Supplement to "Design-based mapping of land use/land cover classes with bootstrap estimation of precision by nearest-neighbour interpolation"

by

A Marcelli[(1)], RM Di Biase[(2)], P Corona[(3)], SV Stehman[(4)], L Fattorini[(5)]

University of Tuscia [(1)], University of Milano Bicocca[(2)],

CREA – Research Centre for Forestry and Wood[(3)], SUNY College of Environmental Science and Forestry[(4)],

and University of Siena [(5)]

Appendix 1. Some consistency results.

For any $\delta > 0$ and for any location $p \in A$, denote by $B(p,\delta) = \{q: q \in A, \|p - q\| < \delta\}$ the δ-ball of p within A. Moreover, denote by $V(p,\delta) = \bigcap_{i=1}^{n}\{P_i \notin B(p,\delta)\}$ the event that a void, i.e., no sample location, occurs within the δ-ball of p.

If p is an interior point, i.e. $p \in A\backslash\Delta$, then the greatest distance δ_p such that $B(p,\delta_p) \cap \Delta = \emptyset$ defines the event $V^c(p,\delta_p)$, i.e. the event that at least a sample point falls within $B(p,\delta_p)$, in such a way that if $V^c(p,\delta_p)$ occurs, the NN interpolar at p guesses the true class. In other words

$$\Pr\{\hat{y}(p) = y(p)\} \geq \Pr\{V^c(p,\delta_p)\}$$

that is equivalent to

$$\mathrm{Err(p)} \leq \Pr\{V(p,\delta_p)\} \qquad (A.1)$$

Now, denote by $a(p) \leq |A|$ the size of the δ_p-ball of p. Under URS, the probability that the i-th sample location falls outside the δ_p-ball of p is given by

$$\Pr\{P_i \notin B(p,\delta_p)\} = 1 - \frac{a(p)}{|A|}\ , \qquad i = 1,\dots,n$$

in such a way that, owing to independence of sample locations under URS, the probability that no sample location falls within the δ_p-ball of p is given by

$$\Pr\{V(p,\delta_p)\} = \left\{1 - \frac{a(p)}{A}\right\}^n \qquad (A.2)$$

Then, substituting (A.2) into (A.1), under URS it holds that

$$\mathrm{Err(p)} \leq \left\{1 - \frac{a(p)}{A}\right\}^n \qquad (A.3)$$

i.e., under URS the NN interpolator is pointwise consistent for each interior point p with an error probability that decreases at least at a c^n rate, with $c \in (0,1)$. Therefore, under URS the NN interpolator is also consistent in mean.

Regarding consistency under TSS and SGS, for a sample size n, denote by $A_{1,n}, \dots, A_{n,n}$ the n patches of equal size $|A|/n$ that partition A, and denote by $i(p)$ the label identifying the patch containing p. Suppose that as n increases the $A_{i,n}$s decrease in size in such a way that $\lim_{n\to\infty} \min_{i=1,\dots,n} \mathrm{diam}(A_{i,n}) = 0$. Therefore, there exists a sample size n_0 such that, for each $n > n_0$ it holds that $A_{i(p),n} \subset B(p,\delta_p)$, in such a way that

$$\mathrm{Err(p)} \leq \Pr\{V(p,\delta_p)\} \leq \Pr\{P_{i(p)} \notin B(p,\delta_p)\} \leq \Pr\{P_{i(p)} \notin A_{i(p),n}\} = 0 \qquad (A.4)$$

In practice, inequality (A.4) states that for a sufficiently large size, the NN interpolator does not provide errors. That obviously proves pointwise consistency and consistency in mean for the NN interpolator under TSS and SGS.

Appendix 2. Features of bootstrap estimators of precision.

Owing to the dichotomous nature of $z(p)$, the bootstrap estimator of $\mathrm{Err(p)}$ can be rewritten as

$$\widehat{Err}_B^*(p) = \frac{1}{B}\sum_{b=1}^{B} z_b^*(p) = \frac{1}{B}\sum_{b=1}^{B} I[\hat{y}_b^*(p) \neq \hat{y}(p)]$$

Accordingly, for a sufficiently large B, owing to the strong law of large numbers it holds that

$$\widehat{Err}_B^*(p) \sim \mathrm{E}^*\{I[\hat{y}^*(p) \neq \hat{y}(p)] | P_1, \dots, P_n\}$$

where E^* denotes expectation with respect to the bootstrap experiment and conditional to the original sample $P_1, \dots, P_n$, and $\hat{y}^*(p)$ denotes the estimate of $y(p)$ occurred in a generic bootstrap resampling.

Because each $\widehat{D}_k$ (see equation 14) can be rewritten as

$$\widehat{D}_k = \{p: p \in A, y(P_{NN(p)}) = c_k\} = \{p: p \in A, P_{NN(p)} \in D_k\}$$

in such a way that

$$\widehat{Err}_B^*(p) \sim E^*\{I[\hat{y}^*(p) \neq \hat{y}(p)] | P_1, \dots, P_n\}$$
$$= \sum_{k=1}^{K} I(P_{NN(p)} \in D_k) \sum_{h \neq k=1}^{K} \Pr\{P^*_{NN(p)} \in \widehat{D}_h | P_1, \dots, P_n\} \qquad (B.1)$$

Then, if $p \in A \backslash \Delta$ and $p \in D_{k_0}$, i.e., $y(p) = c_{k_0}$, from (B.1) and from the identity that

$$I(P_{NN(p)} \in D_{k_0}) = 1 - I(P_{NN(p)} \in D^c_{k_0})$$

it follows that

$$\widehat{Err}_B^*(p) \sim \sum_{h \neq k_0=1}^{K} \Pr\{P^*_{NN(p)} \in \widehat{D}_h | P_1, \dots, P_n\}$$
$$- I(P_{NN(p)} \in D^c_{k_0}) \sum_{h \neq k_0=1}^{K} \Pr\{P^*_{NN(p)} \in \widehat{D}_h | P_1, \dots, P_n\} +$$
$$\sum_{k \neq k_0=1}^{K} I(P_{NN(p)} \in D_k) \sum_{h \neq k=1}^{K} \Pr\{P^*_{NN(p)} \in \widehat{D}_h | P_1, \dots, P_n\} \qquad (B.2)$$

Once again, as stated in Appendix A, under URS, $\Pr(P_{NN(p)} \in D^c_{k_0})$, i.e., the error probability quickly approaches 0 at a rate of at least c^n with $c \in (0,1)$, while under SGS and TSS, it is definitively equal to 0 for a sufficiently large n. Therefore, the random variable $I(P_{NN(p)} \in D^c_{k_0})$ converges almost surely to 0, and, *a fortiori*, each $I(P_{NN(p)} \in D_k)$ for $k \neq k_0$ converges almost surely to 0. Then, from (B.2) it holds

$$\widehat{Err}_B^*(p) \sim \sum_{h \neq k_0 = 1}^{K} \Pr\left\{P_{NN(p)}^* \in \widehat{D}_h | P_1, \dots, P_n\right\}$$

in such a way that

$$\frac{\widehat{Err}_B^*(p)}{\mathrm{Err}(p)} \sim \frac{\sum_{h \neq k_0 = 0}^{K} \Pr\left\{P_{NN(p)}^* \in \widehat{D}_h | P_1, \dots, P_n\right\}}{\sum_{h \neq k_0 = 0}^{K} \Pr\left\{P_{NN(p)} \in D_h\right\}} \tag{B.3}$$

Also in this case, in accordance with Appendix A, the random variable (B.3) is the ratio of two quantities that approach 0 at rates at least of exponential nature and as such it may be very unstable especially in the interior zones, those far by Δ where interpolation is precise and the denominator of (B.3) approaches 0.

Therefore, owing to the volatility of (B.3), the quantity

$$borat_B(p) = \frac{\mathrm{E}\left\{\widehat{Err}_B^*(p)\right\}}{\mathrm{Err}(p)} = \mathrm{E}\left\{\frac{\widehat{Err}_B^*(p)}{\mathrm{Err}(p)}\right\}$$

does not admit any upper bound greater than one, as that achieved by Fattorini et al. (2021, Theorem 3) that proves the conservative nature of the bootstrap means squared error estimator of the NN interpolator for quantitative variables under suitable assumptions.

Appendix 3. Figures from the simulation study.

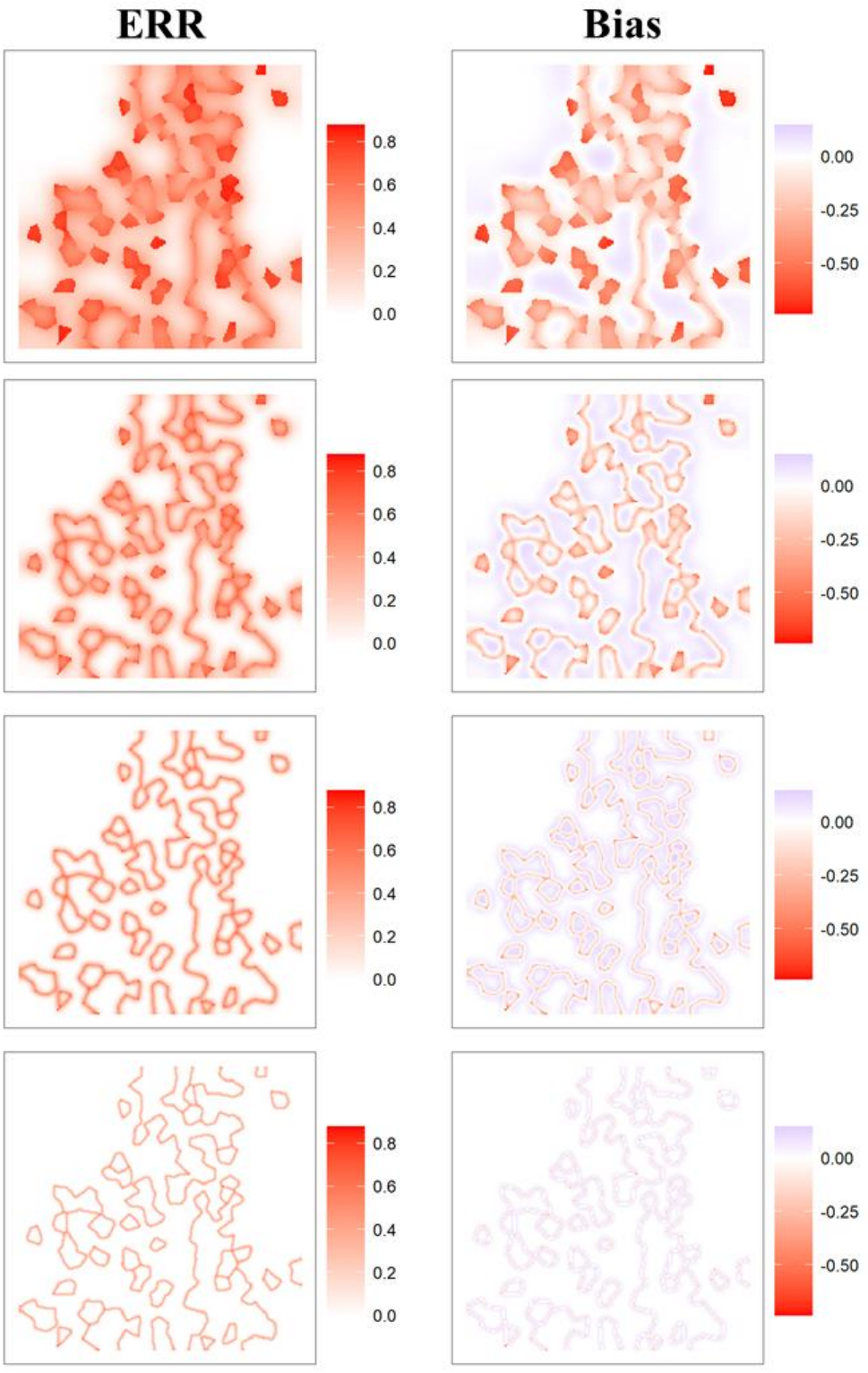


Figure C1. Spatial patterns of the error probabilities (left column) and the bias of their bootstrap estimator (right column) evaluated at each node of the regular grid of 201×201 locations within the quadrat of Figure 1 under URS and sample sizes $n = 100;\ 400;\ 1{,}600;\ 10{,}000$ (rows).

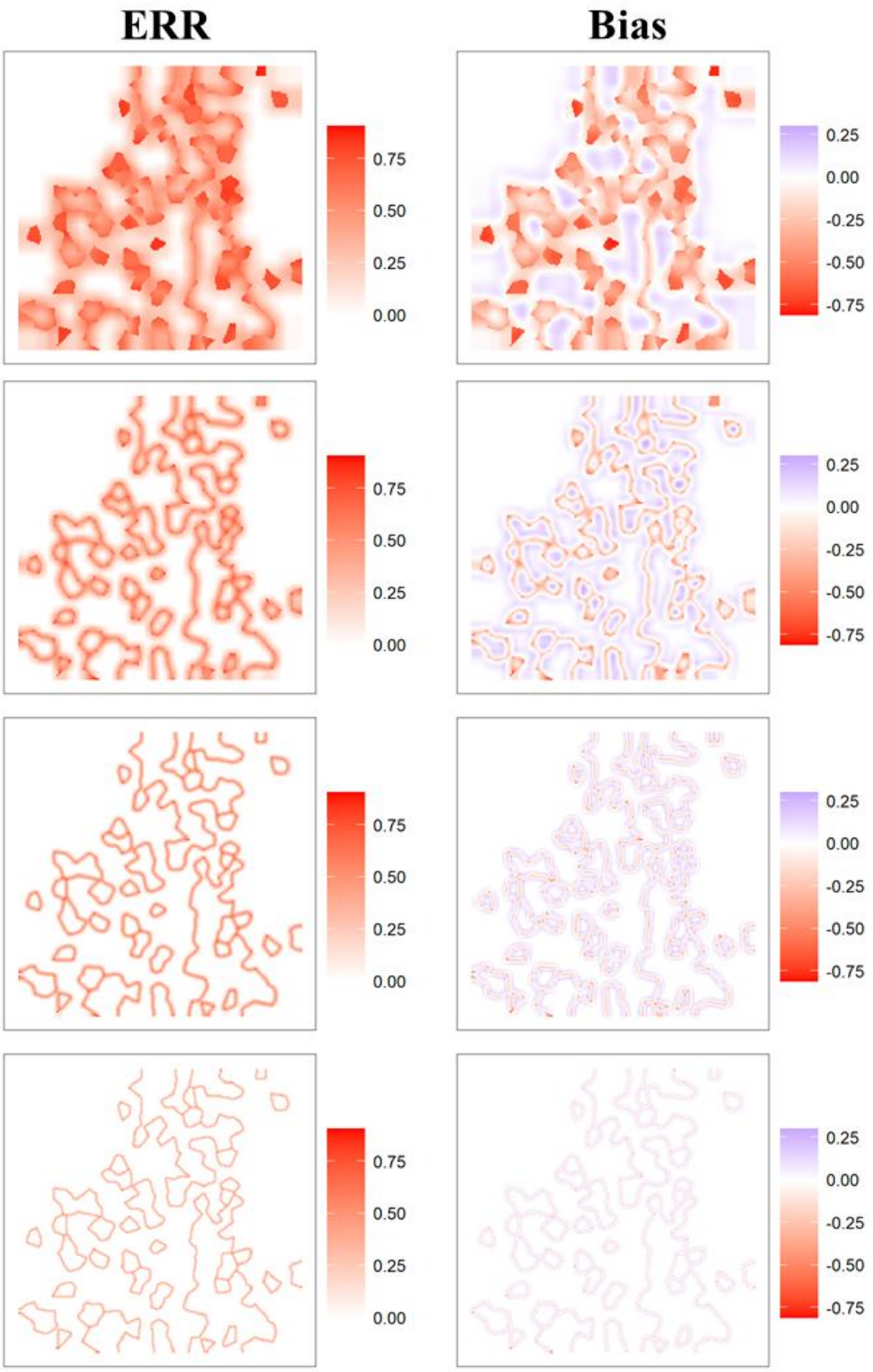
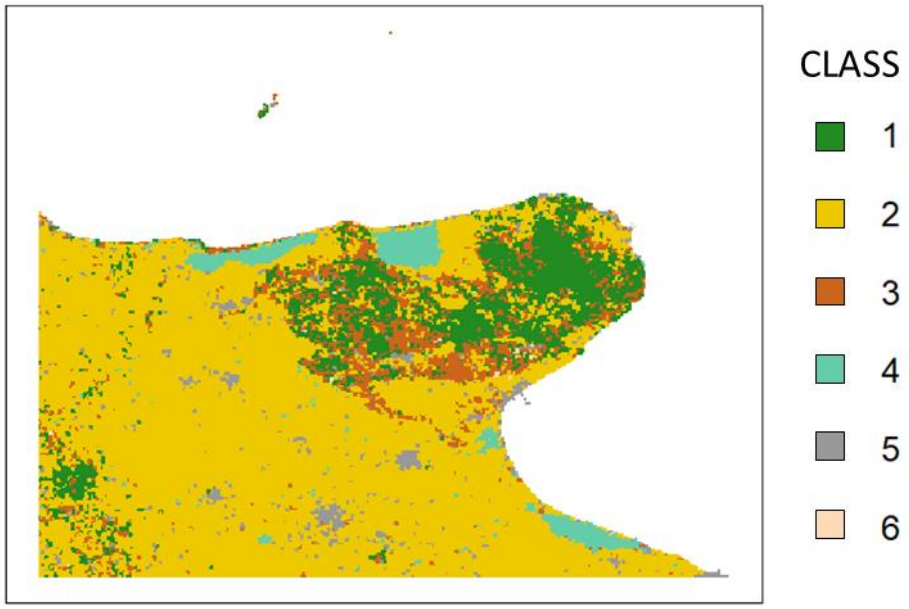


Figure C2. Spatial patterns of the error probabilities (left column) and the bias of their bootstrap estimator (right column) evaluated at each node of the regular grid of 201×201 locations within the quadrat of Figure 1 under SGS and sample sizes $n = 100;\ 400;\ 1{,}600;\ 10{,}000$ (rows).

Appendix 4. Figures from cases studies.

(a)



(b) (c)

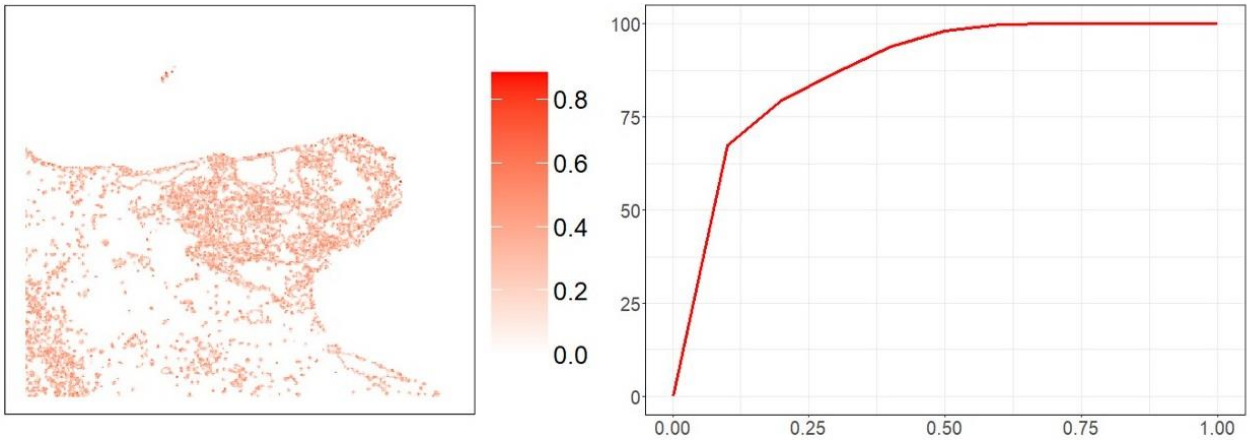
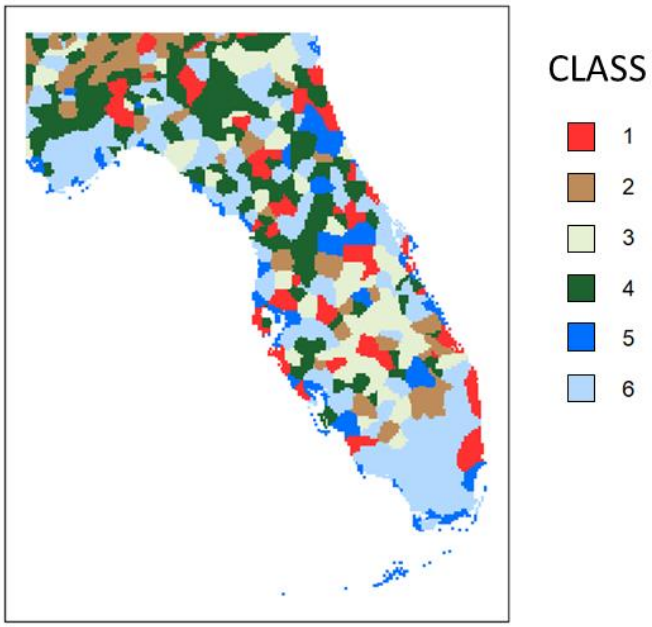
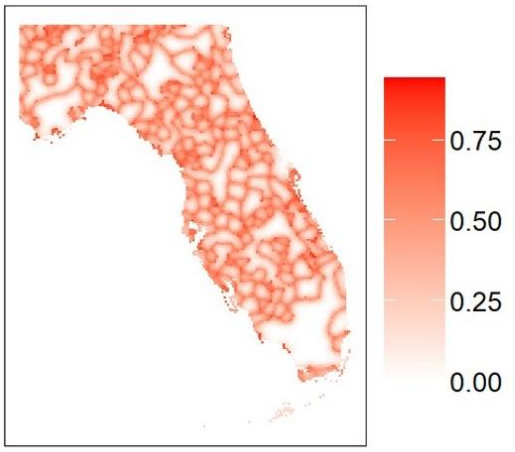


Figure D1. (a) Map of the six land use classes estimated from the IUTI TSS sample at the year 2008 regarding the zone of Gargano Promontory (Puglia Region, Southern Italy); (b) Map of the estimates of the error probabilities achieved by $B = 1{,}000$ bootstrap samples; (c) Cumulative frequencies of the estimates of the error probabilities.

(a)



(b)



(c)

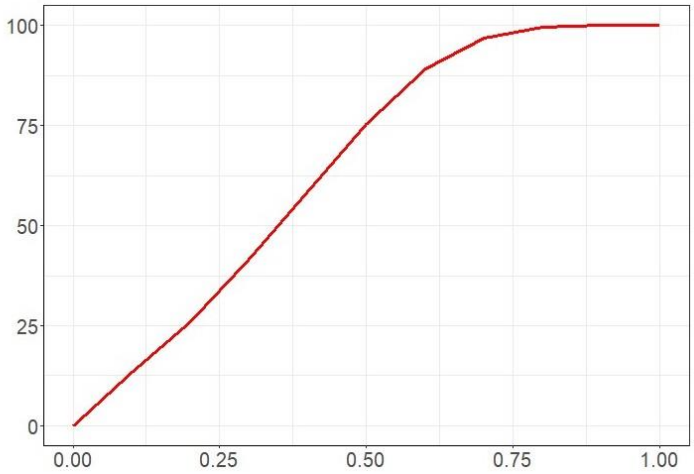
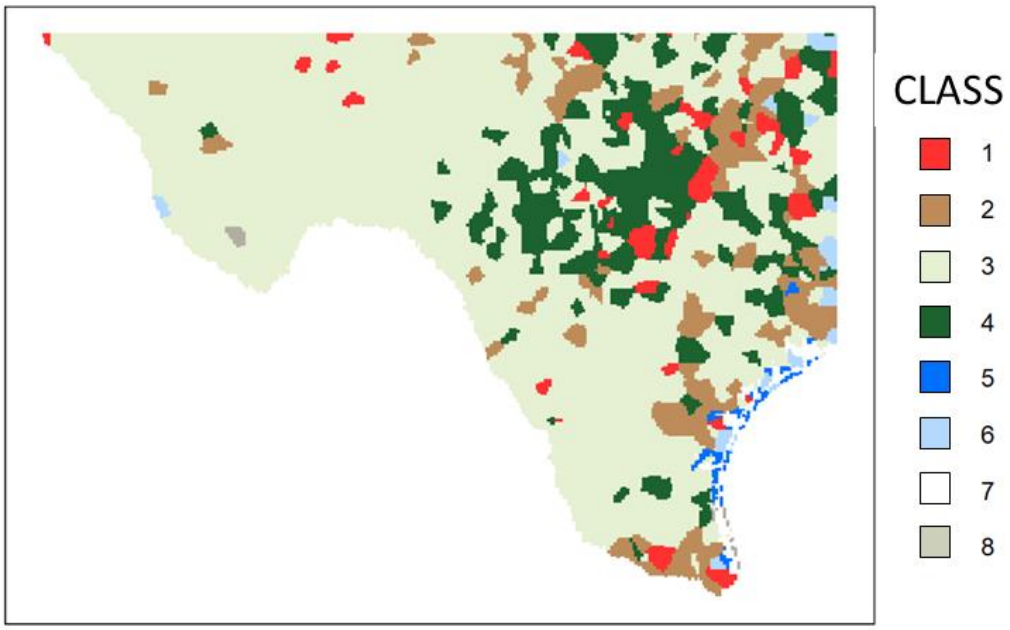


Figure D2. (a) Map of the eight land use classes estimated from the LCMAP SRSWOR sample at the year 2017 regarding the state of Florida; (b) Map of the estimates of the error probabilities achieved by B =1,000 bootstrap samples; (c) Cumulative frequencies of the estimates of the error probabilities.

(a)



(b) (c)

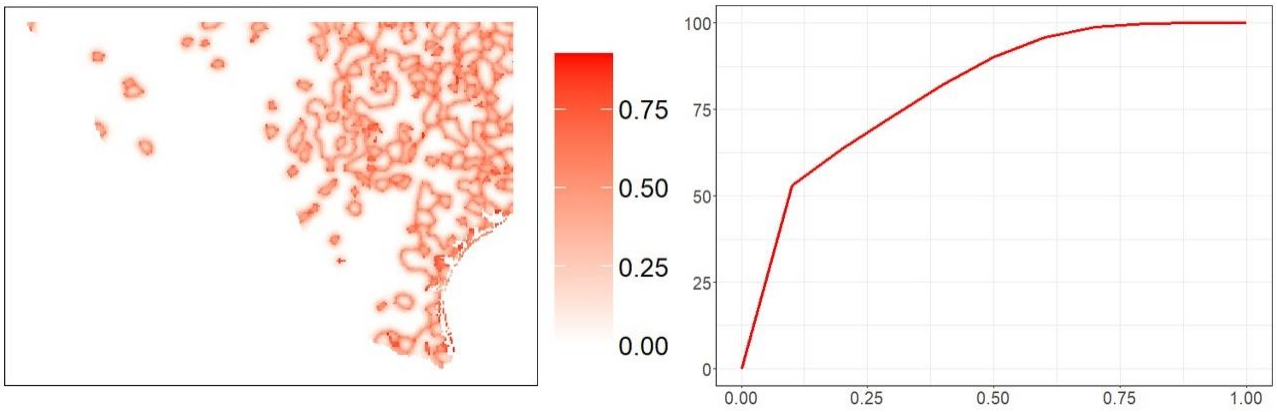


Figure D3. (a) Map of the eight land use classes estimated from the LCMAP SRSWOR sample at the year 2017 regarding the state of Texas; (b) Map of the estimates of the error probabilities achieved by $B = 1{,}000$ bootstrap samples; (c) Cumulative frequencies of the estimates of the error probabilities.