SUPPLEMENTARY MATERIALS

Design-based consistent strategies exploiting auxiliary information in environmental mapping

S1. A finite sample result for bounding kNN estimation errors in the geographic space

Using the notation of Section 3, $Q_{g(1)}(\mathbf{p},\delta) \subset \cdots \subset Q_{g(k)}(\mathbf{p},\delta)$, in such a way that $I\{Q_{g(1)}(\mathbf{p},\delta)\} \leq \cdots \leq I\{Q_{g(k)}(\mathbf{p},\delta)\}$. Therefore, from (5) it follows that

$$
\begin{aligned}
|\hat{f}_g(\mathbf{p}) - f(\mathbf{p})| &= |I(Q_{\mathbf{p}})f(\mathbf{p}) + I(Q^c_{\mathbf{p}}) \sum_{i \in H_{g,k}(\mathbf{p})} w_{g,i}(\mathbf{p}) f(\mathbf{P}_i) - f(\mathbf{p})| \\
&= |I(Q^c_{\mathbf{p}}) \sum_{i \in H_{g,k}(\mathbf{p})} w_{g,i}(\mathbf{p}) f(\mathbf{P}_i) - I(Q^c_{\mathbf{p}}) f(\mathbf{p})| \\
&= I(Q^c_{\mathbf{p}}) | \sum_{i \in H_{g,k}(\mathbf{p})} w_{g,i}(\mathbf{p}) f(\mathbf{P}_i) - f(\mathbf{p})| \\
&\leq | \sum_{i \in H_{g,k}(\mathbf{p})} w_{g,i}(\mathbf{p}) f(\mathbf{P}_i) - f(\mathbf{p})| \\
&= | \sum_{i \in H_{g,k}(\mathbf{p})} w_{g,i}(\mathbf{p}) \{f(\mathbf{P}_i) - f(\mathbf{p})\}| \leq \sum_{i \in H_{g,k}(\mathbf{p})} w_{g,i}(\mathbf{p}) |f(\mathbf{P}_i) - f(\mathbf{p})| \\
&= \sum_{i \in H_{g,k}(\mathbf{p})} [I(Q_{g,i}(\mathbf{p},\delta)) + I(Q^c_{g,i}(\mathbf{p},\delta))] w_{g,i}(\mathbf{p}) |f(\mathbf{P_i}) - f(\mathbf{p})| \\
&= \sum_{i \in H_{g,k}(\mathbf{p})} I(Q_{g,i}(\mathbf{p},\delta)) w_{g,i}(\mathbf{p}) |f(\mathbf{P}_i) - f(\mathbf{p})| \\
&+ \sum_{i \in H_{g,k}(\mathbf{p})} I(Q^c_{g,i}(\mathbf{p},\delta)) w_{g,i}(\mathbf{p}) |f(\mathbf{P}_i) - f(\mathbf{p})| \\
&\leq L \sum_{i \in H_{g,k}(\mathbf{p})} I(Q_{g,i}(\mathbf{p},\delta)) w_{g,i}(\mathbf{p}) + \sum_{i \in H_{g,k}(\mathbf{p})} w_{g,i}(\mathbf{p}) \Delta_g(\mathbf{p},\delta) \\
&= L \sum_{l=1}^{k} I(Q_{g(l)}(\mathbf{p},\delta)) \sum_{i \in H_{g(l)}(\mathbf{p})} w_{g,i}(\mathbf{p}) + \Delta_g(\mathbf{p},\delta)
\end{aligned}
$$

$$
\begin{aligned}
&\leq LI(Q_{g(k)}(\mathbf{p},\delta)) \sum_{l=1}^{k} \sum_{i \in H_{g(l)}(\mathbf{p})} w_{g,i}(\mathbf{p}) + \Delta_g(\mathbf{p},\delta) \\
&= LI(Q_{g(k)}(\mathbf{p},\delta)) + \Delta_g(\mathbf{p},\delta).
\end{aligned} \tag{S1.1}
$$

Taking the expectations of both sides of (S1.1), result (8) follows.

S2. Pointwise kNN consistency in geographic space under URS, TSS, and SGS (continuous populations), and OPSS and SYS (finite populations of areas)

The event $Q^c_{g(k)}(\mathbf{p},\delta_m)$ can be rewritten as

$$
Q^c_{g(k)}(\mathbf{p},\delta_m) = \left\{ \sum_{i=1}^{n} I(d_g(\mathbf{P}_i,\mathbf{p}) \leq \delta_m) \geq k \right\}.
$$

Therefore, if k groups of $[n/k]$ locations, say $H_{1,n},\ldots,H_{k,n}$, are considered among the n sample locations, then one way in which the event $Q^c_{g(k)}(\mathbf{p},\delta)$ can occur is

$$
\bigcap_{l=1}^{k} \left\{ \sum_{i \in H_{l,n}} I(d_g(\mathbf{P}_i,\mathbf{p}) \leq \delta_m) \geq 1 \right\} \subset Q^c_{g(k)}(\mathbf{p},\delta_m). \tag{S2.1}
$$

Because under URS the $\mathbf{P}_i$s are independent and equally distributed, the k events of the intersection in the left side of (S2.1) are independent and have the same probability. Accordingly, from (S2.1) it follows that

$$
\begin{aligned}
\Pr\left\{Q^c_{g(k)}(\mathbf{p},\delta_m)\right\} &\geq \Pr\left[\bigcap_{l=1}^{k} \left\{ \sum_{i \in H_{l,n}} I(d_g(\mathbf{P}_i,\mathbf{p}) \leq \delta_m) \geq 1 \right\}\right] \\
&= \left[\Pr\left\{ \sum_{i \in H_{1,n}} I(d_g(\mathbf{P}_i,\mathbf{p}) \leq \delta_m) \geq 1 \right\}\right]^k
\end{aligned}
$$

from which

$$
\begin{aligned}
\Pr\left\{Q_{g(k)}(\mathbf{p},\delta_m)\right\} &\leq 1-\left[\Pr\left\{\sum_{i\in H_{1,n}} I(d_g(\mathbf{P}_i,\mathbf{p})\leq\delta_m)\geq 1\right\}\right]^k \\
&\leq k\left[1-\Pr\left\{\sum_{i\in H_{1,n}} I(d_g\left(\mathbf{P}_i,\mathbf{p}\right)\leq\delta_m)\geq 1\right\}\right] \qquad \text{(S2.2)} \\
&= k\Pr\left\{\sum_{i\in H_{1,n}} I(d_g\left(\mathbf{P}_i,\mathbf{p}\right)\leq\delta_m)=0\right\}.
\end{aligned}
$$

As proven in the Appendix C by Fattorini et al. (2022), under URS the probability of the event in the right side of (S2.2) approaches 0 with δ_m. Therefore (S2.2) proves condition (11).

Under TSS and SGS, any point $\mathbf{p}\in\mathcal{A}$ identifies the polygon that contains $\mathbf{p}$ and $k-1$ surrounding polygons. Each of these k polygons contains a sample location selected at random within them, so that the k sample locations constitute the locations involved in the kNN interpolation. Usually, it is supposed that polygons do not have stretched shapes, so that, as m increases, the diameters of the m polygons partitioning $\mathcal{A}$ approach 0 (Barabesi et al., 2012). Therefore, because polygons become smaller and smaller, there exist a real $\delta>0$ and an integer m_0 such that for $m>m_0$ the ball of radius δ contains the k neighboring polygons, i.e.,

$$
\Pr\left\{Q_{g(k)}\left(\mathbf{p},\delta\right)\right\}=0,\ m>m_0. \qquad \text{(S2.3)}
$$

Obviously, condition (S2.3) joined with the continuity condition (10) ensures (9).

Similarly, in the case of finite populations of areas, under SYS and OPSS for each m there is a population of N_m areas partitioned into n_m blocks of contiguous areas. Areas and blocks must be of equal shape and each block must contain an equal number $b_m=N_m/n_m$ of areas in the case of SYS (Figure 1a), while areas and blocks can be of different shapes and blocks can contain a different number of areas $b_{m,1},\ldots,b_{m,n_m}$, whose sum gives N_m, in the case of OPSS (Figure 1b). Then, in the case of SYS an area is randomly selected within one block and repeated in the remaining blocks, while under OPSS an area is randomly and independently selected within each block. In this framework, any area j identifies the block that contains

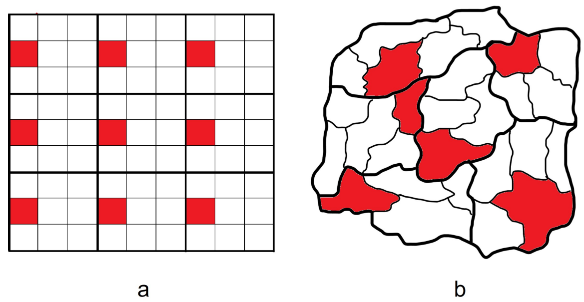

Figure 1: Graphical representation of a sample of areas (in red) selected by means of SYS (a) and OPSS (b).

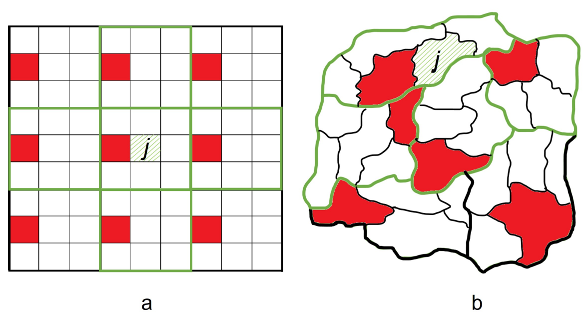

Figure 2: Graphical representation of area j and surrounding blocks (highlighted in green) for $k = 4$ under SYS (a) and OPSS (b).

j and a set of surrounding (neighboring) blocks whose number may be equal or greater than k, especially in regular grids of areas (Figures 2a and 2b). If j is not a selected area (in the opposite case interpolation is without error), each of these blocks contains a sample area involved in the kNN interpolation. Therefore, because the diameters of the areas partitioning $\mathcal{A}$ approach 0 as m increases (see Section 2), areas become smaller and smaller and also blocks become smaller and smaller if it is supposed that the number of areas in each block is bounded by b_0 for each m. Then, for any $\delta > 0$ there exists an integer m_0 such that for any $m > m_0$ the ball of radius δ contains at least k blocks, i.e., condition (S2.3) holds, and joined with the continuity condition (10), it ensures (9).

S3. Considerations on the one-to-one nature of auxiliary variables in the geographic space

Let $x_1, \ldots, x_G$ be G bounded functions defined on the study region $\mathcal{A}$ corresponding to G auxiliary variables. For each location $\mathbf{p} \in \mathcal{A}$, let $E_G(\mathbf{p})$ be the Borelian set of points $\mathbf{q} \neq \mathbf{p} \in \mathcal{A}$ such that $\mathbf{x}(\mathbf{p}) = \mathbf{x}(\mathbf{q})$ and let $I_G(\mathbf{p})$ be the indicator of $E_G(\mathbf{p})$. Therefore, the set of G auxiliary variables failed to be a one-to-one function on $\mathcal{A}$ if

$$I_G(\mathcal{A}) = \max_{\mathbf{p} \in \mathcal{A}} I_G(\mathbf{p}) = 1 \tag{S3.1}$$

Now, consider an additional auxiliary variable x_{G+1}, so that the Borelian set $E_{G+1}(\mathbf{p})$ can be expressed as

$$E_{G+1}(\mathbf{p}) = E_G(\mathbf{p}) \cap \{\mathbf{q} \neq \mathbf{p} \in \mathcal{A} : x_{G+1}(\mathbf{q}) = x_{G+1}(\mathbf{p})\} \subset E_G(\mathbf{p})$$

in such a way that $I_{G+1}(\mathbf{p}) \leq I_G(\mathbf{p})$. Accordingly, from (S3.1) it follows that $I_{G+1}(\mathcal{A}) \leq I_G(\mathcal{A})$. In conclusion, as the number G of auxiliary variables increases the sequence $\{I_G(\mathcal{A})\}$ is decreasing, i.e. as the number of auxiliary variable increases it is more difficult that the corresponding function fail to be one-to-one onto $\mathcal{A}$, even if no sufficiency condition can be claimed.

S4. Equivalence of euclidean distances in geographic and auxiliary spaces

Let the survey region $\mathcal{A}$ be a bounded set of $\mathbb{R}^2$, and for simplicity, suppose that it coincides with the closure of its interior $int(\mathcal{A})$. The following result holds.

Proposition 1. *Let $\mathbf{p} \in int(\mathcal{A})$. If for each $l = 1, \ldots, G$ the auxiliary variable x_l is differentiable at $\mathbf{p}$ and the vector space generated by the gradients $\nabla x_l(\mathbf{p})$ is bidimensional, then there exist two real numbers $\gamma, \delta > 0$ such that for any $\mathbf{q} \in B_g(\mathbf{p}, \delta)$ it holds that*

$$d_g(\mathbf{p}, \mathbf{q}) \leq \gamma \max_{l=1,\ldots,G} |x_l(\mathbf{p}) - x_l(\mathbf{q})|. \tag{S4.1}$$

Proof. Denoting by $\nabla x_{l_1}(\mathbf{p})$ and $\nabla x_{l_2}(\mathbf{p})$ two linearly independent gradients, for each $\mathbf{q} \neq \mathbf{p}$ it holds that

$$\begin{aligned} \max_{l=1,\ldots,G} \frac{|\nabla^t x_l(\mathbf{p})(\mathbf{p} - \mathbf{q})|}{d_g(\mathbf{p}, \mathbf{q})} &\geq \max_{l=l_1,l_2} \frac{|\nabla^t x_l(\mathbf{p})(\mathbf{p} - \mathbf{q})|}{d_g(\mathbf{p}, \mathbf{q})} \\ &\geq \min_{\mathbf{v} \in \partial B_g(\mathbf{0},1)} \max_{l=l_1,l_2} |\nabla^t x_l(\mathbf{p})\mathbf{v}| = \lambda > 0 \end{aligned} \tag{S4.2}$$

where $\partial B_g(\mathbf{0},1)$ is the boundary of the $B_g(\mathbf{0},1)$ ball. Because the x_ls are differentiable at $\mathbf{p}$, it holds that

$$\begin{aligned} \max_{l=1,\dots,G} \frac{|x_l(\mathbf{q}) - x_l(\mathbf{p})|}{d_g(\mathbf{p},\mathbf{q})} &= \max_{l=1,\dots,G} \left| \frac{\nabla^t x_l(\mathbf{p})(\mathbf{p}-\mathbf{q})}{d_g(\mathbf{p}-\mathbf{q})} + o_l(1) \right| \\ &\geq \max_{l=l_1,l_2} \left| \frac{\nabla^t x_l(\mathbf{p})(\mathbf{p}-\mathbf{q})}{d_g(\mathbf{p}-\mathbf{q})} + o_l(1) \right| . \end{aligned} \tag{S4.3}$$

Therefore, if $\delta > 0$ is such that $|o_{l_1}(1)|+|o_{l_2}(1)| \leq \lambda/2$ for each $\mathbf{q} \in B_g(\mathbf{p},\delta)$, it holds that

$$\max_{l=l_1,l_2} \left| \frac{\nabla^t x_l(\mathbf{p})(\mathbf{p}-\mathbf{q})}{d_g(\mathbf{p},\mathbf{q})} + o_l(1) \right| \geq \frac{\lambda}{2}\ , \quad \mathbf{q} \in B_g(\mathbf{p},\delta), \mathbf{q} \neq \mathbf{p}. \tag{S4.4}$$

From the relationships (S4.2), (S4.3), and (S4.4), (S4.1) holds for $\gamma = 2/\lambda$. $\square$

Now suppose that $\mathbf{x}(\mathbf{p})$ is one-to-one on $\mathcal{A}$, a situation likely to occur for G sufficiently large (see Section S3). In this case, $d_x(\mathbf{p},\mathbf{q})$ is a distance. Moreover, because

$$\max_{l=1,\dots,G} |x_l(\mathbf{p}) - x_l(\mathbf{q}))| \leq d_x(\mathbf{p},\mathbf{q})$$

then for each $\mathbf{q} \in B_g(\mathbf{p},\delta)$, (S4.1) implies that

$$d_g(\mathbf{p},\mathbf{q}) \leq \gamma d_x(\mathbf{p},\mathbf{q}).$$

Then, owing the continuity of $\mathbf{x}(\mathbf{p})$ at $\mathbf{p}$, for each sequence $\{\mathbf{q}_m\} \in \mathcal{A}$ converging to $\mathbf{p}$, it holds that

$$\lim_{m\to\infty} d_g(\mathbf{p},\mathbf{q}_m) = 0 \iff \lim_{m\to\infty} d_g(\mathbf{p},\mathbf{q}_m) = 0$$

i.e., the Euclidean distances d_g and d_x are equivalent.

S5. Details on asymptotic scenarios

To achieve design consistency of the considered interpolators, the following asymptotic scenarios are needed.

In particular, in the context of continuous populations, the surface to be interpolated is fixed and a sequence of designs $\{D_m\}$ is considered for

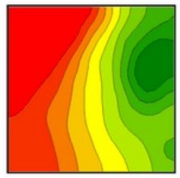

Figure 3: Surface to be interpolated.

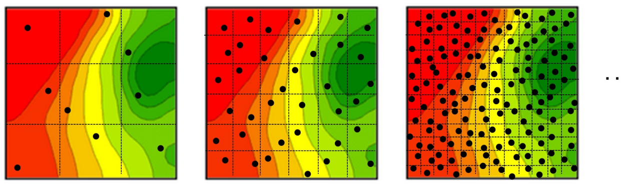

Figure 4: Example of 3 samples of the sequence, selected according to TSS

selecting a sample of n_m locations $\mathbf{P}_{m,1}, \ldots, \mathbf{P}_{m,n}$ on the study area, with $n_m \to \infty$ as m increases.

As an example, in Figure 3 a fixed surface to be interpolated is depicted. Considering a sequence of TSS designs, Figure 4 depicts an example of 3 possible selected samples of locations when considering $m = 1$ and the corresponding sample size $n_1 = 9$, $m = 2$ and the corresponding sample size $n_2 = 36$ and $m = 3$ and the corresponding sample size $n_3 = 144$.

When dealing with finite populations of areas, the study region is fixed but it is partitioned into an increasing number of N_m areas $a_{m,1}, \ldots, a_{m,N_m}$, with $N_m \to \infty$ as m increases and a sequence of designs $\{D_m\}$ is considered for selecting a sample of $n_m < N_m$ areas with $n_m \to \infty$ as m increases.

As an example, considering a sequence of OPSS designs, Figure 5 depicts an example of 3 possible partitions of the study region into the increasing number of $N_1 = 64$ ($m = 1$), $N_2 = 144$ ($m = 2$) and $N_3 = 256$ ($m = 3$) areas. The areas outlined in red represent the realization of a possible sequence of selected samples of areas of increasing size $n_1 = 16$ ($m = 1$), $n_2 = 36$ ($m = 2$) and $n_3 = 64$ ($m = 3$).

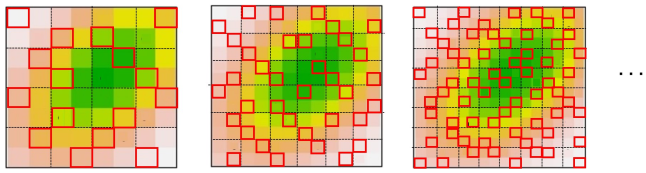

Figure 5: Example of 3 samples of the sequence, selected according to OPSS

References

Barabesi L., Franceschi S., Marcheselli M. (2012) Properties of design-based estimation under stratified spatial sampling with application to canopy coverage estimation. *Ann Appl Stat* 6:210–228

Fattorini L., Marcheselli M., Pisani C., Pratelli L. (2022) Design-based properties of the nearest neighbor spatial interpolator and its bootstrap mean squared error estimator. *Biometrics* 78:1454–1463

Table 1: Correlations between AGB and Landsat spectral bands values for the 30mx30m pixels covering the 27ha portion of Harvard forest.

Landsat spectral band	Correlation coefficient
Band 4	-0.67
Band 3	-0.63
Band 7	-0.57
Band 2	-0.45
Band 6	-0.38
Band 1	-0.28
Band 5	-0.11

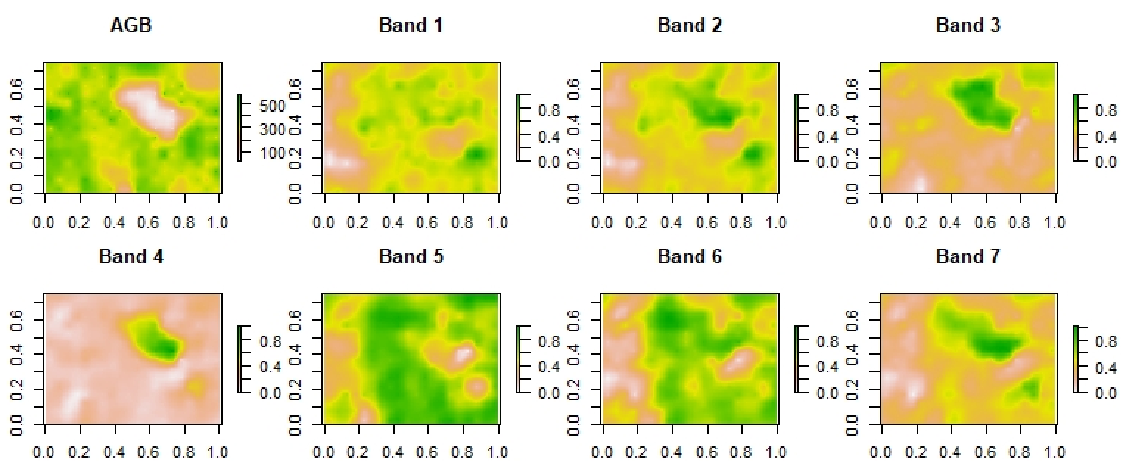

Figure 6: AGB and spectral bands values in the continuum of the 27ha portion of Harvard Forest.

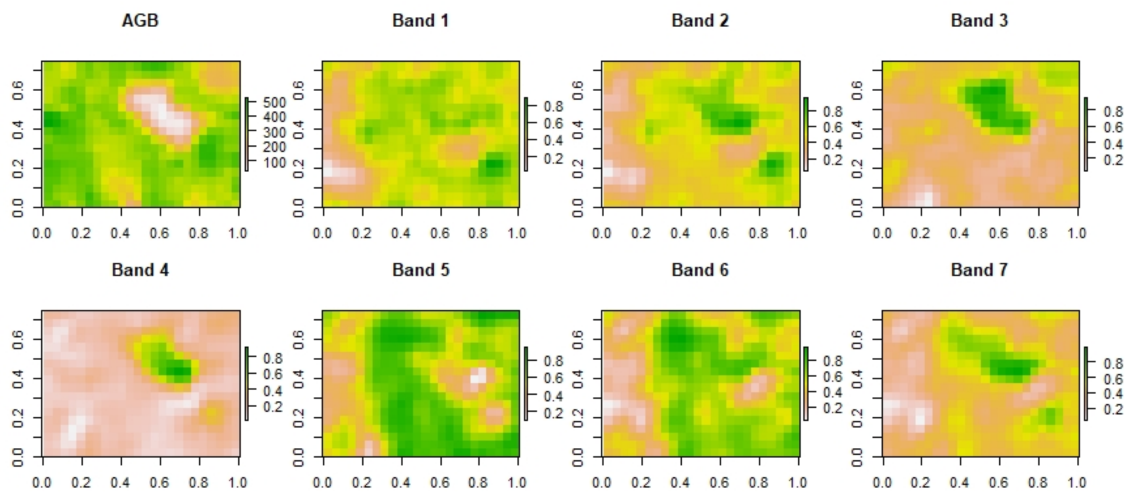

Figure 7: AGB and spectral bands values for the finite population of $N = 432$ areas.

Table 2: Continuous population. Values of ARMSE when kNN, RFI and IDW interpolations are performed by considering the geographic space (column 3), the auxiliary space with the first (column 4), the two (column 5), the three (column 6), the four (column 7), the five (column 8), the six (column 9) auxiliary variables more correlated with ABG and all the seven (column 10) auxiliary variables and the corresponding composite spaces (columns 11-17).

		Geographic space	Auxiliary space							Composite space						
	n		1	2	3	4	5	6	7	1	2	3	4	5	6	7
kNN	100	37.75	54.16	47.75	43.38	42.75	43.11	42.86	43.75	33.06	32.80	33.13	33.59	34.26	34.64	35.10
	200	27.67	53.82	45.46	40.16	38.26	38.58	38.16	38.06	25.43	26.18	26.81	27.10	27.69	28.00	28.10
	300	22.49	53.76	44.17	38.05	35.58	35.63	35.16	34.45	21.37	22.35	23.00	23.38	23.90	24.18	24.08
	400	19.05	53.75	43.33	36.43	33.61	33.38	32.82	31.73	18.45	19.52	20.27	20.69	21.23	21.42	21.28
	500	16.47	53.75	42.70	35.12	32.03	31.61	30.93	29.55	16.25	17.26	18.03	18.49	19.06	19.19	19.05
RFI	100	46.73	57.05	46.22	41.65	41.51	41.42	40.26	40.43	37.58	36.17	36.33	36.65	37.10	37.53	37.52
	200	36.66	56.88	44.08	38.40	37.22	37.10	35.73	35.71	30.53	30.16	30.47	30.31	30.91	31.26	31.26
	300	31.22	56.82	42.87	36.57	34.84	34.66	33.16	32.96	26.77	26.79	27.15	26.70	27.33	27.65	27.51
	400	27.62	56.81	42.03	35.21	33.16	32.90	31.24	30.88	24.12	24.37	24.75	24.07	24.68	25.00	24.72
	500	24.99	56.80	41.38	34.17	31.87	31.52	29.75	29.25	22.24	22.55	22.93	22.07	22.26	22.98	22.57
IDW	100	-	-	-	-	-	-	-	-	28.95	28.47	28.70	28.06	28.82	29.08	29.47
	200	-	-	-	-	-	-	-	-	22.42	22.07	22.18	21.63	22.17	22.36	22.52
	300	-	-	-	-	-	-	-	-	19.12	18.77	18.83	18.33	18.72	18.92	19.00
	400	-	-	-	-	-	-	-	-	16.58	16.29	16.36	15.88	16.22	16.40	16.44
	500	-	-	-	-	-	-	-	-	14.59	14.33	14.38	13.98	14.35	14.47	14.49

Table 3: Populations of areas. Values of ARMSE when kNN, RFI and IDW interpolations are performed by considering the geographic space (column 3), the auxiliary space with the first (column 4), the two (column 5), the three (column 6), the four (column 7), the five (column 8), the six (column 9) auxiliary variables more correlated with ABG and all the seven (column 10) auxiliary variables and the corresponding composite spaces (columns 11-17).

		Geographic space	Auxiliary space							Composite space						
	n		1	2	3	4	5	6	7	1	2	3	4	5	6	7
kNN	18×24	54.34	53.53	49.60	46.44	46.69	46.95	46.98	48.31	44.45	42.89	42.57	43.28	43.10	43.46	44.12
	27×36	39.87	51.15	45.62	41.60	41.34	41.60	41.43	42.52	33.36	32.61	32.74	33.38	33.90	34.34	34.86
	36×48	30.66	50.95	44.05	39.35	38.14	38.49	38.15	38.51	26.78	27.05	27.64	27.82	28.56	28.87	29.12
	45×60	24.55	50.99	42.72	37.27	35.21	35.47	35.02	34.71	22.15	23.02	23.64	23.94	24.42	24.71	24.75
	54×72	20.29	50.99	41.87	35.63	33.13	33.08	32.57	31.74	18.89	19.93	20.61	21.00	21.48	21.70	21.62
RFI	18×24	55.55	54.72	47.06	44.38	45.30	45.49	45.27	45.51	45.34	42.49	42.37	43.31	43.65	44.44	44.28
	27×36	46.19	53.76	44.35	40.36	40.62	40.53	39.64	39.84	37.20	35.74	35.84	36.56	36.94	37.46	37.47
	36×48	38.14	53.76	42.74	37.82	37.19	37.14	35.93	36.05	31.41	30.84	31.12	31.34	31.95	32.37	32.45
	45×60	32.23	53.81	41.46	35.94	34.69	34.56	33.22	33.15	27.23	27.15	27.54	27.46	28,13	28.48	28.51
	54×72	27.92	53.77	40.63	34.46	32.71	32.56	31.09	30.86	24.18	24.42	24.80	24.44	25.10	25.44	25.33
IDW	18×24	-	-	-	-	-	-	-	-	38.16	37.11	37.46	37.46	39.05	39.49	40.99
	27×36	-	-	-	-	-	-	-	-	30.10	29.23	28.80	28.80	29.69	30.00	30.51
	36×48	-	-	-	-	-	-	-	-	24.01	23.50	23.04	23.04	23.63	23.96	24.17
	45×60	-	-	-	-	-	-	-	-	19.88	19.48	19.01	19.01	19.47	19.75	19.84
	54×72	-	-	-	-	-	-	-	-	17.11	16.80	16.40	16.40	16.81	16.99	17.04