

## A ESTIMATORS OF INTRINSIC DIMENSION

### A.1 DISTANCE-BASED ESTIMATORS

**MLE (Levina & Bickel, 2005)** Let  $z \in \mathbb{R}^D$  be an embedding (feature vector) and denote by radii  $R_1(z) \leq \dots \leq R_k(z)$  the Euclidean distances to its  $k$  nearest neighbors. A broad family of intrinsic-dimension estimators rests on the assumption that, in a sufficiently small neighbourhood of  $z$ , samples are drawn from an approximately uniform Poisson process on a  $d$ -dimensional manifold. Under this hypothesis one can write down a closed-form density for the joint distribution of the radii, which depends on  $d$  only through the volume of a unit  $d$ -ball. The *Levina–Bickel MLE* (Levina & Bickel, 2005) estimator leverages this fact to define a local estimator

$$\hat{d}_{\text{MLE}}(z) = \left[ \frac{1}{k-1} \sum_{j=1}^{k-1} \ln \frac{R_k(z)}{R_j(z)} \right]^{-1},$$

whose global counterpart pools the local estimates using its harmonic mean across embeddings  $z_i$ ,

$$\hat{d}_{\text{global}} = \left( \frac{1}{n} \sum_{i=1}^n \frac{1}{\hat{d}_{\text{MLE}}(z_i)} \right)^{-1} = \frac{n}{\sum_{i=1}^n \left[ \frac{1}{k-1} \sum_{j=1}^{k-1} \ln \frac{R_k(z_i)}{R_j(z_i)} \right]}.$$

**MOM (Amsaleg et al., 2018)** Under the same Poisson–ball model, conditioning on the outer radius  $R_k(z)$  makes the inner neighbors approximately uniform inside the  $d$ -ball of radius  $R_k(z)$ . In higher dimension more mass concentrates near the boundary, so the average inner radius moves closer to  $R_k(z)$ . Let

$$\bar{R}(z) = \frac{1}{k} \sum_{j=1}^k R_j(z).$$

The method of moments estimator equates the empirical ratio  $\bar{R}(z)/R_k(z)$  to its model value and solves for  $d$ , which yields the local estimator

$$\hat{d}_{\text{MOM}}(z) = \frac{\bar{R}(z)}{R_k(z) - \bar{R}(z)}.$$

Large  $\hat{d}_{\text{MOM}}(z)$  therefore corresponds to neighborhoods where most points lie close to the outer shell.

**TLE (Amsaleg et al., 2019).** Tight Local Estimation keeps the Levina–Bickel likelihood identity but extracts logarithmic spacing information from the full  $k$ -NN configuration, not only the ordered radii. Inside the ball of radius  $R_k(z)$  it constructs pairwise geometric surrogates  $S_{ij}(z)$  and  $T_{ij}(z) \in (0, R_k(z)]$  that summarize how neighbours are positioned relative to one another. Degenerate or near-zero terms are discarded by a fixed threshold to stabilize the estimate. Schematically, the MLE log-ratios  $\{\ln(R_k/R_j)\}$  are replaced by  $\{\ln(R_k/S_{ij}), \ln(R_k/T_{ij})\}$  and combined through the same harmonic-mean aggregation to produce  $\hat{d}_{\text{TLE}}(z)$ . The result is less sensitive to very small spacings while remaining driven by how tightly the neighborhood fills the  $d$ -ball.

**ESS (Johnsson et al., 2014).** Expected Simplex Skewness uses the shape of the neighbor constellation rather than only radial spacings. After mean-centering the  $k$ -NN patch at  $z$ , one forms vectors to neighbors and builds many small simplices from these vectors. A normalized “skewness” statistic  $s(z)$  combines their volumes and edge lengths. For a uniform set of finite-neighbor samples around  $z$  in a  $d$ -ball this statistic concentrates around a dimension-specific reference curve. The local estimate  $\hat{d}_{\text{ESS}}(z)$  is obtained by inverting the observed  $s(z)$  against that curve, using linear interpolation between consecutive integer

dimensions. Because ESS is driven mainly by relative geometry and angles, it tends to be less sensitive to mild density gradients while still reflecting curvature and noise through systematic changes in the neighbor shape.

## A.2 ANGLE-BASED ESTIMATORS

We provide a detailed explanation of the angle-based FisherS estimator below (Albergante et al., 2019).

Let  $\{z_i\}_{i=1}^n \subset \mathbb{R}^D$  denote embeddings. The FisherS estimator starts from a concentration-of-measure assumption: in high dimensions, points sampled uniformly on a sphere tend to be almost orthogonal, so a typical point can be linearly separated from the rest by a Fisher discriminant<sup>1</sup>. FisherS quantifies how separable each point is from all others at a similarity margin  $\alpha \in (0, 1)$ , then inverts a closed-form reference curve to recover an effective dimension.

Following Albergante et al. (2019), inputs are standardized to make the separability model comparable with a uniform  $n$ -sphere: (i) center  $Z$  by subtracting the sample mean; (ii) reduce dimension using PCA, retaining components with eigenvalues at least  $\lambda_{\max}/C$  (default  $C = 10$ ), which removes ill-conditioned directions; (iii) whiten the retained coordinates so the covariance is the identity; and (iv) project each vector to the unit sphere,  $x_i := z_i / \|z_i\|$ . These steps yield an array  $X \in \mathbb{R}^{n \times d}$  with rows  $x_i$ .

For each  $\alpha$  on a grid, we form the normalized Gram matrix  $G = XX^\top$  and test, for every pair  $(i, j)$  with  $i \neq j$ , whether

$$\langle x_i, x_j \rangle > \alpha \langle x_i, x_i \rangle.$$

Since  $\|x_i\| = 1$ , this reduces to  $\langle x_i, x_j \rangle > \alpha$ , that is, whether the cosine similarity exceeds  $\alpha$ . Point  $x_i$  is called *inseparable* at level  $\alpha$  if it has at least one neighbor whose similarity crosses this threshold. Empirically we record, for each  $i$ , the fraction

$$p_i(\alpha) = \frac{1}{n} \#\{j \neq i : \langle x_i, x_j \rangle > \alpha\}$$

and then average over points to obtain  $\bar{p}(\alpha) = \frac{1}{n} \sum_i p_i(\alpha)$ .

If points are sampled uniformly on the unit  $n$ -sphere, the mean inseparability probability at margin  $\alpha$  is well-approximated by

$$\bar{p}_{\text{sphere}}(\alpha; n) \approx \frac{(1 - \alpha^2)^{(n-1)/2}}{\alpha \sqrt{2\pi n}}.$$

Given the measured  $\bar{p}(\alpha)$ , FisherS solves for  $n$  by inverting this expression with the Lambert  $W$  function:

$$\hat{n}(\alpha) = \frac{W\left(-\frac{\ln(1 - \alpha^2)}{2\pi \bar{p}(\alpha)^2 \alpha^2 (1 - \alpha^2)}\right)}{-\ln(1 - \alpha^2)}.$$

Varying  $\alpha$  yields a profile  $\alpha \mapsto \hat{n}(\alpha)$ ; FisherS then picks a single global estimate by selecting the largest  $\alpha$  for which  $\bar{p}(\alpha) > 0$  (i.e., not all points are fully separable) and reporting  $\hat{n}(\alpha^*)$  for  $\alpha^* \approx 0.9 \max\{\alpha : \bar{p}(\alpha) > 0\}$ . The method also supports pointwise ID by plugging  $p_i(\alpha)$  into the same inversion at a chosen  $\alpha$ .

FisherS does not rely on nearest-neighbor radii. It asks instead: “At a given angular margin  $\alpha$ , how often do points fail to be linearly separable from the rest?” On a true  $n$ -sphere, this failure rate decays at a rate governed by  $n$ . If the dataset is more clustered or locally low-dimensional, inseparability persists to larger

<sup>1</sup>Fisher’s linear discriminant is the single direction (a line) that best separates two groups by making their projected means far apart relative to their spread.



Table 2: **Global intrinsic dimensions of geographic INRs derived from SatCLIP (Klemmer et al., 2025), GeoCLIP (Cepeda et al., 2023), and CSP (Mai et al., 2023a)** For all distance-based estimators, we use  $k = 20$  nearest neighbors. Results shown for 6 sampling schemes, estimated by six total methods. Variation shows as  $1 \times$  standard deviation over 3 sub-sampling runs containing  $N = 50,000$  points each.

Scheme	Model	CorrInt	FisherS	MLE	MOM	TLE	TwoNN
Fibonacci	SatCLIP-L10	$2.16 \pm 0.02$	$5.94 \pm 0.00$	$2.45 \pm 0.05$	$2.15 \pm 0.00$	$2.43 \pm 0.02$	<b><math>44.14 \pm 0.80</math></b>
	SatCLIP-L40	$2.36 \pm 0.02$	$9.95 \pm 0.03$	$2.58 \pm 0.05$	$4.68 \pm 0.10$	$2.78 \pm 0.03$	$12.80 \pm 0.30$
	<b>GeoCLIP</b>	<b><math>22.70 \pm 0.05</math></b>	<b><math>11.16 \pm 0.10</math></b>	<b><math>13.17 \pm 0.10</math></b>	<b><math>23.70 \pm 0.30</math></b>	<b><math>13.20 \pm 0.20</math></b>	<b><math>24.15 \pm 0.08</math></b>
	CSP-FMoW	$2.50 \pm 0.03$	$1.87 \pm 0.01$	$6.28 \pm 0.10$	$5.07 \pm 0.08$	$7.02 \pm 0.12$	$9.43 \pm 0.20$
	CSP-iNat	$5.65 \pm 0.06$	$0.92 \pm 0.01$	$4.96 \pm 0.08$	$6.34 \pm 0.10$	$5.90 \pm 0.10$	$5.07 \pm 0.15$
Sphere	SatCLIP-L10	$2.01 \pm 0.02$	$5.94 \pm 0.03$	$2.01 \pm 0.02$	$2.12 \pm 0.02$	$2.22 \pm 0.02$	$2.00 \pm 0.06$
	SatCLIP-L40	$2.18 \pm 0.02$	$9.89 \pm 0.03$	$2.14 \pm 0.03$	$4.51 \pm 0.10$	$2.52 \pm 0.02$	$2.02 \pm 0.06$
	<b>GeoCLIP</b>	<b><math>20.79 \pm 0.04</math></b>	<b><math>11.10 \pm 0.01</math></b>	<b><math>11.50 \pm 0.02</math></b>	<b><math>22.48 \pm 0.03</math></b>	<b><math>12.08 \pm 0.02</math></b>	<b><math>15.07 \pm 0.08</math></b>
	CSP-FMoW	$2.50 \pm 0.03$	$1.87 \pm 0.01$	$5.43 \pm 0.06$	$4.96 \pm 0.06$	$6.33 \pm 0.10$	$4.53 \pm 0.15$
	CSP-iNat	$5.58 \pm 0.05$	$0.92 \pm 0.01$	$4.27 \pm 0.05$	$6.15 \pm 0.08$	$5.31 \pm 0.08$	$3.19 \pm 0.10$
Land	SatCLIP-L10	$1.96 \pm 0.00$	$5.00 \pm 0.01$	$1.96 \pm 0.00$	$2.02 \pm 0.00$	$2.16 \pm 0.00$	$1.98 \pm 0.01$
	SatCLIP-L40	$2.05 \pm 0.00$	<b><math>8.08 \pm 0.00</math></b>	$2.03 \pm 0.00$	$2.39 \pm 0.00$	$2.32 \pm 0.00$	$1.99 \pm 0.01$
	<b>GeoCLIP</b>	<b><math>12.07 \pm 0.05</math></b>	$7.68 \pm 0.10$	<b><math>11.22 \pm 0.10</math></b>	<b><math>13.02 \pm 0.20</math></b>	<b><math>11.53 \pm 0.15</math></b>	<b><math>11.25 \pm 0.50</math></b>
	CSP-FMoW	$2.16 \pm 0.03$	$1.70 \pm 0.01$	$5.18 \pm 0.08$	$5.23 \pm 0.08$	$6.25 \pm 0.10$	$3.05 \pm 0.12$
	CSP-iNat	$3.93 \pm 0.05$	$0.92 \pm 0.01$	$3.37 \pm 0.06$	$4.65 \pm 0.08$	$4.14 \pm 0.08$	$2.70 \pm 0.10$
Grid	SatCLIP-L10	$1.64 \pm 0.02$	$4.45 \pm 0.00$	$2.00 \pm 0.00$	$2.15 \pm 0.00$	$2.22 \pm 0.00$	$2.61 \pm 0.01$
	SatCLIP-L40	$1.72 \pm 0.02$	<b><math>8.76 \pm 0.02</math></b>	$2.15 \pm 0.00$	$3.40 \pm 0.00$	$2.56 \pm 0.00$	$2.58 \pm 0.03$
	<b>GeoCLIP</b>	<b><math>18.87 \pm 0.08</math></b>	$7.08 \pm 0.01$	<b><math>13.00 \pm 0.00</math></b>	<b><math>22.22 \pm 0.00</math></b>	<b><math>13.55 \pm 0.01</math></b>	<b><math>15.42 \pm 0.10</math></b>
	CSP-FMoW	$2.65 \pm 0.03$	$1.81 \pm 0.01$	$5.69 \pm 0.08$	$5.25 \pm 0.08$	$6.56 \pm 0.10$	$5.86 \pm 0.20$
	CSP-iNat	$5.57 \pm 0.06$	$0.92 \pm 0.01$	$4.64 \pm 0.08$	$6.32 \pm 0.10$	$5.63 \pm 0.10$	$3.68 \pm 0.10$
Naive	SatCLIP-L10	$1.76 \pm 0.02$	$4.47 \pm 0.03$	$2.01 \pm 0.02$	$2.12 \pm 0.02$	$2.22 \pm 0.02$	$2.01 \pm 0.08$
	SatCLIP-L40	$1.95 \pm 0.02$	<b><math>8.93 \pm 0.03</math></b>	$2.15 \pm 0.03$	$4.53 \pm 0.10$	$2.54 \pm 0.02$	$2.02 \pm 0.08$
	<b>GeoCLIP</b>	<b><math>18.02 \pm 0.08</math></b>	$7.09 \pm 0.00$	<b><math>12.08 \pm 0.00</math></b>	<b><math>21.53 \pm 0.03</math></b>	<b><math>12.76 \pm 0.02</math></b>	<b><math>13.63 \pm 0.07</math></b>
	CSP-FMoW	$2.68 \pm 0.03$	$1.81 \pm 0.01$	$5.60 \pm 0.08$	$5.26 \pm 0.08$	$6.53 \pm 0.10$	$4.86 \pm 0.15$
	CSP-iNat	$5.51 \pm 0.05$	$0.92 \pm 0.01$	$4.41 \pm 0.08$	$6.24 \pm 0.10$	$5.45 \pm 0.10$	$3.20 \pm 0.10$
Stratified	SatCLIP-L10	$1.75 \pm 0.01$	$4.47 \pm 0.00$	$2.01 \pm 0.00$	$2.26 \pm 0.00$	$2.22 \pm 0.00$	$2.00 \pm 0.01$
	SatCLIP-L40	$1.92 \pm 0.01$	$8.91 \pm 0.01$	$2.15 \pm 0.00$	$2.49 \pm 0.00$	$2.54 \pm 0.00$	$2.01 \pm 0.01$
	<b>GeoCLIP</b>	<b><math>17.93 \pm 0.08</math></b>	<b><math>7.07 \pm 0.01</math></b>	<b><math>12.06 \pm 0.10</math></b>	<b><math>21.53 \pm 0.03</math></b>	<b><math>12.75 \pm 0.02</math></b>	<b><math>13.58 \pm 0.08</math></b>
	CSP-FMoW	$2.67 \pm 0.03$	$1.81 \pm 0.01$	$5.59 \pm 0.08$	$5.25 \pm 0.08$	$6.53 \pm 0.10$	$4.79 \pm 0.15$
	CSP-iNat	$5.51 \pm 0.05$	$0.92 \pm 0.01$	$4.42 \pm 0.08$	$6.24 \pm 0.10$	$5.45 \pm 0.10$	$3.22 \pm 0.10$

$\alpha$ , inflating  $\bar{p}(\alpha)$  and thus lowering the inferred  $n$ ; if the dataset spreads more uniformly,  $\bar{p}(\alpha)$  drops and the inferred  $n$  rises. The PCA+whitening+sphere projection makes this comparison meaningful by matching the spherical reference.

## B HOW DO THE NUMBER OF NEAREST NEIGHBORS CHANGE GEOGRAPHIC INR ID?

In distance-based estimators, the neighbor count  $k$  controls a bias–variance tradeoff. With small  $k$ , the estimate is anchored to an extremely local neighborhood of  $z$ , so it is sensitive to the true local geometry and density but suffers high variance because it averages (harmonic mean in (Levina & Bickel, 2005) and arithmetic mean in (Amsaleg et al., 2019; 2018)) over a few nearest neighbors. As  $k$  increases, variance

Table 3: Global geographic INR ID versus number of nearest neighbors ( $k$ ) for distance-based estimators on **scikit-dimension**. 100,000 points sampled with land, fibonacci, and spherical sampling.

(a) MLE (Levina & Bickel, 2005)								(b) MOM (Amsaleg et al., 2018)							
Sampling	Encoder	k						Sampling	Encoder	k					
		5	10	20	50	100	200			5	10	20	50	100	200
land	GeoCLIP	12.781	12.690	11.234	10.327	11.935	15.090	land	GeoCLIP	25.754	17.070	12.988	11.282	13.015	16.269
	CSP-FMoW	4.050	4.770	5.169	5.109	4.736	4.245		CSP-FMoW	8.823	7.037	6.440	5.802	5.217	4.567
	CSP-iNat	2.940	3.151	3.375	3.691	3.930	4.095		CSP-iNat	5.752	4.493	4.250	4.396	4.643	4.886
	SatCLIP-L10	1.982	1.969	1.960	1.952	1.951	1.963		SatCLIP-L10	3.389	2.511	2.211	2.058	2.017	2.017
	SatCLIP-L40	1.993	2.004	2.024	2.099	2.225	2.497		SatCLIP-L40	3.407	2.565	2.296	2.248	2.392	2.804
fibonacci	GeoCLIP	17.070	13.041	13.173	17.795	22.188	25.560	fibonacci	GeoCLIP	26.250	15.685	14.756	19.302	23.700	27.114
	CSP-FMoW	8.027	7.087	6.280	5.352	4.749	4.231		CSP-FMoW	13.145	8.991	7.191	5.808	5.069	4.476
	CSP-iNat	4.787	4.792	4.964	5.269	5.457	5.588		CSP-iNat	8.372	6.484	6.100	6.196	6.340	6.411
	SatCLIP-L10	4.480	2.610	2.445	2.207	2.161	2.170		SatCLIP-L10	7.298	2.799	2.496	2.191	2.151	2.178
	SatCLIP-L40	4.747	2.765	2.576	2.847	3.852	5.497		SatCLIP-L40	7.389	2.968	2.649	3.134	4.683	6.989
sphere	GeoCLIP	12.301	11.047	11.494	15.979	20.590	24.393	sphere	GeoCLIP	20.905	14.051	13.451	17.942	22.484	26.187
	CSP-FMoW	5.158	5.483	5.432	5.016	4.591	4.151		CSP-FMoW	10.176	7.579	6.490	5.553	4.959	4.423
	CSP-iNat	3.480	3.848	4.272	4.851	5.208	5.443		CSP-iNat	6.980	5.651	5.541	5.877	6.147	6.306
	SatCLIP-L10	2.007	2.009	2.013	2.032	2.058	2.110		SatCLIP-L10	3.411	2.551	2.262	2.133	2.120	2.161
	SatCLIP-L40	2.043	2.072	2.140	2.553	3.529	5.135		SatCLIP-L40	3.508	2.658	2.453	3.006	4.522	6.802
(c) TLE (Amsaleg et al., 2019)								(d) ESS (Johnsson et al., 2014)							
Sampling	Encoder	k						Sampling	Encoder	k					
		5	10	20	50	100	200			5	10	20	50	100	200
land	GeoCLIP	21.665	14.764	11.547	10.218	11.669	14.278	land	GeoCLIP	12.043	25.661	29.229	30.517	37.955	50.230
	CSP-FMoW	7.851	6.630	6.236	5.720	5.197	4.602		CSP-FMoW	5.621	7.107	7.523	7.239	6.570	5.718
	CSP-iNat	5.070	4.259	4.149	4.326	4.551	4.753		CSP-iNat	3.860	4.192	4.414	4.781	5.108	5.418
	SatCLIP-L10	2.786	2.319	2.155	2.085	2.084	2.117		SatCLIP-L10	2.148	2.097	2.072	2.097	2.166	2.290
	SatCLIP-L40	2.892	2.446	2.319	2.355	2.523	2.923		SatCLIP-L40	2.222	2.249	2.356	2.632	2.970	3.682
fibonacci	GeoCLIP	23.199	14.034	13.196	16.840	20.173	22.568	fibonacci	GeoCLIP	12.031	25.870	39.545	63.910	84.450	98.897
	CSP-FMoW	12.282	8.618	7.023	5.777	5.100	4.546		CSP-FMoW	7.524	9.148	8.400	7.033	6.123	5.344
	CSP-iNat	7.756	6.209	5.898	5.962	6.049	6.057		CSP-iNat	5.158	5.878	6.179	6.590	6.848	7.005
	SatCLIP-L10	3.937	2.498	2.427	2.273	2.279	2.340		SatCLIP-L10	2.726	2.326	2.293	2.326	2.438	2.622
	SatCLIP-L40	5.858	2.970	2.784	3.305	4.779	6.884		SatCLIP-L40	2.852	2.753	2.868	4.075	6.557	10.122
sphere	GeoCLIP	18.149	12.514	12.064	15.750	19.257	21.906	sphere	GeoCLIP	11.801	24.231	34.033	54.676	74.804	91.335
	CSP-FMoW	9.241	7.206	6.330	5.521	4.985	4.488		CSP-FMoW	6.269	7.630	7.492	6.698	5.987	5.292
	CSP-iNat	6.180	5.301	5.305	5.637	5.856	5.952		CSP-iNat	4.438	5.116	5.596	6.236	6.627	6.881
	SatCLIP-L10	2.827	2.372	2.224	2.188	2.225	2.306		SatCLIP-L10	2.168	2.138	2.145	2.250	2.398	2.601
	SatCLIP-L40	3.039	2.586	2.519	3.106	4.532	6.607		SatCLIP-L40	2.320	2.443	2.686	3.809	6.064	9.429

decreases by pooling more radii. The effective neighborhood increases and begins to mix in curvature, density gradients, and boundary effects. These departures from local homogeneity induce bias whose sign and magnitude depend on the estimator. For the Levina–Bickel MLE estimator specifically, which is derived under a local homogeneous Poisson model, the estimator is approximately unbiased when the model holds and  $k$  is small to moderate, and its variance scales like  $\approx d^2/(k-3)$  for fixed intrinsic dimension  $d$  (Levina & Bickel, 2005); pushing  $k$  larger suppresses variance further but also violates the locality assumptions, allowing global structure to pull the estimate upward or downward.

Across estimators, increasing  $k$  changes the balance between local detail and global structure, and our results reflect this clearly. ESS grows rapidly with  $k$ , most strikingly for GeoCLIP: under fibonacci sampling it rises from about 12 at  $k=5$  to nearly 99 at  $k=200$ . This pattern is consistent with ESS aggregating more directions and longer chords as the neighborhood widens, which inflates the measured simplex skewness and maps to a higher effective dimension. MOM and TLE typically drop from  $k=5$  to the  $k \approx 20$ –50 range, then flatten or rebound slightly. The initial decrease suggests that adding a few more neighbors stabilizes

local spacing statistics and trims extreme ratios, while very large neighborhoods begin to mix curvature and density variation, reintroducing upward bias. MLE is comparatively stable. For GeoCLIP it is often flat to gently increasing with  $k$  and occasionally U-shaped on land sampling. For SatCLIP encoders all distance-based estimators are remarkably steady, hovering near 2–3 across  $k$ , which indicates a tightly regular local geometry in those embeddings. Sampling also matters: fibonacci and spherical layouts amplify sensitivity to  $k$  for GeoCLIP and the CSP encoders, whereas land sampling tends to mute it. Overall, the table shows that methods that directly emphasize extreme interpoint ratios or long-range chords move the most with  $k$ , while estimators that average log-radii more evenly are less affected.

## C EXPERIMENT DETAILS: INTRINSIC DIMENSION OF GEOSPATIAL IMAGE ENCODERS

For all image encoders displayed on Table 1, we compute intrinsic dimension (ID) on features extracted from the S2-100K Sentinel-2 multispectral image dataset (Klemmer et al., 2025). Models that accept multispectral inputs use all 13 Sentinel-2 bands; models that only support RGB receive bands B4/B3/B2 in that order. All inputs are resized/cropped to  $224 \times 224$  pixels unless noted, and RGB models use standard ImageNet normalization. We estimate global ID from the resulting feature matrices using multiple estimators as in Table 1 (for distance-based methods,  $k = 20$  nearest neighbors). Below, we list key experimental details for each image encoder.

**Random Convolutional Filters (RCF) (Rolf et al., 2021).** We use training-free RCFs as a simple baseline: 13-channel inputs are convolved with a random filter bank, passed through a nonlinearity, and spatially pooled into a fixed-length descriptor. No external weights are loaded. The feature dimension is set to 512 for accurate comparison with larger image encoders.

**CROMA (Fuller et al., 2023).** We evaluate the optical branch of CROMA’s multimodal encoder (pre-trained on Sentinel-1/2 pairs). Inputs are the 12 optical Sentinel-2 bands (with an excluded B10 cirrus band), resized to  $120 \times 120$  pixels, and normalized to  $[0, 1]$  following official pre-trained model usage guidelines. Intrinsic dimension is calculated on the pooled optical embedding produced by the pre-trained encoder weights.

**DOFA (Xiong et al., 2024).** DOFA is a masked-autoencoder foundation model that conditions its patch embedding on channel wavelengths, enabling a single model to handle arbitrary band sets. We pass the 13 Sentinel-2 central wavelengths with the 13-band inputs and extract encoder features from the publicly released pre-trained weights.

**ScaleMAE (Reed et al., 2023).** ScaleMAE is a scale-aware masked autoencoder trained on RGB imagery with ground-sampling-distance cues and a multiscale reconstruction objective. We feed RGB tiles (B4/B3/B2), apply ImageNet normalization, and use the pre-trained encoder’s feature output.

**ResNet baselines (He et al., 2016).** For multispectral ResNets, we use 13-band Sentinel-2 MoCo weights (ResNet-18/50) and extract the global-average-pooled penultimate features. Pre-trained Sentinel-2 weights are loaded with TorchGeo package (Stewart et al., 2025). For RGB variants, we use the corresponding Sentinel-2 RGB MoCo weights with B4/B3/B2 inputs. For ResNet-152, we rely on ImageNet-1K pre-trained weights (RGB only) and take the pooled backbone features.

**ViT-Small (multispectral) (Dosovitskiy et al., 2021).** We use the Sentinel-2 MoCo-pre-trained ViT-Small that accepts 13 bands. We run a forward pass to obtain patch tokens and mean-pool them (excluding any prefix/CLS token) to form a single feature vector per tile.

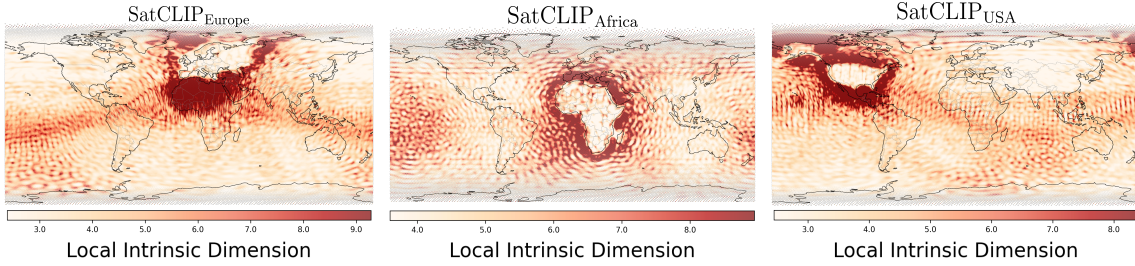


Figure 7: **Local intrinsic dimension can identify pre-training dataset coverage of local models.** We visualize local ID maps with the MLE estimator using  $k = 100$  nearest neighbors. 100,000 geographic coordinates are sampled with the fibonacci sampling scheme.

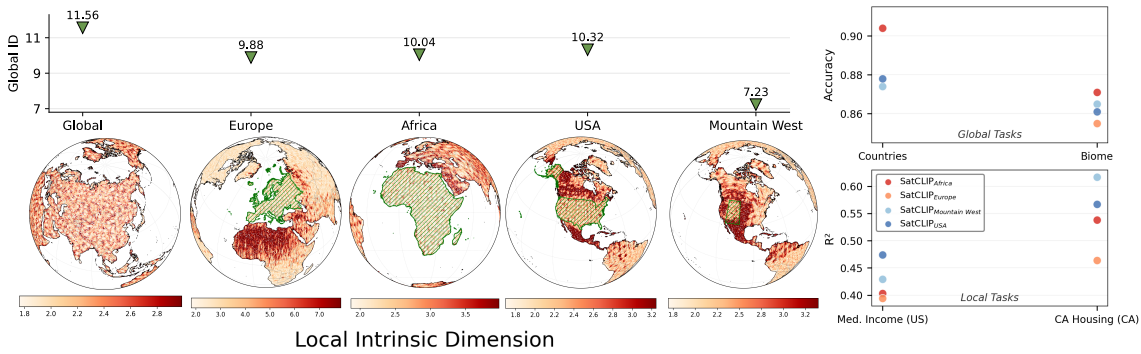


Figure 8: **Effect of pre-training dataset coverage on local and global geographic INR ID:** We generate several variants of the S2-100K datasets constrained to (i) continent-scale, (ii) country-scale, and (iii) region-scale boundaries, and train SatCLIP with  $L = 40$  Legendre polynomials on each regional dataset. (Top) Global geographic INR ID for each regional location encoder. (Bottom) Local ID with the MLE estimator using  $k = 100$  neighbors. Marked green outlines indicate spatial coverage of pre-training dataset. (Right) Performance of regional SatCLIP location encoders on (top) global and (bottom) local evaluation tasks.

## D EXPERIMENT DETAILS: EFFECT OF SPATIAL RESOLUTION AND SPATIAL COVERAGE ON GEOGRAPHIC INR ID

We detail our experimental framework that measures changes in the geographic INR ID with changing spatial resolution of the location encoder. We also include an additional experiment where we show the relationship between local and global geographic INR ID and spatial coverage of the pre-training data.

### D.1 MODIFYING THE SPATIAL RESOLUTION OF GEOGRAPHIC LOCATION ENCODERS

We study the effect of increasing spatial resolution in SatCLIP on ID by pre-training variants that differ only in the spherical-harmonic band-limit  $L$  (higher  $L$  allows finer spatial variation), while holding the location-embedding dimension and all other hyperparameters fixed. After pre-training, we compute location embeddings for  $N=100,000$  *land-only* coordinates (uniform on  $S^2$  then restricted to land) and report the global intrinsic dimension using the FisherS estimator.

GeoCLIP aligns images with GPS using a CLIP-style contrastive objective and a continuous location encoder built from hierarchical random Fourier features (RFF) (Cepeda et al., 2023). As the public codebase provides inference and pre-trained models rather than end-to-end pre-training,<sup>2</sup> we fine-tune *only* the location encoder on an image-to-GPS retrieval loss, freezing the CLIP image tower and the temperature (logit scale), with early stopping on validation loss (patience= 20 epochs). To modulate spatial resolution we (i) densify the RFF hierarchy by increasing the number of log-spaced frequency levels  $M$  while holding  $\sigma_{\min}, \sigma_{\max}$  fixed, and (ii) append one or two higher-frequency RFF branches to extend  $\sigma_{\max}$  while keeping existing branches frozen. Note that In the hierarchy-level-sweep wwith  $M$ , we instantiate the location encoder from scratch for each  $M$  (no warm-start), avoiding inherited bias from a previous hierarchy layout. When adding higher RFF frequency branches, however, we copy weights for overlapping branches and freeze them, so any ID change can be attributed to the new high-frequency branches. After fine-tuning, we estimate global intrinsic dimension with the FisherS estimator on embeddings of uniformly sampled land-only coordinates, and report all results under this land-only sampling.

We study the effect of increased spatial resolution on ID in Sphere2Vec (Mai et al., 2023b) and Space2Vec (Mai et al., 2020) by modifying the multi-scale resolution parameter  $S$ . All encoders are instantiated with trained weights from the MOSAICS nightlights task (Rolf et al., 2021). For **Sphere2Vec**, we modulate spatial resolution by sweeping the number of scales  $S$  (TorchSpatial: `freq`)— $S \in \{2, 4, 8, 16, 32\}$ . We keep the output width fixed (`spa_embed_dim`=256). We apply this to SPHEREM, SPHEREM+, SPHEREC, and SPHEREC+. Increasing  $S$  (and decreasing `min_radius`) adds higher-frequency components, yielding finer spatial detail. For **Space2Vec** we control resolution via either the *grid* sampling density (e.g.,  $W \times H \in \{(180, 90), (360, 180), (720, 360), (1440, 720)\}$ ) or the number of theory frequencies  $S \in \{2, 4, 8, 16, 32\}$ . We sample  $N=100,000$  coordinates uniformly on  $S^2$ , compute embeddings (deduplicating exact repeats), and estimate *global* intrinsic dimension with the FisherS estimator. We report ID as a function of the resolution knob for each encoder family in Figure 5.

## D.2 ADDITIONAL EXPERIMENT: MODIFYING THE SPATIAL COVERAGE OF PRE-TRAINING DATA

Geospatial models with strong global results can still underperform in specific regions; recent studies show that locally trained models or locally fine-tuned variants can match or exceed the accuracy of globally trained systems in the same area, underscoring a local–global divide in practice (Rolf et al., 2024). Motivated by this, we ask how the *spatial coverage* of pre-training data shapes the geometry of location encoders as seen through their global and local IDs. We construct region–constrained pre-training corpora (continent, country, sub-region) from Sentinel-2 using the Google Earth Engine API and train SatCLIP location encoders with fixed capacity and optimization budget, varying only the geographic extent. We then estimate a global FisherS ID on embeddings computed on a global scale and a *local* point-wise ID (MLE) that we visualize on orthographic maps.

To study how pre-training *coverage* shapes geographic INRs, we build region–specific Sentinel-2 datasets and pre-train regional SatCLIP models, then measure the ID of location embeddings derived from each local encoder. Regions are defined by Natural Earth polygons; for each region we sample point locations (random or stratified; cached for determinism) and retrieve COPERNICUS/S2\_SR\_HARMONIZED imagery via Google Earth Engine, selecting the least-cloudy scene in a fixed window between 2023-01-01–2024-12-31. Regions we create local datasets for along with the number of Sentinel-2 tiles queried are displayed in Table 4.

Table 4: Regional Sentinel-2 datasets used for local SatCLIP pre-training (tiles at 10 m,  $256 \times 256$ ).

Region	# Samples
United States	100,000
Europe	100,000
Asia	100,000
Africa	100,000
France	50,000
Mountain West	50,000
Denver Metro	20,000

<sup>2</sup><https://github.com/VicenteVivan/geo-clip>

Around each point we extract a multi-band  $256 \times 256$  tile at 10 m, stack bands  $\{B_1, \dots, B_{12}\}$  into a single GeoTIFF, and resample if needed to preserve exact size. Per region we release the images, an `index.csv` of (lon, lat), and the scripts used to generate them. For regional pre-training, we keep the SatCLIP pre-training hyperparameters fixed and set the spherical-harmonic band-limit to  $L=40$  (location-embedding dimension and all other hyperparameters held constant). The only factor that varies across runs is the pre-training dataset. For ID evaluation, we feed the trained location encoder (no classifier) with  $N$  coordinates sampled globally and report global FisherS ID.

From Figures 7 and 8, as pre-training coverage narrows from global to continental to regional, the global ID decreases. The local ID exhibits a complementary and highly diagnostic pattern: values remain modest *inside* the training extent but rise sharply in its immediate geodesic neighborhood, forming anisotropic “halos” that effectively delineate the pre-training footprint. This finding further adds evidence that the intrinsic dimension can identify the spatial coverage of pre-training data.

## E VALIDATING OUR ESTIMATES OF INTRINSIC DIMENSION

We verify the reliability of intrinsic-dimension (ID) estimators for geographic INRs by constructing a testbed where the “ground truth” ID is effectively known. Pope et al. (2021) validated ID estimators in vision by creating synthetic image manifolds with a known upper bound on ID via GAN latent dimension, then checking whether estimators recover the expected scaling. This established a baseline of estimator reliability for image data. For spherical INRs without additional information augmentation (such as those done in Klemmer et al. (2025); Cepeda et al. (2023); Mai et al. (2023a)), the underlying representation geometry can be described by the 2-sphere ( $S^2$ ). We validate whether common ID estimators behave sensibly under sphere-respecting positional encodings and simple learned mappings (where the “true” ID should remain  $\approx 2$ ).

Our brief experimental framework first measures the intrinsic dimension of raw geographic co-ordinates. This is a toy-example where the ambient (extrinsic) dimension is equal to the intrinsic dimension. We increase the difficulty by passing raw longitude–latitude coordinates through a spherical–harmonics (SH) positional encoder (Rußwurm et al., 2024). Concretely, the encoder maps  $(\lambda, \varphi) \in S^2$  to the concatenated vector of real SH basis values  $\Phi_L(\lambda, \varphi) = [Y_{\ell,m}(\lambda, \varphi)]_{\ell \leq L, -\ell \leq m \leq \ell} \in \mathbb{R}^{\sum_{\ell=0}^L (2\ell+1)}$ , where  $Y_{\ell,m}$  are the (real) spherical harmonics (Eqs. (3)–(5) in (Rußwurm et al., 2024)). While arbitrary nonlinear transformations can alter ID, maps that are smooth and locally invertible (bi-Lipschitz, or more generally  $C^1$  with full-rank Jaco-

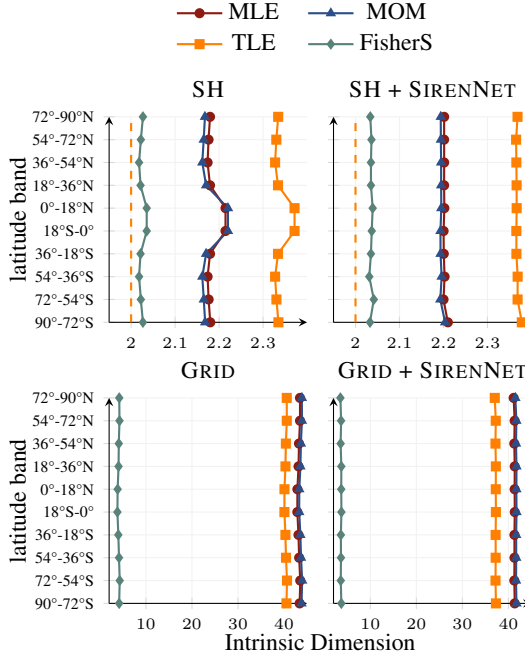


Figure 9: **Latitudinal variation of ID estimators with a Spherical Harmonics (SH) and Grid encoding of geographic co-ordinates.** The FisherS estimator is the most robust to latitudinal effects and different encodings of geographic coordinates.

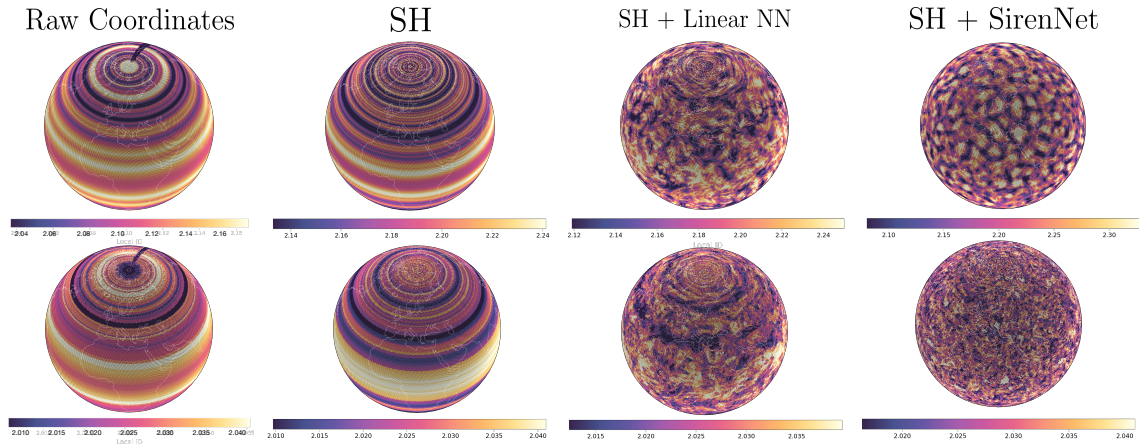


Figure 10: **Verifying local ID estimates of embeddings derived from  $S^2$ .** Top rows shows local ID estimates of the MLE estimator (distance-based) and bottom row shows estimates from the FisherS estimator (angle-based). Columns indicate the local ID plots with (i) raw geographic coordinates ( $\lambda, \phi$ ) (ii) pure spherical-harmonic-based embeddings (SH) of ( $\lambda, \phi$ ), and spherical harmonic embeddings passed through a randomly initialized (iii) linear network and (iv) Sinusoidal representation network (Sitzmann et al., 2020; Rußwurm et al., 2024).

bian—i.e., immersions) do not. They reparameterize the underlying manifold without adding degrees of freedom. The spherical-harmonic encoder  $\Phi_L$  and its compositions with linear layers or smooth pointwise nonlinearities (e.g., Siren) satisfy these conditions generically (Jacobian rank = 2 almost everywhere on  $S^2$ ), so the intrinsic dimension should remain 2.

Table 5 shows that the FisherS estimator is consistently closest to the ground-truth ID = 2 across all stages (raw, SH, SH+linear, SH+Siren), yielding the smallest MAE overall. In contrast, the distance-ratio estimators MLE/MOM, and TLE are biased high—especially after applying SH and learned mappings—indicating greater sensitivity to the embedding and neighborhood construction.

In Fig. 10, interestingly, both the MLE and FisherS ID estimators produce a latitude-based striping with the raw geographic coordinates that persist when these geographic coordinates are passed to the SH positional encoder. A plausible explanation is periodicity and coordinate singularities of longitude-latitude: although the points lie on  $S^2$ ,  $(\lambda, \phi)$  imposes a rectangular chart with a discontinuity at the international date line ( $\lambda = \pm 180^\circ$ ) and singular behavior at the poles. As a result, locations that are geodesically close on the sphere can appear far apart (or vice versa) when distances and neighborhoods are constructed in raw  $(\lambda, \phi)$  space across the wrap and near the poles. Non-nonlinearties introduced by a linear or siren-network change these distinct latitude-based striping patterns for both the MLE and FisherS estimator. We find additional evidence here that local ID maps point to properties of the positional encoding functions used to construct the geographic INRs: a spherical harmonic positional encoding followed by a linear network creates crowding and striping at the poles (column SH + LINEAR NN in Figure 10) which is removed completely when we use a Sinusoidal representation network (column SH + SIRENNET).

Lastly, we find strong evidence from Figure 9 that the FisherS is most robust to latitudinal variation of intrinsic dimension, and is significantly more robust when comparing ID results between geographic INRs constructed with different positional encoders. The Grid positional encoding (employed in Mai et al. (2023a)) causes distance-based estimators of ID to present high values of ID, but FisherS stays close to the true ID of 2.

Table 5: Intrinsic dimension estimates for spherical data (true ID = 2.0) with (i) raw geographic coordinates  $(\lambda, \phi)$  (ii) pure spherical-harmonic-based embeddings of  $(\lambda, \phi)$ , and spherical harmonic embeddings augmented with a randomly initialized (iii) linear network and (iv) Sinusoidal representation network (Sitzmann et al., 2020; Rußwurm et al., 2024). Values show mean  $\pm$  standard deviation. Best values (closest to 2.0) for each stage are shown in bold.

Stage	MLE	MOM	TLE	FisherS
Raw Coordinates	$2.113 \pm 0.056$	$2.097 \pm 0.069$	$2.038 \pm 0.039$	<b><math>2.024 \pm 0.008</math></b>
Pure SH	$2.189 \pm 0.032$	$2.183 \pm 0.046$	$2.344 \pm 0.034$	<b><math>2.025 \pm 0.011</math></b>
SH + Linear	$2.186 \pm 0.040$	$2.179 \pm 0.055$	$2.340 \pm 0.039$	<b><math>2.026 \pm 0.008</math></b>
SH + SIREN	$2.201 \pm 0.075$	$2.194 \pm 0.090$	$2.367 \pm 0.056$	<b><math>2.036 \pm 0.073</math></b>
MAE	0.172	0.163	0.272	<b>0.028</b>



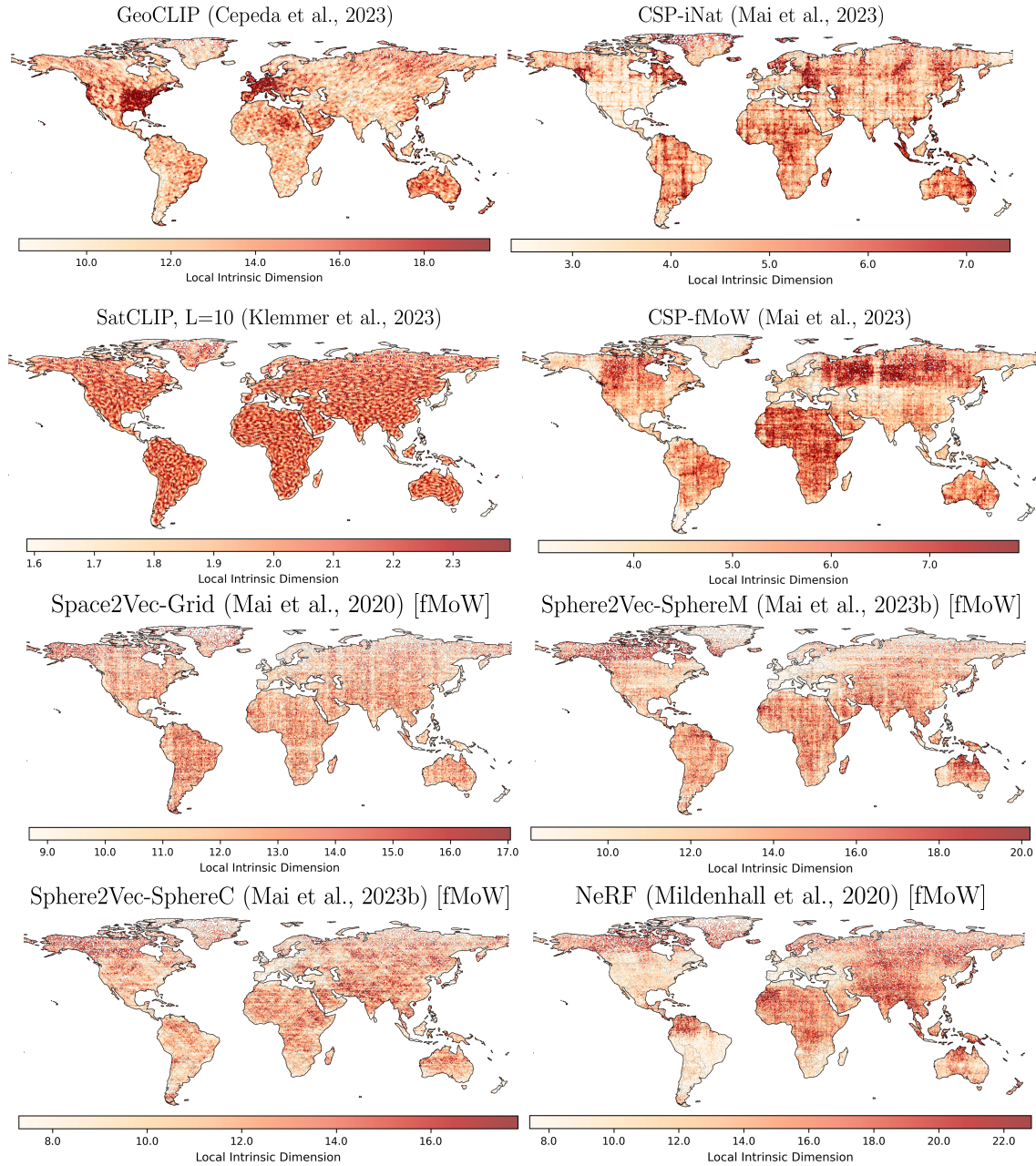


Figure 11: **Full Plot: Local intrinsic dimension of geographic INRs reveal spatial biases.** MLE estimator with  $k = 100$  nearest neighbors used. All Sphere2Vec and Space2Vec models are trained on the functional map of the world dataset (Christie et al., 2018) with supervised learning.