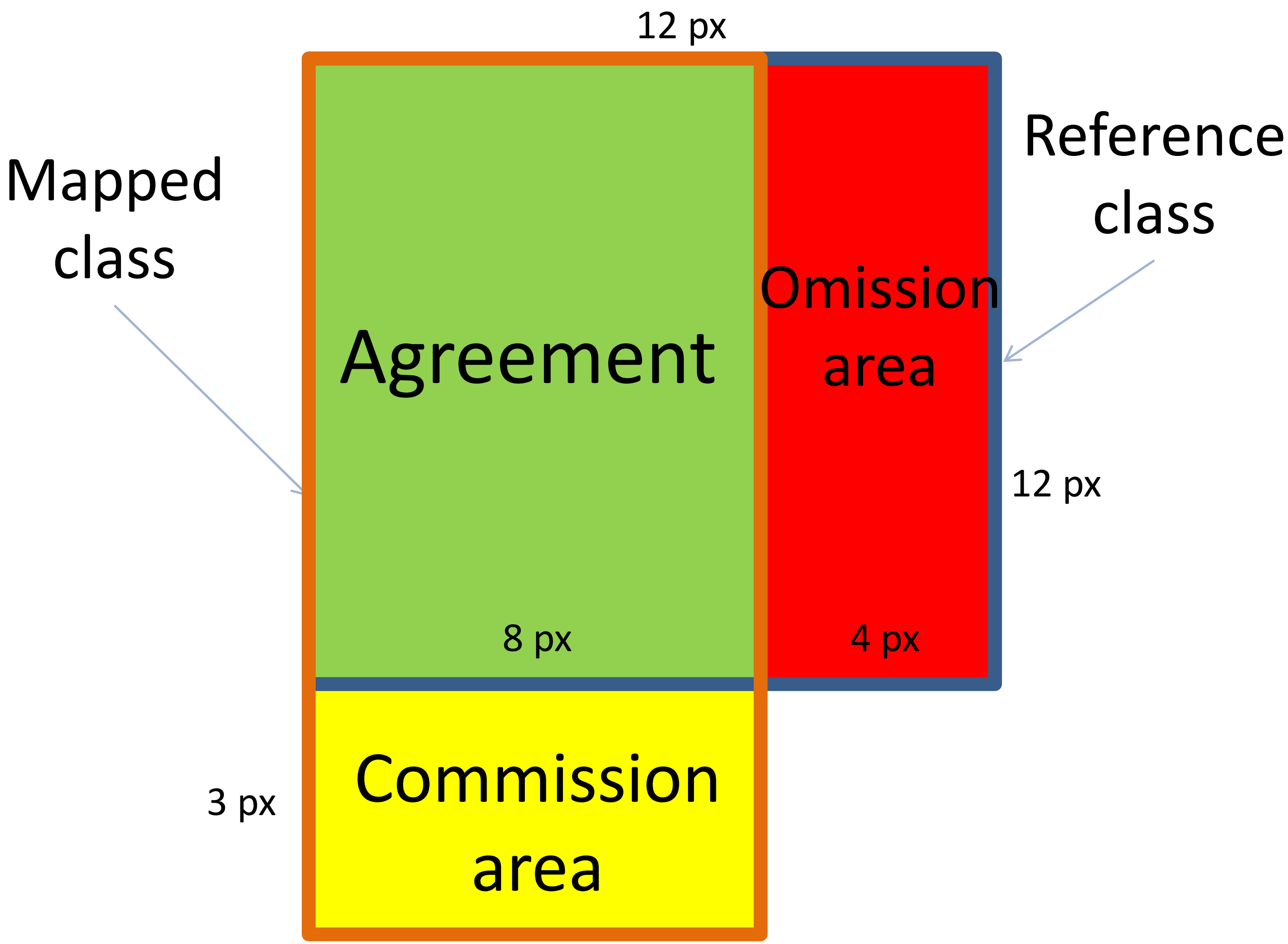
- Data
- Features
- Algorithms

- Spatial context

- Unbiased estimators with uncertainties

Land cover / land use mapping and area estimation

- Pixel counting is a biased estimator



Reference area:
12 x 12 = 144 px

Mapped area:
8 x 15 = 120 px (bias -17%)

PA = 8 x 12 / (12 x 12) = 66.7%

UA = 8 x 12 / (8 x 15) = 80%

Rel. bias = PA/UA - 1

Land cover / land use mapping and area estimation

Remote Sensing of Environment 148 (2014) 42–57

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Review

Good practices for estimating area and assessing accuracy of land change

Pontus Olofsson [a,*], Giles M. Foody [b], Martin Herold [c], Stephen V. Stehman [d],
Curtis E. Woodcock [a], Michael A. Wulder [e]

General steps

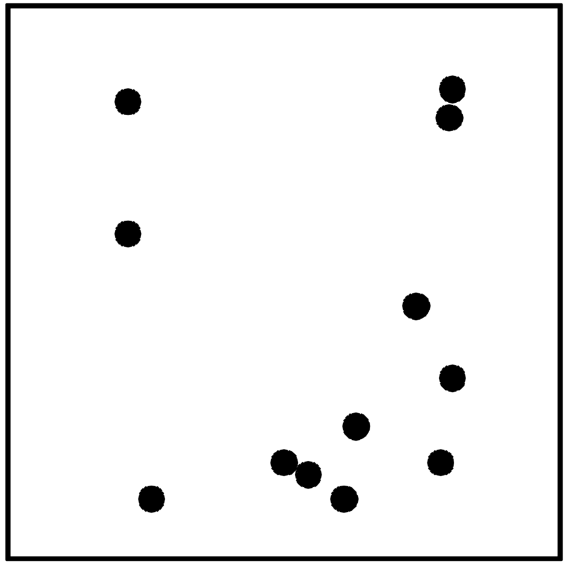
- **Sampling design**
  – The sampling design is the protocol for selecting the **subset of spatial units** (e.g., pixels or polygons) that will form the **basis** of the **accuracy assessment**

- **Response design**
  – Encompasses all aspects of the protocol that lead to determining whether the **map** and **reference** classifications are **in agreement**

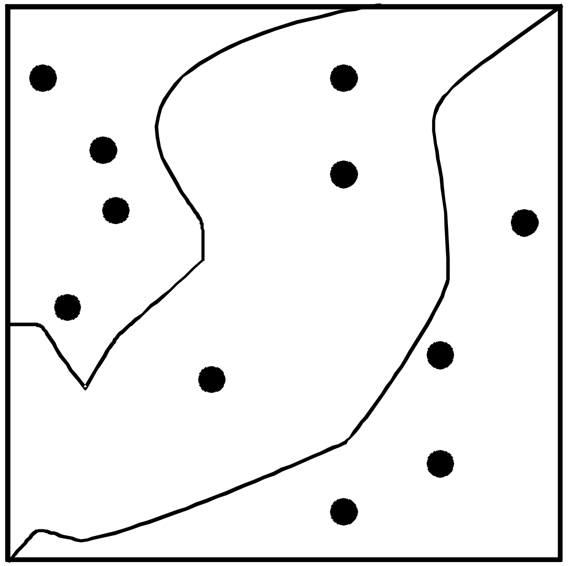
- **Analysis**
  – Includes protocols for defining how to quantify **accuracy** along with the formulas and inference framework for estimating **accuracy** and area and quantifying **uncertainty** of these estimates

Sampling design

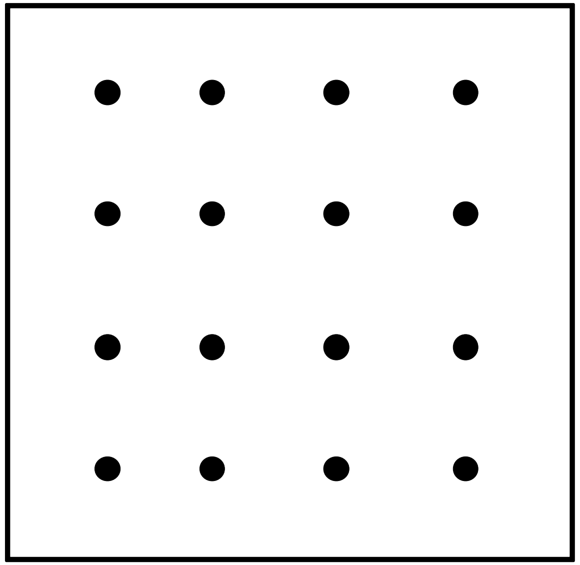
(i)



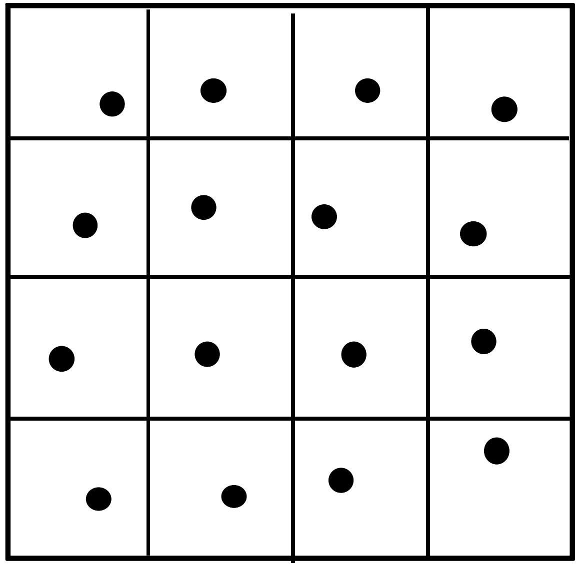
(ii)



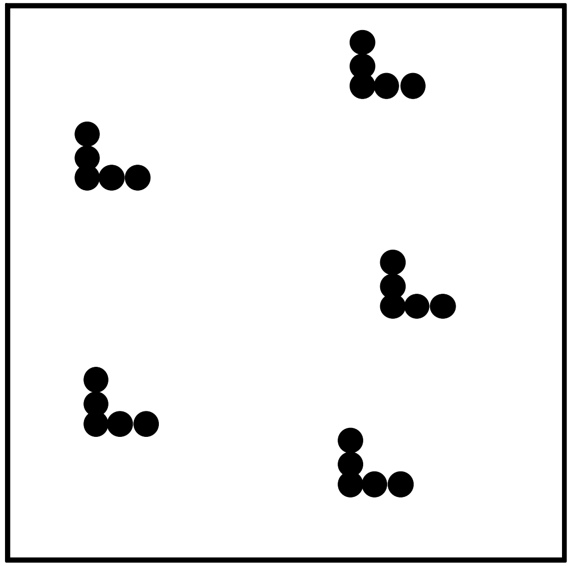
(iii)



(iv)



(v)



(i) Random
(ii) Stratified random
(iii) Systematic
(iv) Unaligned systematic
(v) Cluster

Error matrix

- Population matrix (in terms of **area proportions**)

		Reference				
		Class 1	Class 2	…	Class K	Total
Map	Class 1	f_{11}	f_{12}	…	f_{1K}	$f_{1\cdot}$
	Class 2	f_{21}	f_{22}	…	f_{2K}	$f_{2\cdot}$
	…	…	…	…	…	…
	Class k	f_{K1}	f_{K2}	…	f_{KK}	$f_{K\cdot}$
	Total	$f_{\cdot 1}$	$f_{\cdot 2}$	…	$f_{\cdot K}$	1

$f_{k\cdot}$ is the **mapped** area proportion

$f_{\cdot k}$ is the **reference** area proportion

Producer's accuracy (PA): $v_k = f_{kk} / f_{\cdot k}$ Omission error (OE): $\psi_k = (1 - v_k)$

User's accuracy (UA): $u_k = f_{kk} / f_{k\cdot}$ Commission error (CE): $\phi_k = (1 - u_k)$

Error matrix

- Population matrix (in terms of **area proportions**)

		Reference				
		Class 1	Class 2	…	Class K	Total
Map	Class 1	f_{11}	f_{12}	…	f_{1K}	$f_{1\cdot}$
	Class 2	f_{21}	f_{22}	…	f_{2K}	$f_{2\cdot}$
	…	…	…	…	…	…
	Class k	f_{K1}	f_{K2}	…	f_{KK}	$f_{K\cdot}$
	Total	$f_{\cdot 1}$	$f_{\cdot 2}$	…	$f_{\cdot K}$	1

$\boldsymbol{f_{k\cdot}}$ is the **mapped** area proportion

$\boldsymbol{f_{\cdot k}}$ is the **reference** area proportion

- A **bias of pixel counting estimator**

Gallego et al. (2010): through OE/CE

$$b_k = f_{k\cdot} - f_{\cdot k} = \phi_k f_{k\cdot} - \psi_k f_{\cdot k}$$

this work: through PA/UA

$$b_k = f_{\cdot k} \underbrace{\left({}^{v_k}/_{u_k} - 1\right)}_{\text{Relative bias}}$$

Analysis

- **Elements of the population error matrix should be estimated from <u>sample</u>!**
- Suppose sample-based **<u>estimator</u>** of p_{ij} is $\hat{p}_{ij}$
- The error matrix should be reported in terms of **<u>estimated</u> area proportions**!

- For **equal probability sampling designs** (e.g., **simple random** and **systematic sampling**) and for **stratified random sampling in which the strata correspond to the map classes**,

$$\hat{p}_{ij} = W_i \frac{n_{ij}}{n_{i.}}$$

n_{ij} – number of MMU mapped as class i but reference is class j

$n_{i.}$ – number of MMU mapped as class i

W_i – the proportion of area mapped as class i

Relative efficiency

- Elements of the confusion matrix estimated from the sample

- Sample size [*Olofsson et al., 2014*; *Cochran, 1977, Eq. 5.25*]

$$n = \frac{\left(\sum_{k=1}^{K} w_k s_k\right)^2}{V_{target} + {}^{1}/_{N} \sum_{k=1}^{K} w_k s_k^2} \approx \frac{\left(\sum_{k=1}^{K} w_k s_k\right)^2}{V_{target}}$$

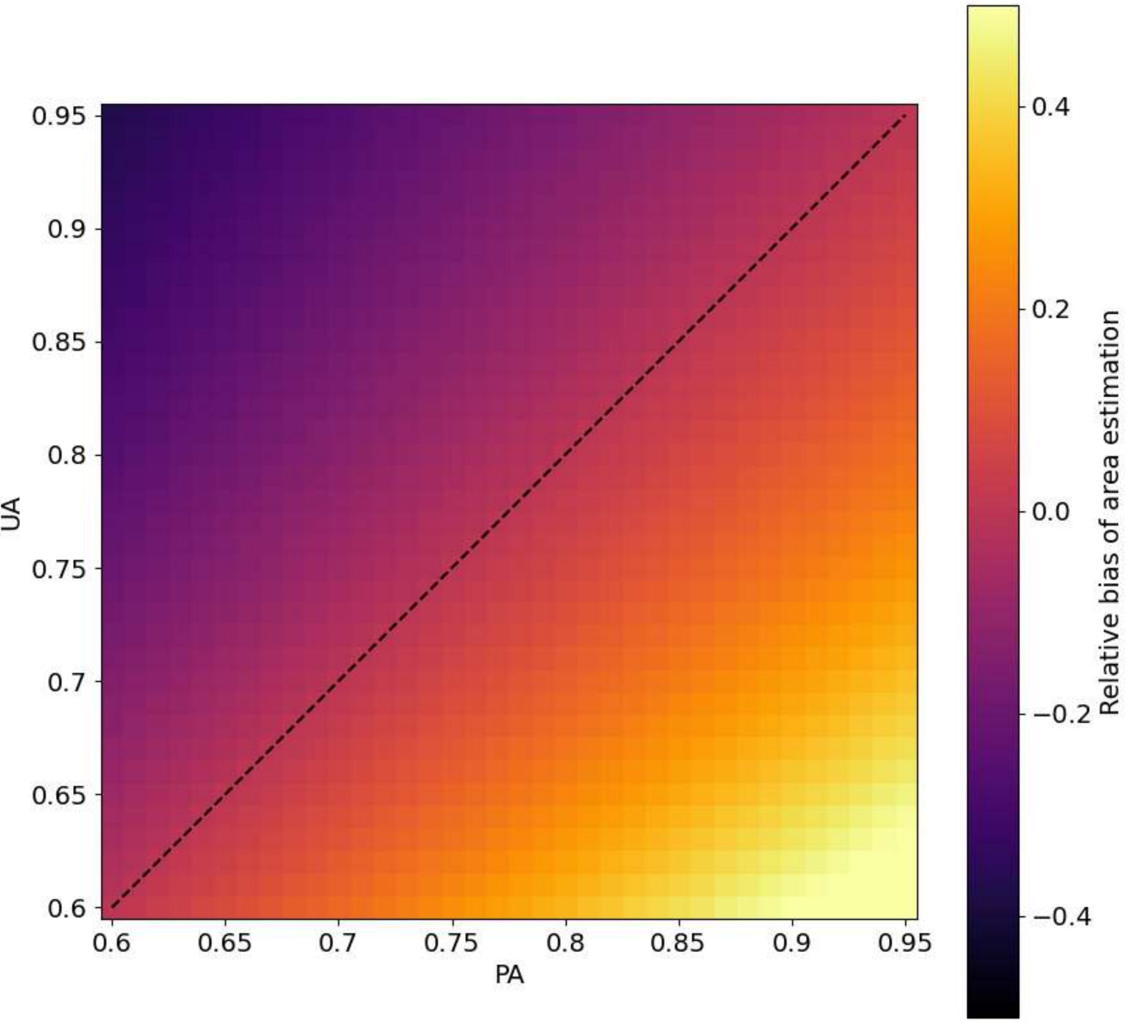
where V_{target} is the target variance, $w_k \equiv f_{k\cdot}$ is the stratum weight, N is population size, and s_k is the standard deviation of stratum k

(for area estimation $s_i(\hat{f}_{\cdot k}) = \sqrt{\frac{n_{ik}}{n_{i\cdot}}\left(1 - \frac{n_{ik}}{n_{i\cdot}}\right)}$)

- Relative efficiency

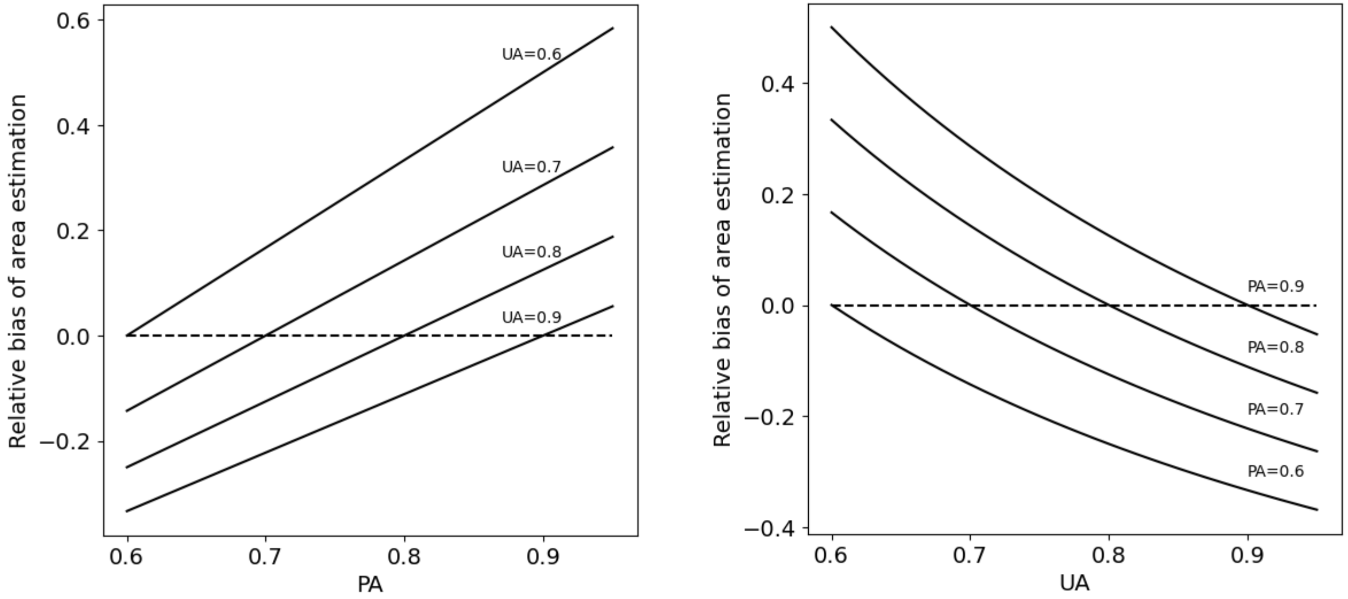
$$\eta_{A,B} = \frac{var(\hat{y}_B) \times n_B}{var(\hat{y}_A) \times n_A} = \frac{\left(\sum_{i=1}^{K} f_{\cdot i} \frac{v_i^B}{u_i^B} s_i^B(\hat{f}_{\cdot k})\right)^2}{\left(\sum_{i=1}^{K} f_{\cdot i} \frac{v_i^A}{u_i^A} s_i^A(\hat{f}_{\cdot k})\right)^2}$$

Pixel counting area estimator



The relative bias of area estimation with pixel counting depending on PA and UA

Relative bias = $\left({}^{PA}/_{UA} - 1\right)$



The relative bias of area estimation with pixel counting under fixed UA (left) and fixed PA (right)

Relative efficiency of stratification

- Two classes *K*=2, true area of the target class $f \equiv f_{\cdot 1}$ =0.15 and target CV=5%.

	A PA=0.9 UA=0.7	B PA=0.7 UA=0.9	C PA=0.8 UA=0.8	D PA=0.6 UA=0.6
	$f = 0.15$			
Overall	0.927	0.943	0.940	0.880
F-score=$\frac{2v_1u_1}{v_1+u_1}$	0.788	0.788	0.800	0.600
Sample size n				1508
Rel. eff. strat.				1.50

Simple random sampling, sample size *n*=2,262

Relative efficiency of stratification

- Two classes *K*=2, true area of the target class $f \equiv f_{.1}$ =0.15 and target CV=5%.

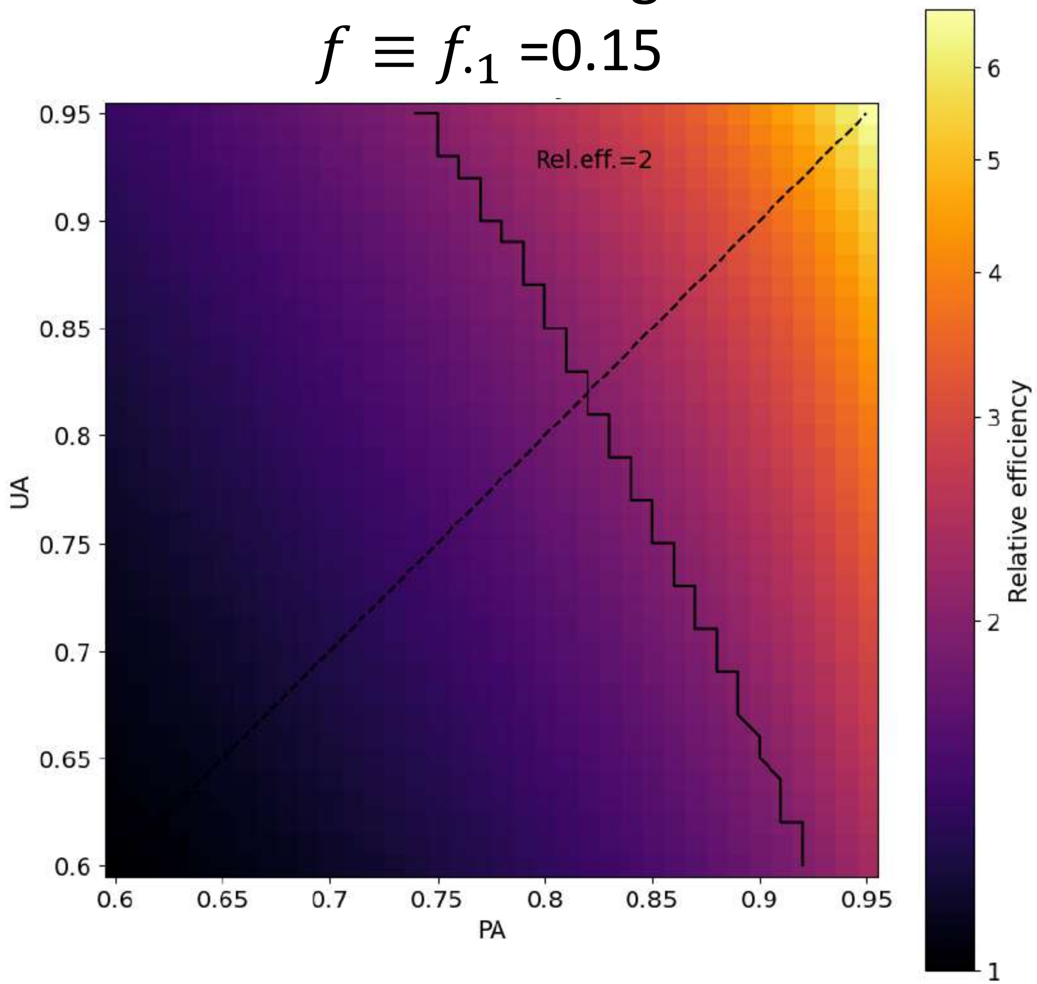
	A PA=0.9 UA=0.7	B PA=0.7 UA=0.9	C PA=0.8 UA=0.8	D PA=0.6 UA=0.6
	$f = 0.15$			
Overall	0.927	0.943	0.940	0.880
F-score=$\frac{2v_1 u_1}{v_1+u_1}$	0.788	0.788	0.800	0.600
Sample size n	693	934	836	1508
Rel. eff. strat.	3.27	2.43	2.71	1.50

Simple random sampling, sample size *n*=2,262

Relative efficiency of stratification

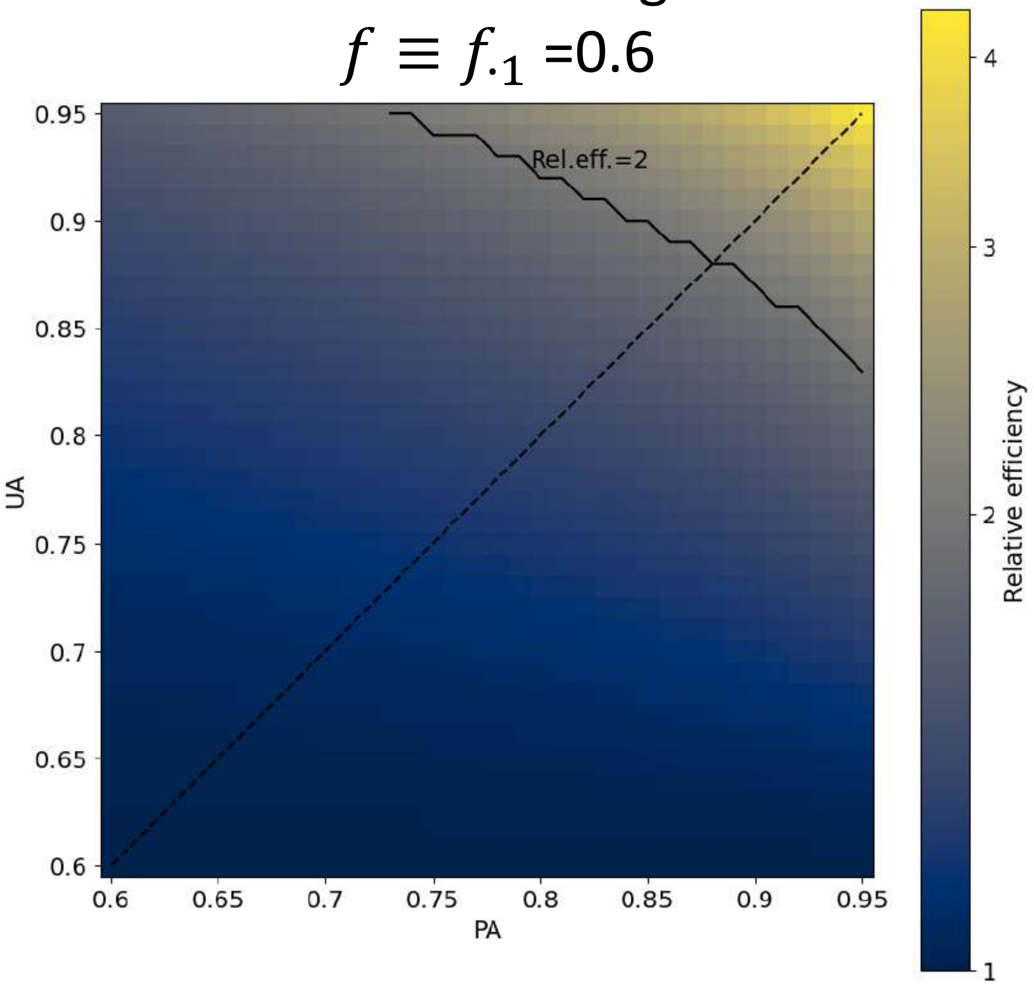
True area of the target class

$f \equiv f_{\cdot 1}$ =0.15

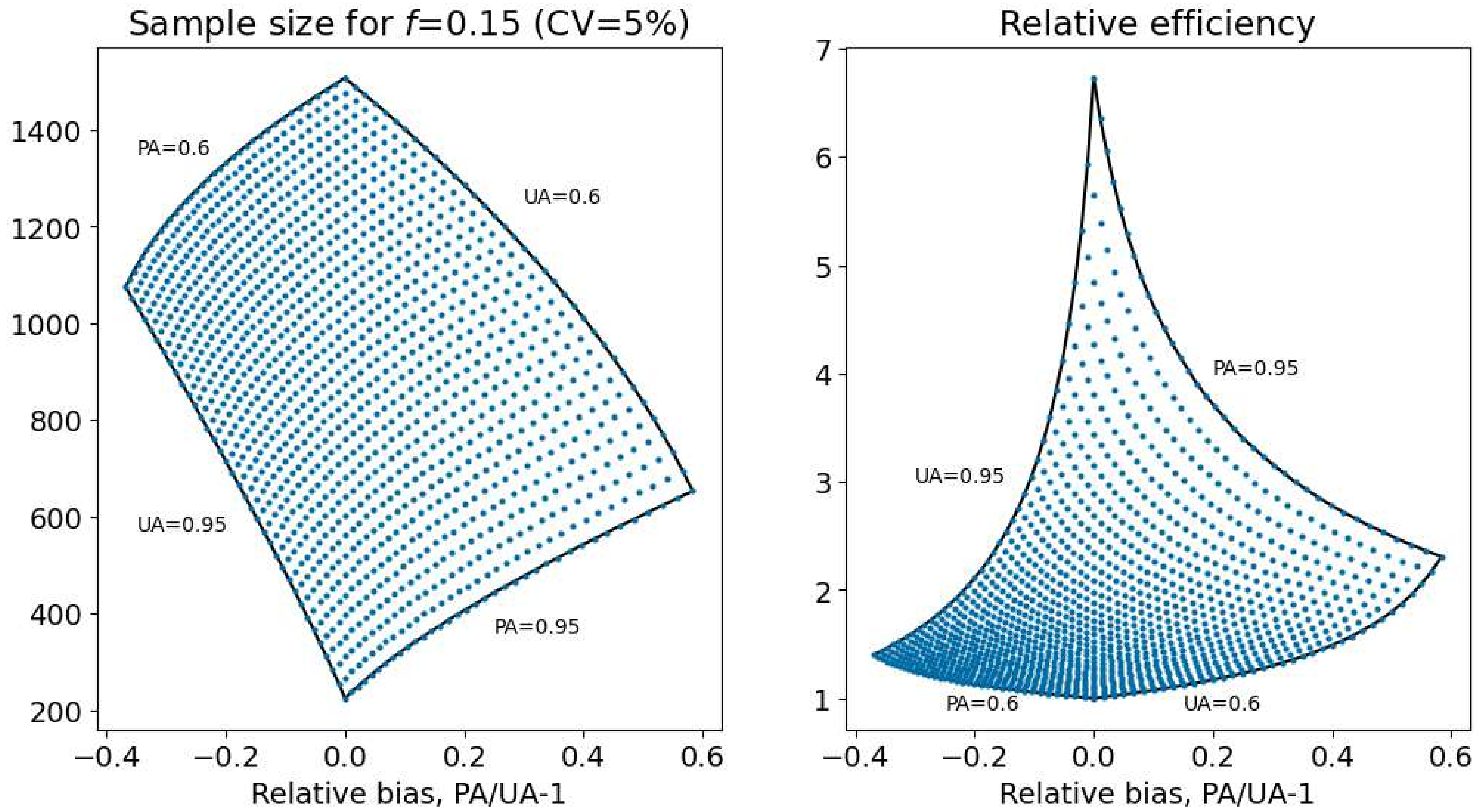


True area of the target class

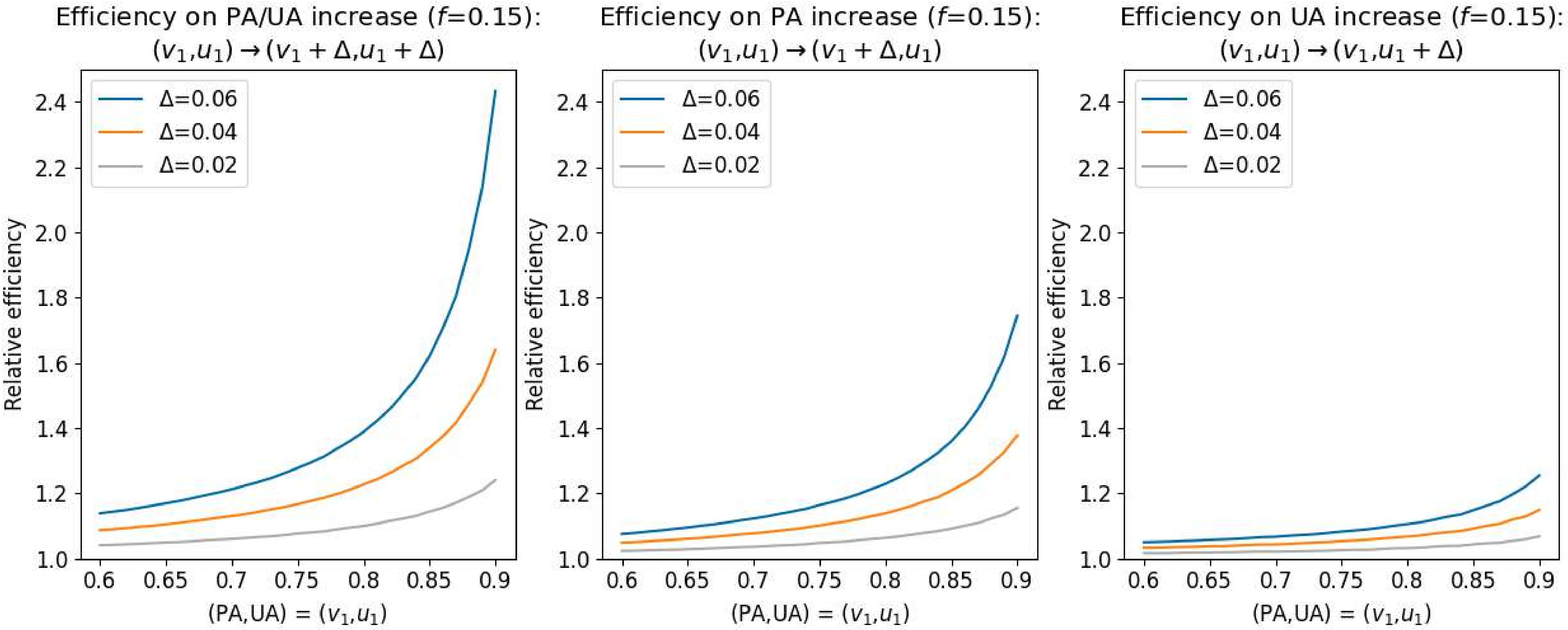
$f \equiv f_{\cdot 1}$ =0.6



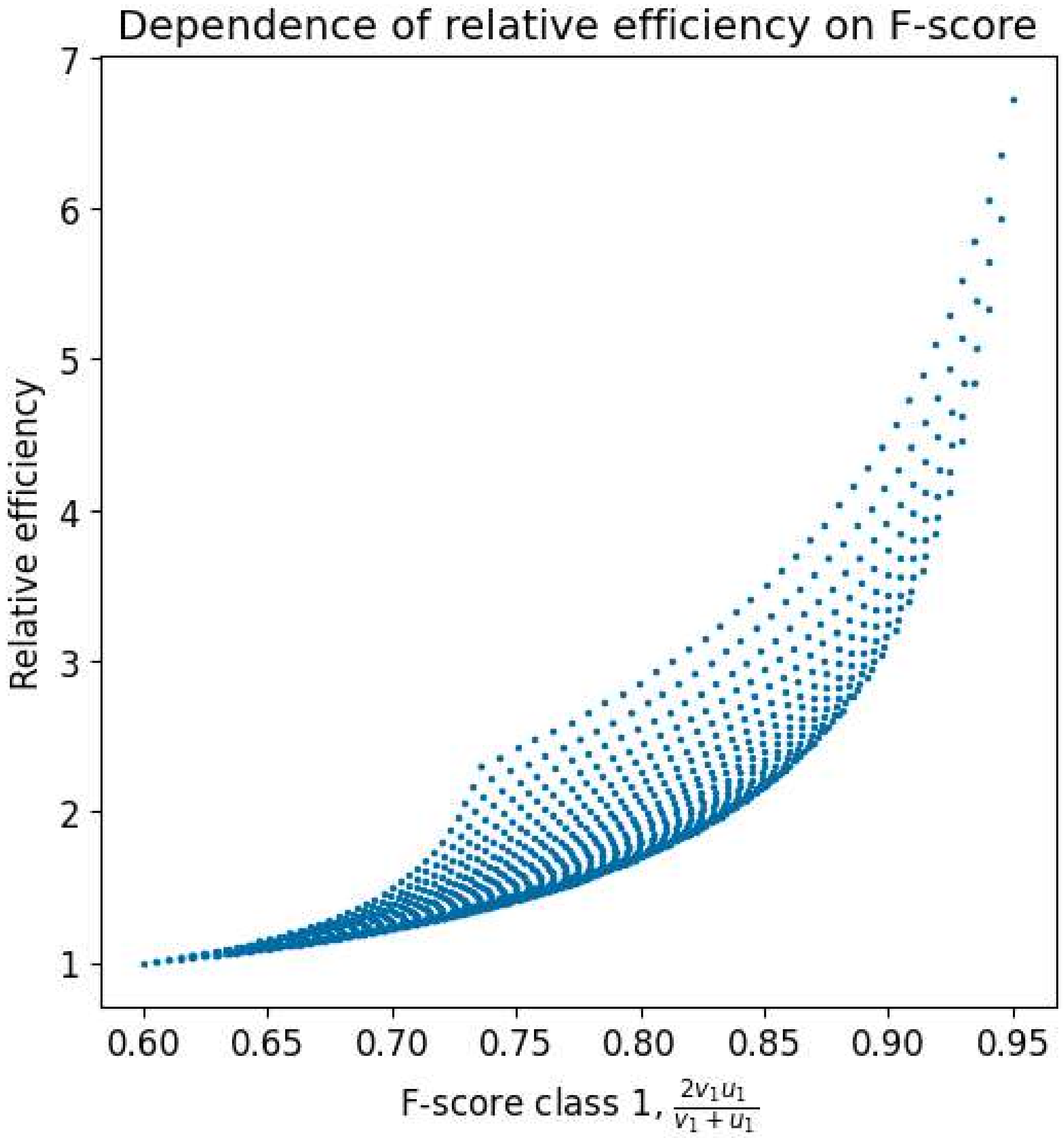
Relative efficiency $\eta_{A,B}$ calculated for each pair (v_1, u_1) (map *A*) in relation to the map at $(v_1, u_1) = (0.6, 0.6)$ (map *B*): $f = 0.15$ (left) and $f = 0.60$ (right). In each figure the solid line shows multiple solutions in the PA/UA space when a relative efficiency of 2 can be reached.



Dependence of the total sample size and relative efficiency (in relation to $(v_1, u_1) = (0.6, 0.6)$) on the relative bias.
Case $f = 0.15$.

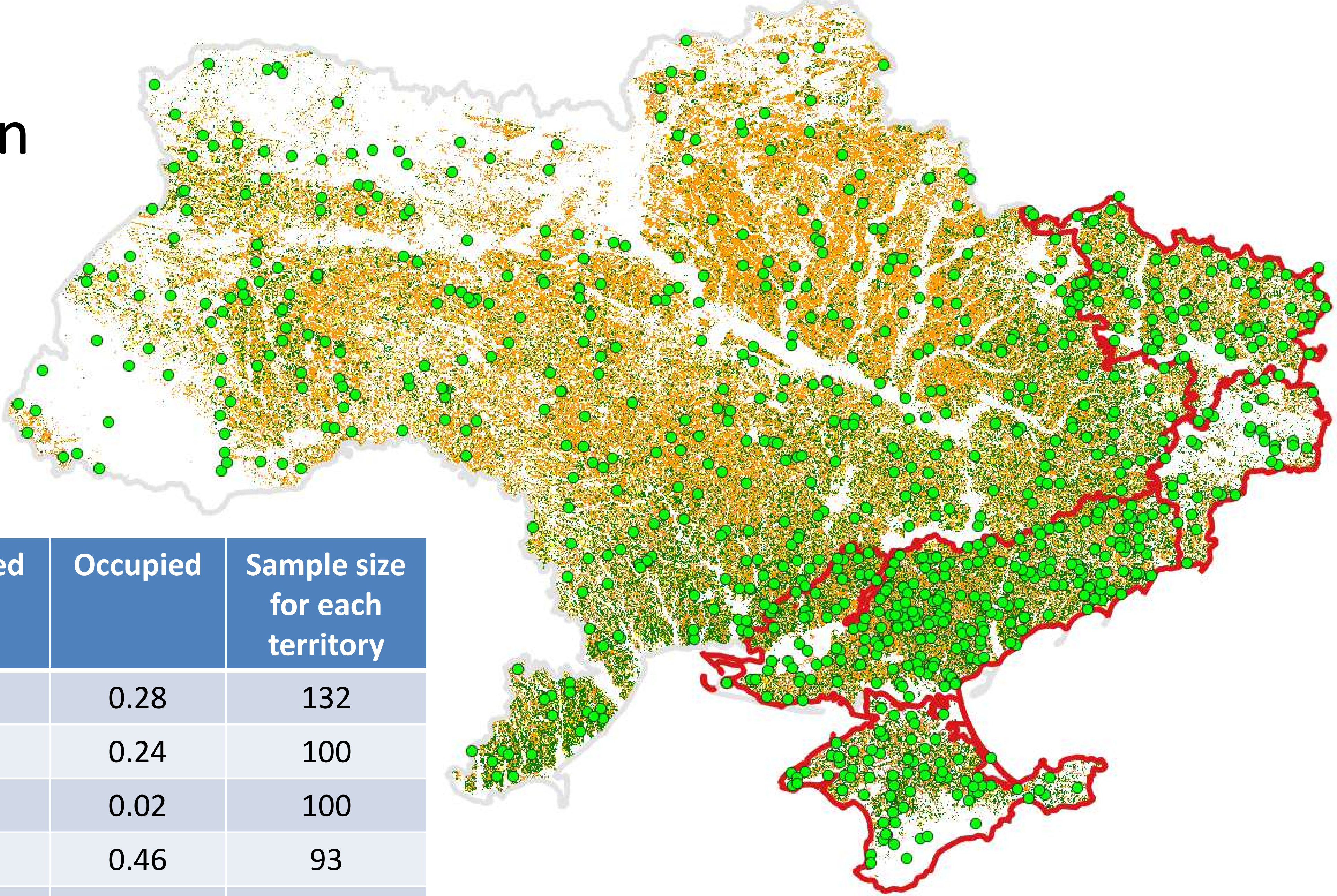


Dependence of improvements of relative efficiency when PA and UA increase by $\Delta = +0.02$, $+0.04$ and $+0.06$. Efficiency increase when both PA and UA increase by Δ is shown in left figure; PA increase only is center figure; and UA increase only is right figure.



Dependence of relative efficiency on F-score
Relative efficiency
F-score class 1, $\frac{2v_1u_1}{v_1+u_1}$
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1
2
3
4
5
6
7

Winter cereal area estimation in 2022



Stratum	Unoccupied	Occupied	Sample size for each territory
Summer crops	0.31	0.28	132
Winter cereal	0.13	0.24	100
Rapeseed	0.02	0.02	100
Non-cropland	0.53	0.46	93
Total	1	1	425

Example

Description of sample data as an error matrix of sample counts, n_{ij}

		Reference						
Map		*1.Summer*	*2. Winter cereals*	*3.Rapeseed*	*4. Non-cropland*	Total	Total mapped [ha]	W_i
	1. Summer	*108*	2	0	22			
	2. Winter cereals	24	*70*	0	6			
	3. Rapeseed	4	5	*87*	4			
	4. Non-cropland	10	4	0	*79*			
	Total							

Estimate of the proportion $\hat{p}_{22}$ - mapped as "2. Winter cereals" and reference "2. Winter cereals"

$$\hat{p}_{22} = W_2 \frac{n_{22}}{n_{2\cdot}} = 0.240 \frac{70}{100} = 0.168$$

Example

The error matrix populated by estimated **proportions of area**

		Reference				
Map		*1.Summer*	*2. Winter cereals*	*3.Rapeseed*	*4. Non-cropland*	Total $\hat{p}_{i\cdot}$
	1. Summer	*0.2283*	0.0042	0.0000	0.0465	
	2. Winter cereals	0.0576	*0.1680*	0.0000	0.0144	
	3. Rapeseed	0.0006	0.0008	*0.0140*	0.0006	
	4. Non-cropland	0.0500	0.0200	0.0000	*0.3949*	
	Total $\hat{p}_{\cdot j}$					

Example

The error matrix populated by estimated proportions of area

		Reference					
Map		*1.Summer*	*2. Winter cereals*	*3.Rape seed*	*4. Non-cropland*	Total $\hat{p}_{i\cdot}$	$\hat{U}_i$
	1. Summer	*0.2283*	0.0042	0.0000	0.0465		
	2. Winter cereals	0.0576	*0.1680*	0.0000	0.0144		
	3. Rapeseed	0.0006	0.0008	*0.0140*	0.0006		
	4. Non-cropland	0.0500	0.0200	0.0000	*0.3949*		
	Total $\hat{p}_{\cdot j}$						
	$\hat{P}_j$						

Overall accuracy $\hat{O} = \sum_{j=1}^{4} \hat{p}_{jj} = 0.2283 + 0.1680 + 0.014 + 0.3949 = 0.8052$

Estimate of **user's accuracy** of class 2 (winter cereals)

$\hat{U}_2 = {\hat{p}_{22}}/{\hat{p}_{2\cdot}} = {0.168}/{0.24} = 0.70$, or **commission error** $1 - \hat{U}_2 = 1 - 0.70 = 0.3$

Estimate of **producer's accuracy** of class 2

$\hat{P}_2 = {\hat{p}_{22}}/{\hat{p}_{\cdot 2}} = {0.168}/{0.1931} = 0.87$, or **omission error** $1 - \hat{P}_2 = 1 - 0.87 = 0.13$

Example

Area estimation

		Reference						
Map		*1.Summer*	*2. Winter cereals*	*3.Rapeseed*	*4. Non-cropland*	Total $\hat{p}_{i\cdot}$	$\widehat{U}_i$	Total mapped [ha]
	1. Summer	*0.2283*	0.0042	0.0000	0.0465	**0.279**	**0.82**	**3,470,363**
	2. Winter cereals	0.0576	*0.1680*	0.0000	0.0144	**0.240**	**0.70**	**2,985,323**
	3. Rapeseed	0.0006	0.0008	*0.0140*	0.0006	**0.016**	**0.87**	**199,502**
	4. Non-cropland	0.0500	0.0200	0.0000	*0.3949*	**0.465**	**0.85**	**5,781,756**
	Total $\hat{p}_{\cdot j}$	**0.3365**	**0.1931**	**0.0140**	**0.4565**	**1**		**12,436,944**
	$\widehat{P}_j$	**0.68**	**0.87**	**1.00**	**0.87**		**0.81**	

Estimate of areas:

Option 1 (directly from the map with pixel counting): **class 2 Winter cereals**

$A_{m,1} = 2{,}985{,}323 \text{ ha}$

Option 2 (from confusion matrix): **class 2 Winter cereals**

$\hat{A}_1 = \hat{p}_{\cdot 1} \times A_{\text{total}} = 0.1931 \times 12{,}436{,}944 \text{ ha} = 2{,}401{,}574 \text{ ha}$

Pixel counting overestimates **winter cereals** by 583,749 ha!

Example

Area estimation

		Reference						
Map		*1.Summer*	*2. Winter cereals*	*3.Rapeseed*	*4. Non-cropland*	Total $\hat{p}_{i\cdot}$	$\widehat{U}_i$	Total mapped [ha]
	1. Summer	*0.2283*	0.0042	0.0000	0.0465	**0.279**	**0.82**	**3,470,363**
	2. Winter cereals	0.0576	*0.1680*	0.0000	0.0144	**0.240**	**0.70**	**2,985,323**
	3. Rapeseed	0.0006	0.0008	*0.0140*	0.0006	**0.016**	**0.87**	**199,502**
	4. Non-cropland	0.0500	0.0200	0.0000	*0.3949*	**0.465**	**0.85**	**5,781,756**
	Total $\hat{p}_{\cdot j}$	**0.3365**	**0.1931**	**0.0140**	**0.4565**	**1**		**12,436,944**
	$\widehat{P}_j$	**0.68**	**0.87**	**1.00**	**0.87**		**0.81**	

Why is it happening?

Since $\widehat{P}_2 = 0.87$ meaning we miss winter cereal areas on average 13% (omission error), while $\widehat{U}_2 = 0.70$, meaning that we commit on average 30% (commission error). Misbalance in omission and commission errors leads to a bias (PA/UA-1 = 0.87/0.7-1 = 24%) if using pixel counting!

Pixel counting overestimates **winter cereals** by 583,749 ha!

Final areas: temporary occupied territories (as of July 2022)

	Summer	Winter cereals	Winter rapeseed	Non-cropland
Estimated area [ha]	4,185,540	2,400,960	173,567	5,676,877
95% CI of area [ha]	±499,631	±367,998	±13,217	±497,106

CI = confidence interval at 95%

Conclusions

- When **maps** are used for **area estimation** using a **sample-based** approach within the **design-based** inference framework:
  - **Map quality** impacts stratification **efficiency**
    - A more accurate map → **smaller sample size** to reach the desired precision of the estimate or
    - A more accurate map → **higher precision** when the sample size is fixed
  - A criterion (**relative efficiency**) ← **implications** of **accuracy** increase
    - Depends on class-specific PA and UA, and target class area
    - The impact is not linear and contributions of PA, UA, and *f* are not equal
    - F-score is not an adequate metric
    - Costs associated with reference data collection (ground surveys, photo-interpretation)
      - E.g, LUCAS [d'Andrimont et al., 2020], sample unit cost=35 Euro
      - Target $f = 0.05$ (CV=5%), relative efficiency of 2 →140,000 Euro (4,000 sample units) to 70,000 Euro (2,000)
      - Target $f = 0.3$ (CV=5%), relative efficiency of 2 → 26,250 Euro (750) to 13,125 Euro (325).

Conclusions (cont')

- Re-emphasized
  - **Rigorous map quality assessment**, e.g., [*Olofsson et al., 2014*], otherwise, a map is a "pretty picture" [*McRoberts, 2011*])
  - Full **area-based confusion matrix** to be reported → critical for further map use (esp., area)
  - Pixel counting estimator → biased → discouraged and should not be used
- Future directions
  - **Costs of map generation**, e.g., data costs & compute/storage costs, not considered

$$\eta_{A,B} = \frac{c_1^B + c_0 n^B}{c_1^A + c_0 n^A}$$

  - Sample unit is a block (segment)