WHEN SPATIAL BLOCKING MAKES DISTANCE BUFFERING REDUNDANT: STRUCTURAL INERTNESS VERSUS SMALL EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial blocking and distance buffering both limit dependence between training and test samples, but their marginal roles are rarely separated. We define the buffer effect L as unbuffered minus buffered AUC on identical test units while matching training size and class counts, then vary only whether those units are scattered or contiguous. Removing near-test training samples costs $+.0693$ AUC for scattered street-view units at $r = 100$ m (95% CI $[+.0357, +.1028]$; 7/7 cities). Pooling two disjoint corrected So2Sat LCZ42 city sets gives $+.0252$ at $r = 200$ m ($[+.0136, +.0369]$; 29/41 cities; $p < 10^{-4}$). Compact blocking produces two distinct cases: in 28/48 primary cities the realised buffered and unbuffered training multisets are identical, so $L_B = 0$ by construction; among the remaining active cities, mean L_B ranges from $-.0012$ to $+.0004$ and each collection-level 90% interval lies inside a prespecified $\pm.005$ AUC tolerance, although one city exceeds that tolerance. A coordinate-only three-state rule separates provably inert, certified-bounded, and must-measure cases for spatially local learners, with no violations in 1,368 full-pool checks. Redundancy is therefore a property of the realised geometry and training intervention, not of the label "spatial CV."

1 INTRODUCTION

Random train–test splits can place nearly duplicate spatial contexts on both sides of an evaluation, producing an interpolation estimate when the intended deployment requires extrapolation. This mismatch has been documented across ecological mapping, remote sensing, and urban representation benchmarks (Roberts et al., 2017; Ploton et al., 2020; Kattenborn et al., 2022; Karasiak et al., 2022; Liu et al., 2026). Two common responses are to hold out contiguous spatial blocks or to exclude training samples within a chosen distance of the test set. Both are useful, but applying both does not imply that both materially alter the comparison.

The missing quantity is the *incremental* effect of a buffer after the test geometry has been chosen. Existing comparisons often change the holdout geometry, training size, and effective distance simultaneously. They therefore cannot tell whether a buffer contributes information beyond a block. We isolate that contribution by fixing the test units, radius, model, metric, training budget, and class ratio, and measuring the change caused by removing only the near-test portion of the training pool.

This distinction also clarifies a long-running disagreement. Random validation can be appropriate for prediction inside an already sampled region, whereas spatial blocking can be pessimistic for that estimand (Wadoux et al., 2021; Rolf, 2023). For transfer into an unsampled region the ordering can reverse. Distance-matching and spatial-plus approaches make the deployment distribution explicit (Milà et al., 2022; Meyer & Pebesma, 2021; 2022; Wang et al., 2023); buffered and block-based tools offer different approximations to it (Valavi et al., 2019). Our question is narrower than selecting one universal validation scheme: once a geometry is fixed, when can a particular buffer still change the score?

We contribute: (i) a matched decomposition of the buffer effect under scattered and contiguous test units; (ii) equivalence tests, geometry-only diagnostics, and a metric-aware three-state certificate that distinguishes a small estimate from a buffer that cannot act; (iii) robustness across two imaging data sources, three city collections, block constructions, representations, probes, and a class-matched

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

synthetic system; (iv) an oracle experiment relating each protocol to interpolation and extrapolation targets; and (v) a coordinate audit of 47 protocol–radius cells from five published benchmarks, including a designed split whose test set is provably inert to buffering while its validation set is not, complemented by measured effects on the official So2Sat and CityRep protocols. The resulting claim is conditional: compact blocks at moderate granularity often subsume a short buffer, but fragmented blocks, larger radii, or a different deployment target can invalidate that redundancy.

2 Related work

Spatial cross-validation was developed to reduce dependence between training and assessment samples (Roberts et al., 2017; Brenning, 2012; Pohjankukka et al., 2017). Empirical studies show that ordinary folds can overstate predictive performance (Ploton et al., 2020; Kattenborn et al., 2022; Karasiak et al., 2022), while design-based critiques show that blocking can understate map accuracy when validation and deployment locations follow the same sampled population (Wadoux et al., 2021). The appropriate protocol consequently depends on the generalisation target (Rolf, 2023; Milà et al., 2022).

Methodological work proposes buffers, blocks, joint geographic–feature-space splits, and combinations of contiguous assessment sets with exclusion zones (Valavi et al., 2019; Wang et al., 2023; Mahoney et al., 2023). Those methods address how to build a validation design. We instead measure whether a fixed buffer has any remaining action once a design already exists. Recent urban benchmarks demonstrate the practical importance of structured splits (Liu et al., 2026; Hou et al., 2026); our matched comparison gives a diagnostic for deciding whether an additional distance rule is consequential.

3 Matched buffer decomposition

3.1 A controlled comparison

For city c and test unit u, let T_{cu} be the test samples, $R_{cu} = D_c \setminus T_{cu}$ the unbuffered training pool, and $R_{cu,r} = \{i \in R_{cu} : d(i, T_{cu}) > r\}$ the pool after applying radius r. For every city–radius cell we take the largest positive and negative counts that are feasible in *all* buffered and unbuffered pools of the two geometries,

$$n_y^* = \min_{g\in\{A,B\},u} \min\{|R_{g,u}^{(y)}|, |R_{g,u,r}^{(y)}|\}, \qquad y \in \{+,-\}. \tag{1}$$

The joint minimum gives every arm the same class counts; $\tilde{N}$ is the across-city median of $n_+^* + n_-^*$. Fixed per-city random keys rank samples once, and each arm takes the n_y^* lowest eligible keys per class, so changes reflect removals and replacements rather than reshuffling (Appendix J). The most restrictive arm, typically the buffered contiguous pool, sets this budget. This removes training size as a confounder but measures L_A with less data than unconstrained scattered training; capacity sweeps preserve its direction, not a full-size effect magnitude. The unit-level buffer effect is

$$L_{cu} = S(\widetilde{R}_{cu} \to T_{cu}) - S(\widetilde{R}_{cu,r} \to T_{cu}), \tag{2}$$

where S is the task score (AUC for the primary classification tasks and R^2 for the CityRep regression audit). Regime A uses scattered folds; regime B uses contiguous blocks. Thus L_A and L_B differ in test geometry, not radius, training-set class ratio, training size, features, learner, or score; test-set class composition necessarily differs between scattered folds and contiguous blocks.

Two coordinate-only summaries expose whether the intervention is active. Writing R, R_r for R_{cu}, $R_{cu,r}$ when the unit is clear, the buffer action rate $A_r = 1 - |R_r|/|R|$ is the share of training samples removed; the boundary exposure $E_r = |\{t \in T : d(t, R) \le r\}|/|T|$ is the share of test samples with a training neighbour inside the radius. Let $Q_0 = \widetilde{R}$ and $Q_1 = \widetilde{R}_r$ denote the two realised training multisets in Equation 2. If $Q_0 = Q_1$, any deterministic learner has $L = 0$ exactly; we call this case *structurally inert*. The full-pool condition $A_r = 0$ is sufficient because $R_r = R$ and common-key subsampling then preserves the multiset, but the converse need not hold: a buffer can remove unselected pool members while leaving Q_0 and Q_1 identical. Standard uniform stability bounds concern average loss rather than AUC and are vacuous under our fitted settings (Appendix A);

we do not use them to justify the empirical effect. Under idealised sampling, scattered thinning gives $1 - A_r = \exp(-cr^d)$, while M compact equal-area regions yield $A_r(M) \approx c_1\sqrt{M} + c_2 M$ (Appendix A). These relations motivate diagnostics; they are not assumed by the empirical estimates.

A coordinate-only local certificate. For a realised training multiset Q, let $D_Q(t)$ be the coordinate-determined subset on which a local learner's prediction at test point t depends. This includes the k nearest members of Q for spatial kNN and the members inside a fixed support for compact-support kernels or geographically weighted learners. For two realised multisets Q_0, Q_1, let $C = \{t \in T : D_{Q_0}(t) \neq D_{Q_1}(t)\}$, $f = |C|/|T|$, and $f_y = |C \cap T_y|/|T_y|$. Only terms incident to C can change. Therefore, for a pointwise score that averages terms in $[0, 1]$, and for a pairwise score whose pair contributions lie in $[0, 1]$, respectively,

$$|\Delta S| \leq \begin{cases} f, & \text{pointwise metrics (e.g., accuracy or Brier score),} \\ f_+ + f_- - f_+ f_-, & \text{pairwise metrics (e.g., AUC).} \end{cases} \tag{3}$$

For the nested full-pool comparison $Q_0 = R, Q_1 = R_r$, a kNN dependency set can change only when an original neighbour is removed; this gives $E_r \leq f_1 \leq f_5 \leq f_{10} \leq \cdots$. For fixed-support learners, C is the set whose support intersects $Q_0 \triangle Q_1$. The bounds require no fitted model; the AUC form additionally requires labels. Appendix B gives the proof, tightness construction, and validation across three local learner families. The 1,368 validation checks use the nested full pools; recomputing C on the realised matched multisets of Equation 2, which are the input of Algorithm 1, adds 23,150 checks in the three primary cells, again with no violations.

Algorithm 1: coordinate-only buffer decision. Given the two realised training multisets, test coordinates, a locality definition, an equivalence margin δ, and labels when required by the metric: (1) return *provably inert* if $Q_0 = Q_1$ (any deterministic learner) or if $f = 0$ (the stated local learner); (2) return *certified bounded* for a local learner when the applicable bound in Equation 3 is below δ; and (3) otherwise return *must measure*. The last state says that coordinates do not settle the effect, not that the buffer necessarily changes the score.

The middle state applies only to local learners, and the certificate is sufficient rather than necessary. Without labels, a pairwise metric can be certified only through the zero-incidence case $f = 0$. Global learners—including the frozen-backbone probes used here and any end-to-end network fitted on the whole training set—do not satisfy this locality; Algorithm 1 can return only *provably inert* or *must measure* for them.

3.2 Test geometries and inference

Scattered units are random five-fold partitions. The headline seven- and ten-city cells average 20 independent partitions; the larger replication and supporting analyses use five. Contiguous units are $K = 5$ k-means blocks in a local metric projection. Blocks whose test fold contains only one class are discarded before Equation 1, because their AUC is undefined; they enter neither arm of the comparison. We also test quantile grids and region growing for $K \in \{3, 5, 10, 20\}$.

Cities are the inferential unit. We report t intervals over city means and two-sided sign-flip tests; for more than 20 cities the latter uses a fixed-seed Monte Carlo estimator with 200,000 draws. A missing rejection is not evidence of a negligible effect, so L_B is assessed with a 90% interval against a prespecified $\pm.005$ AUC reporting tolerance. This is an operational AUC tolerance rather than a universal scientific threshold, and we do not transfer it to R^2. Primary active-subset intervals are reported separately from structural identities. Tests and intervals in radius sweeps, block-construction checks, and benchmark audits are exploratory and unadjusted; they are not interpreted as family-wise error-controlled discoveries. With seven cities the smallest attainable two-sided sign-flip value is .0156.

4 Data and protocol

Street-view. The dataset contains 30,307 panoramas from seven Chinese cities. Point records approximate areal, uncertain ponding (Al-Abadi & Pradhan, 2020), and crowdsourced reports have

coverage and precision limits (Esparza et al., 2024). We therefore treat 150 m (positive) and 600 m (negative) as design radii and exclude the intermediate band. Across four alternative settings in Appendix F, L_A stays positive ($+.0635$–$+.0917$). This is registry-proximity classification, not hydrological susceptibility; each panorama averages four independently encoded horizon crops.

So2Sat LCZ42. We use Sentinel-2 patches from the public LCZ42 benchmark (Zhu et al., 2020). The primary binary task contrasts compact built classes (LCZ 1–3) with open built classes (LCZ 4–6), giving 14,924 patches in the ten official evaluation cities. An earlier analysis treated zero-based class indices as one-based and instead compared LCZ 2–4 with 5–7. All headline numbers use the corrected slice; the as-run slice is retained only for a transparent sensitivity analysis (Appendix J). A disjoint LCZ42 training split supplies 121,576 patches and 31 feasible cities. The conventional built-versus-natural collapse is a near-saturated control. Distances use patch centres: a 32×32 patch at 10 m spans 320 m, so $r = 200$ m may retain footprint overlap and 400 m is the first non-overlapping tested radius (Zhu et al., 2020).

Representation and learner. Main results use frozen SigLIP SO400M features and an ℓ_2-regularised logistic probe ($C = .1$, balanced class weights). DINOv2 ViT-L/14, CLIP ViT-L/14, weaker regularisation, kNN, and random forests are sensitivity analyses. Coordinates are projected per city; no model is fine-tuned.

Primary radii. We use $r = 100$ m for street-view and $r = 200$ m for So2Sat, the largest tested radius retaining all cities under the class-feasibility rule. Larger radii are reported as sensitivity cells rather than discarded. A street-view positive-cap sensitivity analysis is also reported, but its feasible city set changes with the cap and radius; it therefore cannot by itself separate radius from positive scarcity. Only six and two cities remain feasible in the primary radius sweep at 200 and 400 m.

5 Results

5.1 Buffering matters for scattered units, conditionally

At $r = 100$ m (Figure 1), the street-view buffer reduces scattered-unit performance by $L_A = +.0693$ AUC (95% CI $[+.0357, +.1028]$; sign-flip $p = .0156$), with a positive effect in all seven cities; .0156 is the minimum attainable two-sided sign-flip value for $n = 7$. The corrected ten-city So2Sat cell at $r = 200$ m is much smaller and uncertain: $+.0075$ ($[-.0073, +.0223]$; $p = .389$; 4/10 positive). Its disjoint 31-city counterpart gives $+.0309$ ($[+.0165, +.0454]$; $p < 10^{-4}$; 25/31 positive).

The two So2Sat city sets use the same corrected labels, learner, budget rule, and radius and contain no overlapping cities. As a descriptive summary across both collections, pooling gives $L_A = +.0252$ over 41 cities ($[+.0136, +.0369]$; sign-flip $p < 10^{-4}$; 29/41 positive). A permutation test does not distinguish the subset means ($p = .080$), although their spreads differ (.0207 versus .0394; permutation $p = .047$). The subsets also use 20 and five scattered partitions; matching both at five moves the pooled estimate by only $+.0002$ (to $+.0254$). Power simulations at the replication effect $+.0309$ give .98 power for ten cities under the ten-city spread but .59 under the replication spread. We therefore retain the two subsets separately and do not explain their difference by low power or by the non-rejection of equal means. The saturated built-versus-natural control gives only $+.0003$ around a buffered-arm AUC of .974. Thus a buffer effect is neither automatic nor a dataset-wide constant: it depends on the label, scale, and remaining predictive headroom.

The main logistic probe is conservative with respect to learner capacity. In separate five-partition sweeps across street-view and the 31-city replication, weaker regularisation, kNN, and random forests increase L_A from its five-partition value $+.0733$ to as much as $+.1294$ and from $+.0309$ to as much as $+.0528$, respectively (Appendix E). The same models leave L_B near zero, so the headline effects do not arise because a weak linear probe alone is unusually sensitive to scattered leakage.

5.2 Compact blocking leaves little marginal action

The all-city contiguous means in Table 1 mix two different events. In 28/48 primary cities (2/7 street-view, 8/10 ten-city So2Sat, and 18/31 in the replication), every realised buffered and unbuffered

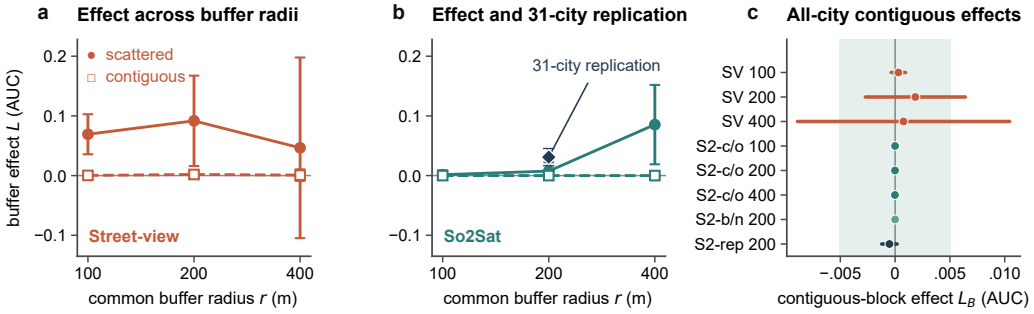

Figure 1: Matched buffer effects across radius. (a,b) Scattered and contiguous test geometries use the same radius and common class budget. The ten-city corrected So2Sat estimate is uncertain at 100–200 m; the disjoint 31-city replication is shown at 200 m. Street-view at 200 and 400 m retains six and two cities, respectively; the ten-city 400 m cell has median budget 104 and is a feasibility boundary, not evidence of a stronger scale effect. (c) L_B all-city intervals against the $\pm.005$ AUC tolerance. These intervals include structurally inert cities and are descriptive; active-subset inference is reported in Section 5.2 and Appendix G. The pooled 41-city So2Sat estimate is reported in the text; the panels retain the subsets to expose their different spreads.

Table 1: Primary specification. The seven- and ten-city SigLIP cells average 20 scattered partitions; the 31-city replication uses five. A and B denote scattered and contiguous test units. All arms within a city share the class budget in Eq. 1; $\tilde{N}$ is the median across cities of the matched training budget $n_+^* + n_-^*$. AUC is the buffered-arm score. B-row intervals pool structural identities and active cities and are not used as active-subset evidence.

Task	Units	Buf. AUC	L	95% CI	p	$L > 0$
Street-view	A: scattered	0.596	$+.0693$	$[+.0357, +.1028]$	0.0156	7/7
(7 cities, $r = 100$ m, $\tilde{N} = 2023$)	B: contiguous	0.611	$+.0003$	$[-.0004, +.0010]$	0.1250	4/7
So2Sat compact/open	A: scattered	0.755	$+.0075$	$[-.0073, +.0223]$	0.3887	4/10
(10 cities, $r = 200$ m, $\tilde{N} = 422$)	B: contiguous	0.609	$-<.0001$	$[-<.0001, <.0001]$	0.5000	0/10
So2Sat c/o (replication)	A: scattered	0.713	$+.0309$	$[+.0165, +.0454]$	$< .0001$	25/31
(31 cities, $r = 200$ m, $\tilde{N} = 368$)	B: contiguous	0.648	$-.0005$	$[-.0012, +.0002]$	0.1631	7/31
So2Sat built/natural	A: scattered	0.974	$+.0003$	$[-.0006, +.0012]$	0.4746	8/10
(saturated, $r = 200$ m, $\tilde{N} = 1622$)	B: contiguous	0.960	$-<.0001$	$[-<.0001, <.0001]$	0.4844	2/10

training multiset is identical under the common-key rule. Their $L_B = 0$ values are structural identities, not measurements of a small model response.

We therefore re-aggregate only cities in which at least one realised training multiset changes. The active-subset means are $+.0004$ for street-view ($n = 5$; 90% CI $[-.0004, +.0012]$), $-.00004$ for ten-city So2Sat ($n = 2$; $[-.0003, +.0002]$), and $-.0012$ for the 31-city replication ($n = 13$; $[-.0027, +.0002]$). Each collection-level interval lies inside the $\pm.005$ AUC tolerance, but this is not a city-wise guarantee: Islamabad has $L_B = -.0084$ (Appendix H). Compact blocking therefore often makes a short buffer either structurally inert or small on average when it acts (Appendix C); it does not make every active city equivalent. Appendix G reports the full stratification. Across all 104 active city–regime cells, the global probe's median $|L|$ is .0001 for contiguous blocks (90th percentile .0035; maximum .0107; $n = 31$), versus .025 for scattered units (.101; .251; $n = 73$), a two-order-of-magnitude separation after structural identities are excluded. These are descriptive statistics of this collection, not bounds. Within the active primary compact-block cities, common-key comparisons exchange only 0.2–11.6 training samples on average (fewer than three in most cities), yet effect size does not track that count: Islamabad exchanges about 1.4 and has the largest active effect, $|L_B| = .0084$. Thus non-zero action is a reason to measure, not a quantity to wave away. Radius sweeps remain conditional; the street-view 200 and 400 m cells have too few cities to support the same active-subset claim.

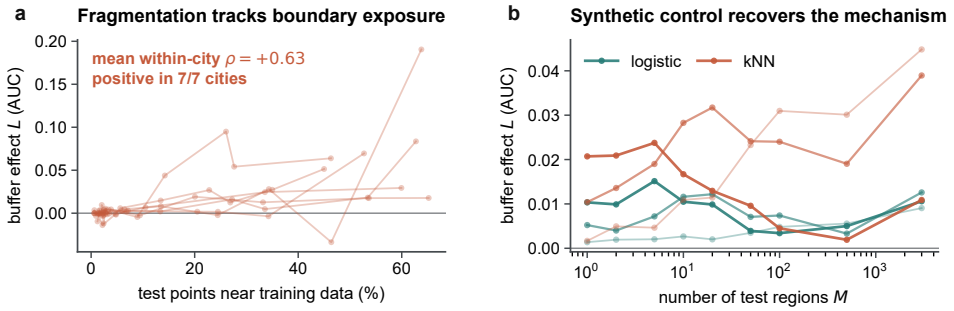

Figure 2: Geometry-only diagnostics. (a) Street-view fragmentation at $r = 100$ m: each line is one city and the horizontal axis is the share of test samples with a training neighbour inside the buffer. (b) Synthetic fields hold the generator, signal, features, class ratios (approximately .50), and learner fixed while changing only the number of test regions.

6 Why geometry changes the value of buffering

6.1 Fragmentation and boundary exposure

We divide each city into an increasing number of equal-size test regions while holding the 20% holdout, radius, class budget, and learner fixed (Figure 2). On street-view at 100 m, within-city rank correlation between boundary exposure and L averages $+.628$ and is positive in 7/7 cities. The association is similar with DINOv2 ($+.648$) and CLIP ($+.694$). Thus scattered and compact tests are useful endpoints of a continuum rather than inherently different protocols.

The continuum is not universal. Corrected So2Sat gives correlations near zero; raising the shared fragmentation budget to $M \leq 100$ increases available samples but leaves the mean correlation only $+.103$ (4/10 cities positive). We treat that null as a result, not as a budget artefact. Geometry is sufficient to move the effect in street-view and the controlled generator, but the mapping from exposure to model score remains data-dependent.

6.2 Theory checks and synthetic control

Fitting the scattered retention law gives $\hat{d} = .99$ for street-view, .82 for So2Sat, and 1.98 for a two-dimensional uniform synthetic process. The real-data exponents are descriptive because their sampling is neither homogeneous nor Poisson. In the synthetic system, varying only the number of test regions moves L smoothly for both logistic and kNN probes, showing that geometry alone can be sufficient. Fits and generator details appear in Appendix A.

6.3 Positive-cap sensitivity is not city-matched across radii

The street-view buffered arm becomes difficult beyond 100 m. Within a fixed city–radius cell, we hold the split, test units, and negative count fixed and vary only the positive cap. The feasible city set, however, changes with both cap and radius: all seven cities contribute at 100 m for caps up to 20, only four reach 100, and at 200 m the cap-100 result contains only Shenzhen. The pooled cap means therefore do not identify a radius effect. In the only cap-100 city available at both radii, buffered AUC is .569 at 100 m and .530 at 200 m; for the cross-cap comparison highlighted previously, Shenzhen instead rises from .504 (100 m, cap 5) to .530 (200 m, cap 100). We retain the analysis as a feasibility diagnostic (Appendix D) but do not use it to separate radius from positive scarcity.

6.4 The validation target still matters

Our synthetic oracle separates deployment inside a sampled region from deployment in an unsampled region. Random CV has the smallest interpolation bias. For extrapolation, block plus buffer has the smallest absolute bias in all four probe–scale cells, improving over an unbuffered block by .0003–

Table 2: Protocol bias against interpolation and extrapolation targets in 25 synthetic field realisations. ρ is each realised field's empirical correlation range in metres—the first lag at which label autocorrelation falls to .5—and is measured from the realisation rather than set on the generator. Positive values overstate deployment AUC; * denotes a two-sided sign-flip $p < .05$.

	Interpolation target		Extrapolation target	
Protocol	ρ=239	ρ=457	ρ=239	ρ=457
logistic probe				
random CV	$-.0047$	$-.0072^*$	$+.0215^*$	$+.0774^*$
scattered + buffer	$-.0135^*$	$-.0182^*$	$+.0148^*$	$+.0670^*$
contiguous block	$-.0258^*$	$-.0787^*$	$+.0004$	$+.0052$
block + buffer	$-.0263^*$	$-.0806^*$	$-.0001$	$+.0028$
knn probe				
random CV	$-.0023$	$-.0022$	$+.1131^*$	$+.2550^*$
scattered + buffer	$-.0610^*$	$-.0354^*$	$+.0524^*$	$+.2145^*$
contiguous block	$-.1083^*$	$-.2152^*$	$+.0033$	$+.0234$
block + buffer	$-.1109^*$	$-.2204^*$	$+.0007$	$+.0169$

.0064 (Table 2; differences from unrounded values). These differences are small but consistently favour the added buffer in this oracle. Buffering a scattered split reduces extrapolation optimism but does not generally eliminate it, and can make an interpolation estimate pessimistic. Redundancy therefore concerns the marginal operation of a buffer; it does not make all protocols estimate the same target.

7 Auditing published benchmark splits

We next apply the coordinate-only quantities to splits we did not design: CityRep (Liu et al., 2026), So2Sat LCZ42 (Zhu et al., 2020), WILDS PovertyMap (Koh et al., 2021), CropHarvest (Tseng et al., 2021), and reBEN (Clasen et al., 2025). Across 47 protocol–radius cells, three patterns emerge (Figure 3; Appendix I). First, disjoint regions can be exactly inert. The official So2Sat cross-city split has $A_r = E_r = f = 0$ through 400 m; extending the radius to 5 km still removes no sample. A fitted logistic probe confirms the consequent identity: unbuffered AUC is .7515 and $L = 0$ exactly at every tested radius.

Second, spatial separation within a city does not guarantee inertness. On So2Sat's own validation-to-testing partition, a 100 m buffer changes the global probe's AUC by $+.0121$ ($[-.0038, +.0281]$; sign-flip $p = .041$; 8/10 cities). None of the 100, 200, or 400 m effects satisfies the $\pm.005$ equivalence criterion. The spatial kNN effects are instead noisy and near zero at 100–200 m. This learner difference is descriptive: geometry permits an intervention but does not determine how a global representation responds.

Third, CityRep supplies a measured within-benchmark contrast. Because it ships no embeddings, we form each feature vector from the other task values available at the same co-registered grid cell; the learner never receives coordinates. The 492 protocol–radius–seed cells form 41 task–city units nested within only eight cities; the two released seeds are averaged before inference. At 400 m, city-level aggregation gives a mean R^2 effect of $+.0016$ ($[-.00005, +.00322]$) for the official block split and $+.0281$ ($[-.0041, +.0603]$) for the official random split. The task–city aggregation is $+.0312$ ($[+.0008, +.0616]$) for the random split, illustrating the narrower interval obtained when shared city structure is ignored. Moreover, median matched training size at 400 m is 10,692 under the block protocol but 999 under the random protocol. We therefore treat CityRep as a descriptive stress test, not a clean causal contrast of geometry and not an R^2 equivalence test. Because R^2 is neither a bounded pointwise average nor a pairwise ranking functional, CityRep's f is a reachability diagnostic, not an application of Equation 3.

Names are not guarantees. WILDS PovertyMap's country holdout is separated at 2 km, but at 10 km $f = .515$ because clusters can lie on opposite sides of land borders; its in-country and random protocols both reach about .92. These are reachability statements, not measured model effects. They

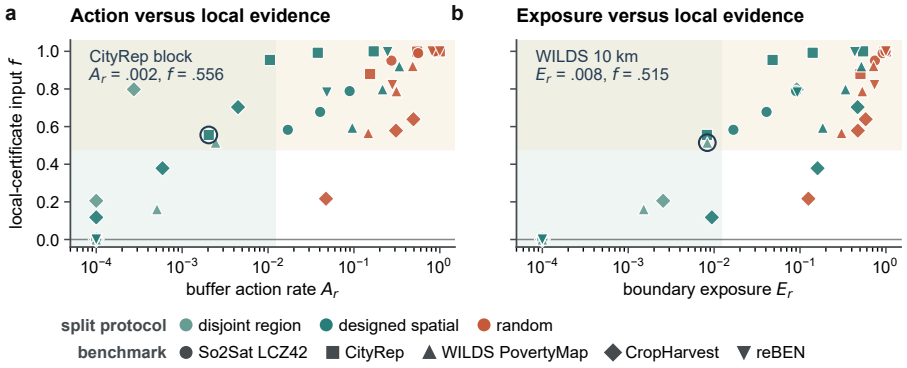

Figure 3: Coordinate-only audit of 47 published protocol–radius cells from five benchmarks. Small A_r can coexist with large local-certificate input f (left); E_r and f also measure different neighbourhood events (right). Points at zero are displayed at 10^{-4}.

demonstrate why A_r (cost), E_r (any-neighbour exposure), and f (local-certificate input) answer different questions.

Running Algorithm 1 on all 217 available cells requires no model fitting. For our 152 city–regime cells, all 76 scattered cells are must-measure; among 76 contiguous cells, 45 are provably inert and 31 remain must-measure. Thus the coordinate decision recovers the regime distinction only in one direction: inertness implies compactness here, but compactness does not imply inertness. Among the 65 available published-protocol cells, 21 are provably inert and 44 require measurement. Figure 3 and Table 11 display 47 prespecified protocol–radius cells, of which 35 are must-measure and 12 provably inert; the decision count adds the 18 cells audited but not displayed (CropHarvest at 200 and 400 m, reBEN at 200, 400 and 2400 m). The displayed count of 35 must-measure cells and the decision count of 44 therefore refer to different sets. In the 17 cells with measured effects, no provably inert cell changes, while seven must-measure cells nevertheless fall below $\pm.005$. Refusing a certificate is therefore not evidence of a nonzero effect. For spatial 10-NN, no non-zero contiguous bound is below .005 and the decision is binary. The middle state does occur for the other validated locality definitions: 10 compact-support-kernel cells at bandwidth r and six at bandwidth $2r$ have non-zero bounds below .005.

The designed split is where structural inertness is sharpest. reBEN (Clasen et al., 2025) replaced BigEarthNet's random assignment with a geographic division; its official training and test patches are never closer than 11.9 km, so $A_r = 0$ through 4.8 km and a buffer at any such radius is provably inert rather than merely small. The protection is not uniform across held-out sets: training and validation patches touch at 1.2 km, giving $A_r = .048$ there and .247 at 4.8 km, so model selection on this benchmark is exposed to the adjacency its test split was built to avoid.

8 Discussion

The actionable check is simple: report $A_r(r)$, $E_r(r)$, and the local exposure f with any spatial split. The three are not interchangeable as screens. Among the same 104 active cells, the smallest boundary exposure that still moves the global probe beyond the equivalence margin is $E_r = .005$: a thin removed band can lie outside almost every test point's immediate neighbourhood yet reach enough of them to matter. No E_r threshold separates compliant cells from exceptions in these data. Incidence orders the exceptions: within the active compact-block subset, all 19 cells with $f_{10} < .40$ stay within the margin (largest $|L| = .0035$), whereas all three exceptions have $f_{10} \geq .41$. Yet 9 of the 12 higher-incidence compact-block cells also stay within the margin. We therefore report an observed ordering and propose no threshold: the certificate in Equation 3 does not cover this probe, and a more sensitive learner can exceed the margin at any incidence. If compact blocks already place almost every training sample beyond the radius, an added buffer may be a nominal safeguard with no material intervention. Under budget matching, this check must be repeated on the realised training multisets: full-pool action can be positive even when the selected buffered and unbuffered multisets

are identical. If blocks are fine or fragmented, boundary-to-area ratio grows and the same buffer can act again. This reconciles our result with simulations where contiguous assessment sets plus exclusion buffers improve error estimates (Mahoney et al., 2023): those designs occupy a different part of the action-rate continuum.

Official protocols bound this conclusion. So2Sat is inert across cities but active within them, and its global and local learners respond differently. CityRep's larger random-split effect is confounded by city nesting and training size. Geometry determines whether a buffer can act; representation, learner, sample size, and target determine the score change.

Benchmark designers should state the deployment target, choose geometry accordingly, and verify that each safeguard changes admissible information. Algorithm 1 can stop at exact inertness, certify a local metric, or require measurement; global-model claims still require observed effects. Split names alone are insufficient: a country fold can be locally exposed, while a cross-city split can make a short buffer exactly inert.

9 LIMITATIONS

Intervals describe selected collections, not global prevalence. The So2Sat subsets have similar means but different spreads, and centre distances permit footprint overlap at 200 m but not 400 m. Two imaging sources omit continuous-field regression and broad suites such as SustainBench and GEO-Bench (Yeh et al., 2021; Lacoste et al., 2023), limiting transfer across tasks, biomes, and scales. Larger radii reduce feasible cities and budgets; fragmentation transfers across street-view backbones but not So2Sat, and AUC ceilings can mask responses. Equation 3 applies only to the stated local learner and bounded metric, with C computed from the two multisets actually compared, not to the global probe or CityRep's R^2; its agreement in 744 paired and 212 zero-bound cells is empirical calibration, not theorem extension. Frozen probes do not substitute for fine-tuning where a buffer acts; structural identities give $L = 0$ for any deterministic learner, fine-tuned networks included. Street-view labels measure registry proximity: their 150 m radius may inflate, but cannot create, L_A at a 100 m buffer, and Appendix F tests only four cities. The compact/open correction followed earlier analyses, and feasibility informed the radii. Primary intervals use one k-means seed; four further seeds leave every computable active-subset 90% interval inside $\pm.005$ (Appendix C); secondary tests are exploratory and unadjusted. CityRep uses nested task–city units, separate 60,000/20,000-sample caps, protocol-specific sizes, and sibling-task tabular features rather than official representations. Because these features lack the texture, illumination, and road-network cues through which an image encoder may leak spatial context, the measured CityRep effects likely understate vision leakage. PovertyMap coordinates have 2–5 km jitter. Appendix J records these decisions; the study was not pre-registered.

10 CONCLUSION

Distance buffering is not automatically complementary to spatial blocking. In our matched design it materially changes scattered-unit estimates for street-view and the pooled 41-city LCZ summary. After compact blocking, 28/48 primary city effects are exact training-multiset identities; among active cities, collection-level mean effects remain inside the stated AUC tolerance, but one city exceeds it. Exact inertness of So2Sat's cross-city split and reBEN's test split extends the structural case to published benchmarks, whereas CityRep remains a descriptive stress test rather than a clean geometry contrast. The boundary is the main result: geometry and sampling determine whether a buffer can act, while the learner, score, and deployment target determine the consequence. Reporting full-pool action, realised training-set action, local reach, and measured score change keeps those questions separate.

AI USE STATEMENT

Generative AI assisted source reconstruction, code and diagnostic development, synthetic analyses, literature and reference checks, and drafting and editing. It did not collect raw data, determine authorship, or make unreviewed scientific decisions. The authors selected all questions and interpretations,

reviewed AI-assisted text and code, verified claims, citations, and reported values against the artefacts, and take responsibility for the final content.

Ethics statement

The street-view task uses imagery from a commercial map provider. Because its terms and the completeness of face and licence-plate blurring do not support redistribution, neither panoramas nor image-derived features are released; only aggregate results and coordinate diagnostics are included. Optimistic evaluation could misallocate infrastructure resources, but registry-proximity labels do not measure household risk or flood vulnerability.

Reproducibility statement

The supplement contains code, source results, schemas, hashes, audits, certificates, benchmark analyses, and the synthetic pipeline. Headline seven- and ten-city cells use 20 scattered partitions; all repeat counts are in the manifest. Validation covers recomputations, checksums, anonymity, consistency, and deterministic packaging. Licensed street-view imagery and features are excluded; `bufferdiag` needs only coordinates and a split.

References

Alaa M. Al-Abadi and Biswajeet Pradhan. In flood susceptibility assessment, is it scientifically correct to represent flood events as a point vector format and create flood inventory map? *Journal of Hydrology*, 590:125475, 2020. doi: 10.1016/j.jhydrol.2020.125475.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Alexander Brenning. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 5372–5375, 2012.

Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sümbül, and Begüm Demir. reBEN: Refined BigEarthNet dataset for remote sensing image analysis. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2025. doi: 10.1109/IGARSS55030.2025.11242834.

Miguel Esparza, Hamed Farahmand, Xueming Liu, and Ali Mostafavi. Enhancing inundation monitoring of road networks using crowdsourced flood reports. *Urban Informatics*, 3(1), 2024. doi: 10.1007/s44212-024-00055-7.

Shuyang Hou, Ziqi Liu, Haoyue Jiao, Zhangyan Xu, Xiaopu Zhang, Lutong Xie, Yaxian Qing, Jianyuan Liang, Xuefeng Guan, and Huayi Wu. Slum detection and density mapping with AlphaEarth foundations: A representation learning evaluation across 12 global cities. *arXiv preprint arXiv:2605.10029*, 2026.

Nicolas Karasiak, Jean-François Dejoux, Claude Monteil, and David Sheeren. Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*, 111:2715–2740, 2022.

Teja Kattenborn, Felix Schiefer, Julian Frey, Hannes Feilhauer, Miguel D. Mahecha, and Carsten F. Dormann. Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 5:100018, 2022. doi: 10.1016/j.ophoto.2022.100018.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, 2021.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiaoxiang Zhu. GEO-Bench: Toward foundation models for earth monitoring. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2223.

Junyuan Liu, Xinglei Wang, Zichao Zeng, Jiazhuang Feng, Quan Qin, Ilya Ilyankou, Guangsheng Dong, and Tao Cheng. CITYREP: A unified benchmark for urban representations across cities, tasks, and modalities. *arXiv preprint arXiv:2605.26036*, 2026.

Michael J. Mahoney, Lucas K. Johnson, Julia Silge, Hannah Frick, Max Kuhn, and Colin M. Beier. Assessing the performance of spatial cross-validation approaches for models of spatially structured data. *arXiv preprint arXiv:2303.07334*, 2023.

Hanna Meyer and Edzer Pebesma. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9):1620–1633, 2021. doi: 10.1111/2041-210X.13650.

Hanna Meyer and Edzer Pebesma. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13:2208, 2022.

Carles Milà, Jorge Mateu, Edzer Pebesma, and Hanna Meyer. Nearest neighbour distance matching leave-one-out cross-validation for map validation. *Methods in Ecology and Evolution*, 13(6): 1304–1316, 2022. doi: 10.1111/2041-210X.13851.

Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, Alexei Lyapustin, Sylvie Gourlet-Fleury, and Raphaël Pélissier. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11:4540, 2020. doi: 10.1038/s41467-020-18321-y.

Jonne Pohjankukka, Tapio Pahikkala, Paavo Nevalainen, and Jukka Heikkonen. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31(10):2001–2019, 2017.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

Esther Rolf. Evaluation challenges for geospatial ML. In *ICLR Workshop on Machine Learning for Remote Sensing*, 2023. arXiv:2303.18087.

Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. CropHarvest: A global satellite dataset for crop type classification. In *NeurIPS Datasets and Benchmarks Track*, 2021. URL `https://openreview.net/forum?id=JtjzUXPEaCu`.

Roozbeh Valavi, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232, 2019. doi: 10.1111/2041-210X.13107.

Alexandre M. J.-C. Wadoux, Gerard B. M. Heuvelink, Sytze de Bruin, and Dick J. Brus. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, 457:109692, 2021. doi: 10.1016/j.ecolmodel.2021.109692.

Yanwen Wang, Mahdi Khodadadzadeh, and Raúl Zurita-Milla. Spatial+: A new cross-validation method to evaluate geospatial machine learning models. *International Journal of Applied Earth Observation and Geoinformation*, 121:103364, 2023. doi: 10.1016/j.jag.2023.103364.

Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David Lobell, and Stefano Ermon. SustainBench: Benchmarks for monitoring the sustainable development goals with machine learning. In *NeurIPS Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=5HR3vCylqD.

Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Häberle, Yuansheng Hua, Rong Huang, Lloyd Hughes, Hao Li, Yao Sun, Guichen Zhang, Shiyao Han, Michael Schmitt, and Yuanyuan Wang. So2Sat LCZ42: A benchmark data set for the classification of global local climate zones. *IEEE Geoscience and Remote Sensing Magazine*, 8(3): 76–89, 2020. doi: 10.1109/MGRS.2020.2964708.

A Coordinate-only theory and descriptive fits

Exact redundancy. Since $R_r \subseteq R$, $A_r = 0$ implies $R_r = R$. A deterministic learner then receives the same training multiset and has $L = 0$ on the same test set. Our common-key subsampling preserves this identity. For the matched estimator, the more direct condition is $Q_0 = Q_1$: it can also hold when $A_r > 0$ if every removed full-pool sample lies outside the realised common-key subsample.

Stability. If a learner has uniform stability β_N, replacing one sample changes an average risk score by at most $2\beta_N$ (Bousquet & Elisseeff, 2002); extending this to m replacements by the triangle inequality gives $2m\beta_N$. For the descriptive primary-cell calculation, $m = A_r\tilde{N}$, where $\tilde{N}$ is the median matched training budget. With scikit-learn's parameterisation $\frac{1}{2}\|w\|^2 + C\sum_i \ell_i$, dividing by Cn gives $\frac{1}{n}\sum_i \ell_i + \lambda\|g\|^2$ with $\lambda = 1/(2Cn)$. Logistic loss is 1-Lipschitz and the frozen features are ℓ_2-normalised, so $\sigma = \kappa = 1$ and Theorem 22 gives $\beta_N = \sigma^2\kappa^2/(2\lambda n) = C$. The resulting bound is $2A_r\tilde{N}C$. It equals 121.4, 46.4, and 40.5 for the primary street-view, ten-city So2Sat, and 31-city So2Sat cells. Worst-case limits of 40–121 loss units do not constrain the observed changes and are therefore vacuous here. Moreover, the bound concerns average logistic loss, whereas AUC is a ranking functional. Uniform stability does not directly bound the reported L and is retained only to document why it cannot support our claim.

Geometry. For a q-thinning of a homogeneous Poisson process in d dimensions, the probability that a retained training point has no test point within radius r is $\exp(-q\lambda\omega_d r^d)$. For M separated convex test regions, Steiner's formula gives a removed band proportional to perimeter plus corner area, producing $A_r(M) \approx c_1\sqrt{M} + c_2 M$. Figure 4 shows the descriptive fits; the synthetic exponent 1.98 recovers its known two-dimensional construction, while real-data exponents summarise sampling geometry only.

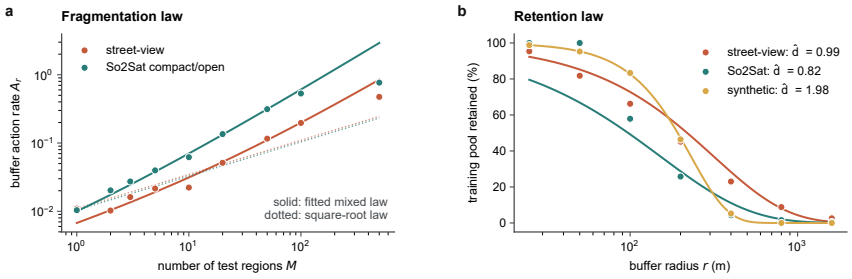

Figure 4: Descriptive checks of the two idealised action-rate laws. Fits exclude saturated fragmentation levels. The real-data exponent is not interpreted as a physical dimension.

B Proof and validation of the local certificate

For two realised training multisets Q_0, Q_1, let $D_{Q_j}(t)$ be the subset on which the stated local learner depends and let $C = \{t : D_{Q_0}(t) \neq D_{Q_1}(t)\}$. A local learner's prediction is unchanged outside C, so at most a fraction f of bounded pointwise terms can change. For class $y \in \{+, -\}$, set $C_y = C \cap T_y$ and $f_y = |C_y|/|T_y|$. AUC is the average contribution over all $|T_+||T_-|$ positive–negative pairs. Only pairs with at least one endpoint in $C_+ \cup C_-$ can change; their fraction is

$$1 - (1 - f_+)(1 - f_-) = f_+ + f_- - f_+ f_-.$$

Each pair contributes within $[0, 1]$, proving Equation 3. The pairwise bound is tight: a spatial 4-NN construction attains equality for five tested (f_+, f_-) combinations, including unequal class incidences, so no uniformly smaller bound depending only on these two fractions is possible. The pointwise bound is attained exactly in 14 contiguous cells across all three city sets (seven street-view), for the compact-support kernel with bandwidth r.

For the nested full-pool buffer $Q_0 = R, Q_1 = R_r$, spatial kNN has the one-sided invariance property: if none of the original k neighbours is removed, its neighbour set and prediction are unchanged. This supplies the kNN definition that a fixed-neighbourhood notation would not cover.

We evaluated the two bounds over 152 city–regime cells for spatial 10-NN and compact-support kernels with bandwidths r and $2r$, using accuracy, Brier score, and AUC on the nested full pools: all 1,368 checks had zero violations. The ordering $E_r \leq f_1 \leq f_5 \leq f_{10}$ also held in all 152 cells. For street-view at 100 m and $k = 10$, the contiguous Shenyang and Wuhan cells have bound zero, certifying $L = 0$ rather than merely observing it; the Kunming cell has bound .78 and measured $|L| = .01$, correctly refusing a certificate where local neighbourhoods can change substantially. Recomputing C on the realised matched multisets of Equation 2—the input of Algorithm 1—for every primary city, scattered repeat, and test unit of both geometries gives 4,630 comparisons and, over $k \in \{1, 5, 10, 25, 50\}$, 23,150 checks of both bounds with zero violations. Of the 2,155 contiguous comparisons, 1,617 have identical multisets and hence $f = 0$; no non-zero bound falls below .005 at any k, so the spatial-kNN decision is binary on the realised multisets as well. Scattered comparisons have a median bound of at least .997: the certificate refuses them, as it should. As an empirical calibration only, we also compare the full-pool bound with the global logistic probe in 744 paired cells. The global effect never exceeds the local bound's numerical value; the smallest observed envelope is $|L_{\text{global}}| \leq .251$ bound. That envelope is driven by one cell; across the 744 pairs, the median global-effect–to–local-bound ratio is .009 and its 90th percentile is .099. In all 212 cells with bound exactly zero, the global probe is unchanged. These are finite-sample observations, not guarantees for non-local learners.

C BLOCK CONSTRUCTION

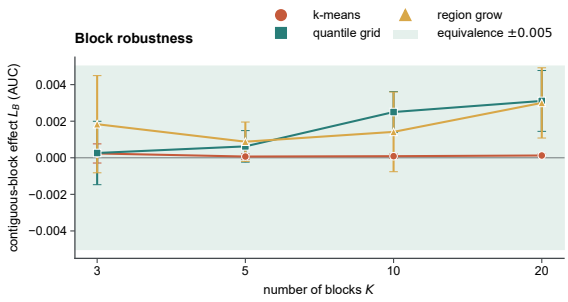

Figure 5: Street-view L_B under three block constructions. All four k-means cells satisfy equivalence at $\pm.005$; finer boundary-rich constructions can move upward even when their point estimate remains small.

Table 3: L_B for three constructions and $K \in \{3, 5, 10, 20\}$ across three city sets (36 cells). All 12 k-means cells pass equivalence; 29/36 cells pass overall. Fine grid/region cells fail most often, and the largest point estimate is $+.0095$ in the 31-city set.

Task	Construction	K	cities	L_B	90% CI	equiv.
So2Sat c/o, 31 cities (r=200)	kmeans	3	30	$+.0005$	$[-.0002, +.0012]$	yes
	kmeans	5	31	$-.0001$	$[-.0003, +.0001]$	yes
	kmeans	10	31	$+.0002$	$[-.0001, +.0005]$	yes
	kmeans	20	31	$+.0001$	$[-.0001, +.0003]$	yes
	quantile grid	3	32	$+.0024$	$[+.0004, +.0045]$	yes
	quantile grid	5	32	$+.0032$	$[+.0010, +.0054]$	**no**
	quantile grid	10	32	$+.0061$	$[+.0030, +.0092]$	**no**
	quantile grid	20	32	$+.0095$	$[+.0043, +.0147]$	**no**
	region grow	3	28	$+.0010$	$[-<.0001, +.0019]$	yes
	region grow	5	31	$+.0016$	$[-.0002, +.0034]$	yes
	region grow	10	31	$+.0014$	$[+.0001, +.0027]$	yes
	region grow	20	32	$+.0062$	$[-.0003, +.0128]$	**no**
So2Sat c/o, 10 cities (r=200)	kmeans	3	9	$+.0003$	$[-.0001, +.0008]$	yes
	kmeans	5	10	$-<.0001$	$[-<.0001, <.0001]$	yes
	kmeans	10	10	$+.0007$	$[-.0005, +.0020]$	yes
	kmeans	20	10	$+.0001$	$[-<.0001, +.0001]$	yes
	quantile grid	3	9	$+.0008$	$[-.0005, +.0022]$	yes
	quantile grid	5	10	$+.0021$	$[-.0008, +.0050]$	**no**
	quantile grid	10	10	$+.0007$	$[-.0018, +.0031]$	yes
	quantile grid	20	10	$+.0038$	$[+.0007, +.0068]$	**no**
	region grow	3	9	$+.0020$	$[-.0010, +.0051]$	**no**
	region grow	5	10	$+.0017$	$[-.0006, +.0041]$	yes
	region grow	10	10	$+.0015$	$[-.0001, +.0032]$	yes
	region grow	20	10	$+.0004$	$[-.0002, +.0009]$	yes
Street-view (r=100)	kmeans	3	7	$+.0002$	$[-.0003, +.0008]$	yes
	kmeans	5	7	$+.0001$	$[-.0001, +.0002]$	yes
	kmeans	10	7	$+.0001$	$[-<.0001, +.0002]$	yes
	kmeans	20	7	$+.0001$	$[<.0001, +.0002]$	yes
	quantile grid	3	7	$+.0003$	$[-.0015, +.0020]$	yes
	quantile grid	5	7	$+.0006$	$[-.0002, +.0015]$	yes
	quantile grid	10	7	$+.0025$	$[+.0014, +.0036]$	yes
	quantile grid	20	7	$+.0031$	$[+.0014, +.0048]$	yes
	region grow	3	6	$+.0018$	$[-.0008, +.0045]$	yes
	region grow	5	7	$+.0009$	$[-.0002, +.0020]$	yes
	region grow	10	7	$+.0014$	$[-.0008, +.0036]$	yes
	region grow	20	7	$+.0030$	$[+.0011, +.0049]$	yes

k-means seed. The primary contiguous units come from one k-means seed. Table 4 reruns the contiguous arm of the matched estimator under four further prespecified seeds, holding the scattered repeats, budget rule, and probe fixed; the first row of each block reproduces Table 9 exactly. The new seeds change the partition in 19–23 of the 48 cities, yet 27–32 remain structurally inert, every computable active-subset 90% interval lies inside $\pm.005$, and Islamabad remains the largest active effect. Where a seed leaves a single active city, no interval is formed.

Table 4: The compact-block summary of Section 5.2 under five prespecified k-means seeds. "Changed" counts cities whose partition differs from the primary seed's; "inert" counts cities whose realised multisets are identical in every comparison, among those with a feasible common budget.

Task	seed	changed	inert	active	mean L_B	90% CI	max. $\lvert L_B \rvert$
Street-view (r=100 m)	1 (primary)	–	2/7	5	$+.0004$	$[-.0004, +.0012]$	.0019 (Kunming)
	2	6/7	1/7	6	$+.0005$	$[-.0005, +.0015]$	.0029 (Kunming)
	3	7/7	2/7	5	$+.0004$	$[-.0004, +.0012]$	.0019 (Kunming)
	4	4/7	2/7	5	$+.0004$	$[-.0004, +.0012]$	.0019 (Kunming)
	5	5/7	2/7	5	$+.0006$	$[-.0002, +.0014]$	.0019 (Kunming)
So2Sat c/o, 10 cities (r=200 m)	1 (primary)	–	8/10	2	$-<.0001$	$[-.0003, +.0002]$	.0001 (Sydney)
	2	4/10	9/10	1	–	–	–
	3	4/10	9/10	1	–	–	–
	4	3/10	8/10	2	$<.0001$	$[-.0003, +.0004]$	.0001 (Sydney)
	5	5/10	9/10	1	–	–	–
So2Sat c/o, 31 cities (r=200 m)	1 (primary)	–	18/31	13	$-.0012$	$[-.0027, +.0002]$	.0084 (Islamabad)
	2	12/31	17/30	13	$-.0011$	$[-.0026, +.0003]$	.0084 (Islamabad)
	3	11/31	21/30	9	$-.0013$	$[-.0031, +.0006]$	.0084 (Islamabad)
	4	12/31	17/30	13	$-.0014$	$[-.0028, -.0001]$	.0084 (Islamabad)
	5	13/31	19/31	12	$-.0015$	$[-.0029, -<.0001]$	.0084 (Islamabad)

D FEASIBILITY AND CLASS-SPECIFIC RETENTION

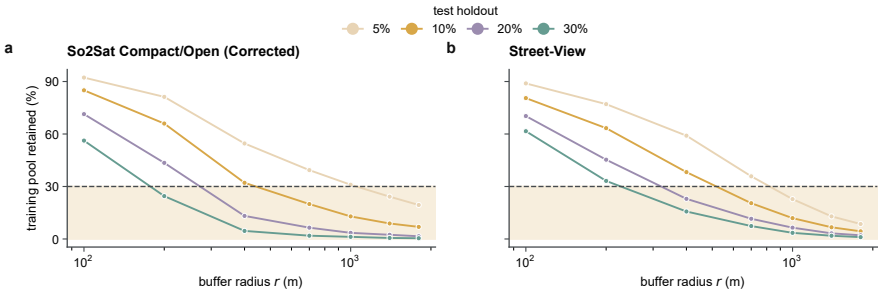

Figure 6: Median training-pool retention under scattered holdouts. Feasibility depends jointly on holdout fraction, radius, task, and class composition; no models are fitted.

Table 5: Median percentage of the scattered-unit training pool retained after buffering. Overall retention can conceal a much smaller surviving positive pool, which is why class-specific feasibility is enforced.

Task	Pool	$r = 100$	$r = 200$	$r = 400$
street-view	overall	70.3%	45.3%	23.0%
	positives	21.7%	4.2%	2.6%
So2Sat compact/open (corrected)	overall	71.4%	43.5%	13.2%
	positives	71.0%	42.6%	13.6%
So2Sat built/natural	overall	72.8%	44.4%	13.6%
	positives	70.8%	41.9%	10.4%

E BACKBONE AND PROBE SENSITIVITY

Table 6: Primary geometry across frozen backbones. Headline SigLIP rows use 20 scattered partitions for the seven- and ten-city cells; DINOv2/CLIP and the 31-city cells use five. Comparisons therefore support direction and scale, not a paired claim about repeat-level variance.

Task	Backbone	L_A	L_B	p	$L_A > 0$
Street-view ($r = 100$ m)	SigLIP SO400M	$+.0693$	$+.0003$	0.0156	7/7
	DINOv2 ViT-L/14	$+.0830$	$+.0002$	0.0156	7/7
	CLIP ViT-L/14	$+.0734$	$+.0004$	0.0156	7/7
So2Sat c/o, 10 cities ($r = 200$ m)	SigLIP SO400M	$+.0075$	$-<.0001$	0.3887	4/10
	DINOv2 ViT-L/14	$+.0192$	$-<.0001$	0.0059	9/10
	CLIP ViT-L/14	$+.0023$	$-<.0001$	0.8613	8/10
So2Sat c/o, 31-city replication ($r = 200$ m)	SigLIP SO400M	$+.0309$	$-.0005$	$< .0001$	25/31
	DINOv2 ViT-L/14	$+.0296$	$-<.0001$	$< .0001$	30/31
	CLIP ViT-L/14	$+.0307$	$+.0001$	$< .0001$	26/31

Table 7: Five-partition probe-capacity sweeps on all three city sets. Stochastic random-forest scores are averaged over seeds before city-level inference. The planned MLP arm was stopped before completion and is not reported. Headline SigLIP cells use 20 partitions where available.

Task	Probe	L_A	95% CI	L_B	equiv.
Street-view	logistic C=0.1	$+.0733$	$[+.0368, +.1098]$	$+.0003$	yes
	logistic C=10	$+.1038$	$[+.0679, +.1397]$	$+.0003$	yes
	kNN k=10	$+.1258$	$[+.0995, +.1521]$	$+.0005$	yes
	random forest	$+.1294$	$[+.0983, +.1604]$	$+.0001$	yes
So2Sat c/o, 10 cities	logistic C=0.1	$+.0084$	$[-.0073, +.0240]$	$-<.0001$	yes
	logistic C=10	$+.0260$	$[+.0118, +.0402]$	$+.0001$	yes
	kNN k=10	$+.0387$	$[+.0165, +.0608]$	$<.0001$	yes
	random forest	$+.0345$	$[+.0175, +.0515]$	$+.0001$	yes
So2Sat c/o, 31 cities	logistic C=0.1	$+.0309$	$[+.0165, +.0454]$	$-.0005$	yes
	logistic C=10	$+.0410$	$[+.0278, +.0542]$	$+.0006$	yes
	kNN k=10	$+.0515$	$[+.0419, +.0612]$	$+.0008$	yes
	random forest	$+.0528$	$[+.0397, +.0659]$	$+.0007$	yes

F STREET-VIEW LABEL SENSITIVITY

Table 8: Label-threshold and imagery-year sensitivity at $r = 100$ m. Only four cities with surviving verified point lists can be reconstructed, so the two-sided sign-flip floor is .125 and the table is directional rather than a powered significance analysis.

Setting	n	Pos. rate	L_A	p	L_B
pos<100 / neg>600	19,315	10.4%	$+.0917$	0.125	$-.0001$
pos<150 / neg>600	21,171	18.2%	$+.0793$	0.125	$+.0002$
pos<100 / neg>800	17,426	11.5%	$+.0893$	0.125	$<.0001$
pos<150 / neg>800	19,282	20.0%	$+.0814$	0.125	$-<.0001$
pos<150 / neg>600 / yr 2017	12,751	21.6%	$+.0635$	0.125	$+.0003$

G STRUCTURAL VERSUS ACTIVE COMPACT-BLOCK EFFECTS

A city is structurally inert here only when every realised contiguous comparison, across all common-budget repeats and usable blocks, selects the same buffered and unbuffered training multiset. Active

means that at least one comparison changes the realised multiset. The active-subset interval treats the city as the independent unit; the largest active value is shown to prevent a mean-equivalence statement from being read as a city-wise bound.

Table 9: Primary contiguous effects separated by realised training-set action. The inert counts are identities and receive no interval. Active-subset 90% intervals assess the collection mean against the $\pm.005$ AUC tolerance.

Task	inert	active	mean L_B	90% CI	equiv.	max. $\lvert L_B \rvert$
Street-view (r=100 m)	2/7	5	$+.0004$	$[-.0004, +.0012]$	yes	.0019 (km)
So2Sat c/o, 10 cities (r=200 m)	8/10	2	$-<.0001$	$[-.0003, +.0002]$	yes	.0001 (sydney)
So2Sat c/o, 31 cities (r=200 m)	18/31	13	$-.0012$	$[-.0027, +.0002]$	yes	.0084 (islamabad)

H PER-CITY PRIMARY RESULTS

Table 10: Every city underlying the primary table. The Street-view and ten-city So2Sat rows use the canonical 20-partition results; the 31-city replication uses five. "Buffered AUC" is the score of the buffered arm, not a claim that the protocol is unbiased for every deployment target.

Task	City	L_A	L_B	$\mathrm{AUC}_{\mathrm{buf,A}}$	$\mathrm{AUC}_{\mathrm{buf,B}}$
Street-view (r=100)	Dalian	+.0607	−<.0001	0.655	0.636
	Guangzhou	+.0989	+.0001	0.522	0.531
	Kunming	+.0444	+.0019	0.650	0.617
	Shenyang	+.0474	0	0.632	0.661
	Shenzhen	+.0216	<.0001	0.597	0.572
	Wuhan	+.1269	0	0.553	0.650
	Zhengzhou	+.0848	+.0001	0.566	0.612
So2Sat c/o (r=200)	Guangzhou	+.0440	0	0.688	0.755
	Jakarta	−.0033	0	0.754	0.708
	Moscow	−.0003	−<.0001	0.916	0.238
	Mumbai	−.0031	0	0.591	0.488
	Munich	−.0103	0	0.724	0.448
	Nairobi	−.0110	0	0.755	0.675
	San Francisco	+.0108	0	0.801	0.682
	Santiago	+.0456	0	0.737	0.738
	Sydney	+.0037	−.0001	0.729	0.701
	Tehran	−.0011	0	0.852	0.658
So2Sat replication (r=200)	Amsterdam	+.0254	0	0.729	0.579
	Beijing	+.0610	0	0.681	0.743
	Berlin	+.0570	−.0050	0.769	0.775
	Cairo	+.0002	<.0001	0.812	0.716
	Cape Town	−.0077	0	0.700	0.588
	Changsha	+.0694	−.0035	0.705	0.716
	Cologne	+.0090	0	0.762	0.730
	Dongying	−.0174	0	0.723	0.601
	Hong Kong	+.0042	+.0001	0.688	0.715
	Islamabad	+.0257	−.0084	0.581	0.485
	Istanbul	+.0113	+.0001	0.752	0.618
	Kyoto	+.0060	0	0.687	0.682
	Lisbon	+.1006	0	0.531	0.453
	London	+.0058	−.0007	0.788	0.478
	Los Angeles	+.0216	<.0001	0.765	0.780
	Madrid	+.0885	0	0.710	0.721
	Melbourne	+.0613	+.0006	0.791	0.681
	Milan	+.1145	0	0.678	0.506
	Nanjing	+.0994	0	0.554	0.513
	New York	−.0070	0	0.777	0.747
	Paris	+.0083	0	0.881	0.672
	Qingdao	+.0223	0	0.684	0.782
	Rio de Janeiro	−.0188	0	0.730	0.577
	Rome	+.0347	−.0021	0.723	0.691
	São Paulo	−.0232	+.0005	0.764	0.712
	Shanghai	+.0437	0	0.616	0.553
	Shenzhen	+.0413	0	0.633	0.643
	Vancouver	+.0065	0	0.742	0.676
	Washington DC	−.0105	0	0.842	0.684
	Wuhan	+.1005	+.0024	0.726	0.603
	Zurich	+.0251	−.0001	0.580	0.670

I FULL PUBLISHED-SPLIT AUDIT

Table 11 reports every protocol–radius cell summarised in Section 7. City-scoped distances use local metric projections; continental audits use great-circle chords. CityRep protocols replay the

released split implementation and five seeds; its largest tasks use uniform subsampling capped at 60,000 observations for this coordinate audit. The separate measured CityRep model analysis uses a 20,000-sample cap. DHS coordinates have 2–5 km published jitter. The pooled f column is a label-free certificate input: only $f = 0$ by itself proves inertness. For reBEN, the random 20% row is active at 100 m on a 1.2 km grid because reBEN patches are acquisitions: the official split keeps every repeat of a ground cell on one side, whereas a random split separates repeats of the same location.

Measured effects use the released benchmark partitions without replacing them by our own geometry. For So2Sat, the manifest's validation and testing labels define the same-city comparison, while the official training and evaluation city sets define the cross-city comparison. For CityRep, other co-registered task values at the same grid cell form standardised global features; tasks with insufficient feature coverage are excluded by a fixed 80% rule. Full task–city–seed cells and the six protocol–radius summaries are included in the result files. Inference is over task–city units after averaging the two released seeds; the per-seed-cell summaries are retained to document that aggregation.

Table 11: Coordinate-only audit of five public benchmarks. A_r is training mass removed, E_r is any-neighbour exposure, and f is the fraction of test points whose k-nearest-neighbour set contains a removed point.

Benchmark	Protocol	Geometry	r	A_r	E_r	f
CityRep	block10 (official)	designed spatial	100 m	.002	.008	.556
	block10 (official)	designed spatial	200 m	.011	.048	.954
	block10 (official)	designed spatial	400 m	.038	.140	.992
	block10 (official)	designed spatial	1000 m	.170	.542	1.000
	random (official)	random	100 m	.154	.505	.880
	random (official)	random	200 m	.545	.951	.999
	random (official)	random	400 m	.942	.999	1.000
	random (official)	random	1000 m	1.000	1.000	1.000
CropHarvest	leave-one-dataset-out	disjoint region	100 m	0	0	0
	leave-one-dataset-out	disjoint region	1000 m	$< .0001$	.003	.206
	leave-one-dataset-out	disjoint region	5000 m	$< .001$	.092	.798
	official is_test flag	designed spatial	100 m	$< .0001$	.009	.118
	official is_test flag	designed spatial	1000 m	.001	.160	.379
	official is_test flag	designed spatial	5000 m	.004	.468	.704
	random 20%	random	100 m	.047	.125	.217
	random 20%	random	1000 m	.309	.471	.579
	random 20%	random	5000 m	.493	.581	.639
So2Sat LCZ42	official (34 train areas → 10 test)	disjoint region	100 m	0	0	0
	official (34 train areas → 10 test)	disjoint region	200 m	0	0	0
	official (34 train areas → 10 test)	disjoint region	400 m	0	0	0
	validation → testing (same 10 cities)	designed spatial	100 m	.017	.017	.583
	validation → testing (same 10 cities)	designed spatial	200 m	.040	.041	.678
	validation → testing (same 10 cities)	designed spatial	400 m	.089	.090	.789
	random 80/20 (same pool)	random	100 m	.273	.744	.951
	random 80/20 (same pool)	random	200 m	.560	.917	.990
	random 80/20 (same pool)	random	400 m	.871	.988	1.000
WILDS PovertyMap / DHS	WILDS OOD country folds (official)	disjoint region	2 km	0	0	0
	WILDS OOD country folds (official)	disjoint region	5 km	.001	.002	.161
	WILDS OOD country folds (official)	disjoint region	10 km	.002	.008	.515
	in-country folds (official alternative)	designed spatial	2 km	.096	.184	.594
	in-country folds (official alternative)	designed spatial	5 km	.215	.338	.798
	in-country folds (official alternative)	designed spatial	10 km	.338	.518	.921
	random 80/20	random	2 km	.147	.304	.566
	random 80/20	random	5 km	.313	.531	.788
	random 80/20	random	10 km	.484	.718	.922
reBEN / BigEarthNet v2.0	leave-one-country-out	disjoint region	100 m	0	0	0
	leave-one-country-out	disjoint region	1200 m	0	0	0
	leave-one-country-out	disjoint region	4800 m	0	0	0
	official train->test	designed spatial	100 m	0	0	0
	official train->test	designed spatial	1200 m	0	0	0
	official train->test	designed spatial	4800 m	0	0	0
	official train->val	designed spatial	100 m	0	0	0
	official train->val	designed spatial	1200 m	.048	.093	.781
	official train->val	designed spatial	4800 m	.247	.434	.996
	random 20%	random	100 m	.280	.738	.821
	random 20%	random	1200 m	.813	.994	1.000
	random 20%	random	4800 m	.999	1.000	1.000

J ANALYSIS-DESIGN DISCLOSURE

The LCZ42 labels are stored as zero-based indices. An earlier compact/open analysis applied one-based slice boundaries and therefore compared LCZ 2–4 with 5–7. The corrected comparison is LCZ 1–3 versus 4–6. At $r = 200$ m the as-run slice gives $L_A = +.0199$ (10/10 positive), whereas the corrected ten-city result is $+.0075$ with an interval spanning zero. The independent 31-city corrected replication gives $+.0309$ and is the principal evidence that the corrected task can exhibit a scattered-unit effect. We retain the old slice only to disclose the correction, not as a second confirmatory result.

Other design changes were made after diagnostics exposed confounding: coordinate calculations were moved to per-city metric projections; all geometries were placed on a common class budget; scattered holdouts became complete five-fold partitions rather than independent 20% draws. Each sample receives one pseudo-random key drawn once per city from a deterministic city seed and fixed for the run; each arm takes the lowest eligible keys per class, so pool changes cannot reshuffle samples retained by both arms. The headline small-city cells were subsequently rerun with 20 partitions. These corrections and their artefact hashes are recorded in the reviewer-facing `CORRECTIONS_AND_PROVENANCE.md` file shipped with the supplementary material.