Model Evaluation for Geospatial Problems*

Jing Wang[1], Tyler A. Hallman[2], Laurel M. Hopkins[3], John B. Kilbride[4], W. Douglas Robinson[5], and Rebecca A. Hutchinson[6]

[1,3,4,5,6]Oregon State University, Corvallis, OR, USA
[2]Bangor University, Bangor, Gwynedd, UK
Email ids: [1,3,5,6]{wangji9, hopkilau, douglas.robinson, rah}@oregonstate.edu,
[2]t.hallman@bangor.ac.uk, [4]john.b.kilbride@gmail.com

Abstract

Geospatial problems often involve spatial autocorrelation and covariate shift, which violate the independent, identically distributed assumption underlying standard cross-validation. In this work, we establish a theoretical criterion for unbiased cross-validation, introduce a preliminary categorization framework to guide practitioners in choosing suitable cross-validation strategies for geospatial problems, reconcile conflicting recommendations on best practices, and develop a novel, straightforward method with both theoretical guarantees and empirical success.

1 Introduction

Cross-validation (CV) estimates how reliable and accurate a model's predictions are on new, unseen data. Standard CV provides unbiased model error estimates for independent, identically distributed (iid) data [2]. However, it can substantially underestimate model errors for geospatial problems due to the inherent spatial autocorrelation and the frequent presence of covariate shift. For example, natural resource managers may use environmental features (e.g., soil characteristics, canopy cover, rainfall) from a region that a threatened species currently occupies to learn a species distribution model (SDM) describing the species-habitat relationship. Conservation actions like translocating individuals of the species to a new area may require the SDM to be applied outside the region where it was fit, raising a key question: *How good are the model predictions in this new area?* If spatial autocorrelation within the training region permits information leakage between training and validation folds, CV estimates computed within the training region may be optimistically biased. Furthermore, the distribution of the features in the occupied region may differ from those in the new area, a phenomenon referred to as covariate shift. Standard CV fails to account for the challenges of spatial autocorrelation and covariate shift. This paper addresses the question of how to evaluate models for geospatial problems from both theoretical and empirical perspectives.

2 Background

Geospatial Problems Consider a dataset T_w consisting of a ***training set*** T_{tr} and a ***test set*** T_{te}: $T_w = \{T_{tr}, T_{te}\} = \{\{X_{tr}, \mathbf{y_{tr}}\}, \{X_{te}, \mathbf{y_{te}}\}\} = \{\{\mathbf{x_i}, y_i\}_{i=1}^{n_{tr}}, \{\mathbf{x_j}, y_j\}_{j=1}^{n_{te}}\}$, where $X, \mathbf{x}$ are features (all or some are spatial variables); $\mathbf{y}, y$ are the response variables; i and j subscripts respectively denote training and test samples. A spatial random variable is a stochastic process $Z : \mathfrak{D} \times \Omega \to \mathbb{R}$, where $\mathfrak{D} \subset \mathbb{R}^d$ is a region (typically, $d = 2$ or 3), and Ω is a sample space. When we observe a spatial variable's value, it creates a realization of the process by fixing $\omega \in \Omega$. ***Spatial autocorrelation*** (SAC) refers to the degree of spatial dependence (SD) between feature values measured at locations.

*This submission is an extended abstract of Wang et al. (2023) [1].

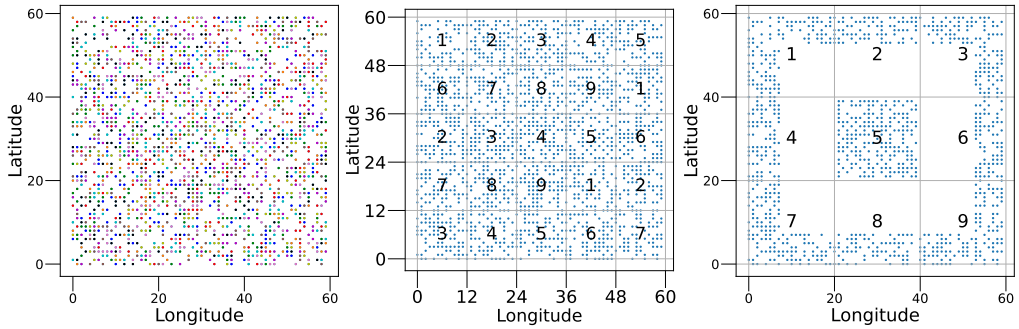
Computational Sustainability Workshop at NeurIPS 2023.

T_{tr} and T_{te} are collected from a ***training region*** $\mathfrak{D}_{tr}$ and a ***test region*** $\mathfrak{D}_{te}$, respectively. ***Covariate shift*** (CS) is defined by $P_{X_{tr}} \neq P_{X_{te}}$ while $P_{\mathbf{y_{tr}}|X_{tr}} = P_{\mathbf{y_{te}}|X_{te}}$, where P denotes distribution [3]. It is likely to happen when $\mathfrak{D}_{tr}$ and $\mathfrak{D}_{te}$ differ. In this paper, we focus on geospatial problems with the common traits: 1) Models are learned and applied at geolocated data points. 2) Geolocation information is available for each data point but not explicitly used in the model, i.e., X_{te} and X_{tr} do not include geocoordinates. 3) T_{tr} and X_{te} are known or estimable, while $\mathbf{y_{te}}$ is unknown.

Model Error Test error and risk are two common quantities for model error [4]. ***Test error*** (Err_T) is the expected loss over test samples, given a fixed training set T. (T is short for T_{tr} when different predictive goals are not emphasized.) ***Risk*** (R) is the expected test error over training sets from the same population. The key distinction is that risk considers expectations over training sets while test error conditions on a single training set. CV is occasionally mistaken for estimating test error, but it truly estimates the risk under the iid assumption [5].

Cross-validation Methods Standard CV approaches split data points into folds uniformly at random. As the most widely used variant, K-Fold Cross-Validation (KFCV) divides a training set randomly into K non-overlapping folds, iteratively holds out one fold as the ***validation fold***, and trains a model on the remaining ***training folds*** (e.g., Fig. 1(a)). Leave-One-Out Cross-Validation (LOOCV) is an extreme case, where it uses n-fold CV with a single data point as the validation fold, while the training folds contain the rest of the data. For iid data, these standard CV estimators are unbiased; the random resampling mimics how a new sample would be drawn from the population.

Additional CV techniques have been developed for non-iid data. Spatial CV strategies handle data dependencies by spatially segregating training and validation folds, and they can be broadly classified into two categories. First, BLock Cross-Validation (BLCV) groups geographically close points into blocks [6, 7]. BLCV reduces much spatial dependence across folds as most nearby points end up in the same fold (e.g., Fig. 1(b)). Second, BuFfered Cross-Validation (BFCV) inserts a buffer between training and validation folds and excludes points within it [8, 9]. This procedure removes spatial dependence between folds at the cost of losing training samples located within the buffer region (e.g., Fig. 1(c)). One concern with spatial CV methods is the possibility of introducing covariate shift between folds. Spatially separated folds may have differing feature spaces, possibly resulting in pessimistically biased error estimates. Importance-Weighted Cross-Validation (IWCV) does not assume identical distributions between training and test sets but was developed in non-spatial contexts [10, 11]. It provides an asymptotically unbiased estimator by adjusting the loss function with the ratio of test and training probability densities. Fig. 1 illustrates how these CV methods assign training samples to folds.



(a) KFCV, IWCV (b) BLCV (bl = 12) (c) BFCV (bl = 20, bf = 12)

Figure 1: Visualization of various CV methods dividing 1800 training data points on a 60×60 landscape into 9 folds. (a) Different colors represent different folds for KFCV or IWCV. (b) An illustration of BLCV with a block size (bl) of 12 grid cells, assigning each block to one of the 9 folds. (c) An example of BFCV using a block size (bl) of 20 grid cells and a buffer size (bf) of 12 grid cells. It highlights fold 5 as the validation fold, with training samples in its adjacent buffer region excluded.

3 Methodology

A Criterion for Unbiased CV Rabinowicz and Rosset (2020) prove that when training and test features are iid and a latent variable induces correlation structure in the response variable, the CV risk estimate (under squared error loss) is unbiased when the joint distribution of the test and training sets matches that of the validation and training folds, for all folds. To extend this result, we consider autocorrelation as the mechanism creating dependence, with a general loss function. We establish a criterion for unbiasedness in Thm. 1, where K is the number of CV folds, and the subscripts $k, -k$ denote the validation fold and the training folds, respectively. (See proof in Appx.)

Theorem 1. *If $P_{X_{te}|X_{tr}} = P_{X_k|X_{-k}}, \forall k \in 1, \ldots, K$, then cross-validation is an asymptotically unbiased estimator of the risk $R^{(n)}$.*

Framework for Selecting a CV Method For the purpose of exploring the relationships between problem characteristics and CV strategies, we introduce a categorization framework determined by two dimensions (Tab. 1). We assess spatial (in)dependence by the maximal ***semivariogram range*** of all features, which refers to the distance at which semivariance reaches its maximum value, indicating that observations at this distance or farther are spatially independent. We characterize covariate shift between training and test features by the Cramér-von Mises two-sample test (Cramér test) [12], a multivariate, distribution-free test with null hypothesis that the samples are identically distributed.

Table 1: Geospatial scenarios determined by semivariogram range and the Cramér test. We consider spatial (in)dependence between training and testing sets by comparing the nearest distance (d) between training and test samples with the semivariogram range (r) of the training features; and covariate shift by comparing the p-value (p) of Cramér test and the user-defined significance level α.

	Spatial Dependence	Spatial Independence
No Covariate Shift	Scenario SD: $d < r$ and $p \geq \alpha$	Scenario SI: $d \geq r$ and $p \geq \alpha$
Covariate Shift	Scenario SD + CS: $d < r$ and $p < \alpha$	Scenario SI + CS: $d \geq r$ and $p < \alpha$

A New CV Method For Scenario SI + CS (Tab. 1) which requires extrapolation to a new, spatially independent test region, none of the existing CV methods are quite appropriate. KFCV, BLCV, and BFCV are ineffectively in dealing with covariate shift, while IWCV is not tailored for autocorrelated data. Therefore, we propose Importance-weighted Buffered Cross-Validation (IBCV), combining the strengths of BFCV and IWCV. IBCV separates training and validation folds with a buffer region (as Fig. 1c) and employs density ratio weighting to correct covariate shift. The K-fold IBCV estimator is

$$\hat{R}^{(n)}_{KIBCV} \equiv \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in k^{th}} \frac{p_{te}(\mathbf{x_i})}{p_{tr}(\mathbf{x_i})} \mathcal{L}(y_i, \hat{y_i}(\mathbf{x_i}; T_{-k-bf})),$$

where bf is the buffer region, T_{-k-bf} is the training fold, the subscript (n) denotes the size of training set, $\frac{p_{te}(\mathbf{x_i})}{p_{tr}(\mathbf{x_i})}$ is the density ratio of a validation sample T_i, and $\mathcal{L}$ can be any loss function. We claim that IBCV is asymptotically unbiased. (See proof and Algo. 1 in Appx.)

Proposition 2. *IBCV is asymptotically unbiased: $\mathbb{E}_T[\hat{R}^{(n)}_{IBCV}] = R^{(n)}$ when $n \to \infty$.*

Prospective users of IBCV should weigh a few caveats. First, IBCV may not be well-suited for small datasets. Eliminating buffer points would have a stronger impact on smaller datasets, potentially resulting in a pessimistic bias and high variance. Second, IBCV should be expected to struggle with severe covariate shift just as other importance-weighed methods do. Such methods perform poorly when the support of $P_{X_{te}}$ has little overlap with the support of $P_{X_{tr}}$.

4 Experiments

Simulation Experiments We generated data with the model $\mathbf{y} = \mathbf{x^{(1)}} + \mathbf{x^{(2)}} + \mathbf{x^{(1)}} \cdot \mathbf{x^{(2)}} + \epsilon$, where $\mathbf{x^{(1)}}$ and $\mathbf{x^{(2)}}$ are two features with varying degrees of spatial autocorrelation and ϵ is an iid normal error term. We generated 100 simulations for each scenario, with each simulation comprising

Table 2: Simulations: average biases of 9-fold CV estimates of RMSE with respect to risk, across varying spatial autocorrelation ranges (r). Bias is calculated as the mean CV estimate minus the risk except for Scenario SD + CS, where bias is calculated as the average of absolute differences between CV estimates and test errors. The smallest biases in each row of the same scenario block are in bold, and the best CV methods are summarized alongside the scenario names.

	KFCV	BLCV	BFCV	IWCV	IBCV	KFCV	BLCV	BFCV	IWCV	IBCV
r	**Scenario SD: KFCV**					**Scenario SI: BLCV**				
4	**0.0103**	0.0174	0.0280	-0.0517	-0.0354	**0.0023**	0.0094	0.0200	-0.0611	-0.0449
8	**-0.0024**	0.0030	0.0691	-0.0605	0.0062	-0.0299	**0.0046**	0.0388	-0.1106	-0.0482
12	**-0.0010**	0.0778	0.1602	-0.0550	0.0956	-0.0438	**0.0347**	0.1123	-0.1778	-0.0477
	Scenario SD + CS: BFCV					**Scenario SI + CS: IBCV**				
4	0.1043	0.1068	**0.0973**	0.1651	0.1453	0.4883	0.4954	0.5060	**-0.0018**	0.0059
8	0.1938	0.1861	**0.1853**	0.2524	0.1872	0.4603	0.4955	0.5320	-0.0652	**-0.0347**
12	0.2472	0.2492	0.2583	0.2472	**0.1870**	0.4429	0.5217	0.6041	-0.0646	**0.0070**

1800 training points and 500 test points. We fitted linear models without the interaction term to mimic the common case of model misspecification.

We measured bias in the CV estimates of RMSE across all scenarios and varying degrees of spatial autocorrelation (Tab. 2). In Scenario SD, KFCV is the least biased. Its internal random partitioning mechanism intersperses training points among validation points, aligning with the spatial structure between training and test sets in this case. Instead, BLCV and BFCV always produce pessimistic bias via introduced covariate shift, especially for strongly autocorrelated data (when $r = 12$). In Scenario SI, BLCV is advantageous. BLCV retains some spatial dependence across fold boundaries, potentially causing error underestimation. However, it can also introduce covariate shift across folds, which could lead to error overestimation. When the two offset each other, as may be controlled with block size, BLCV can achieve an accurate error estimate. In Scenario SD + CS, BFCV seems the best choice though it displays the highest variability among all methods (Appx. Fig. 3). In Scenario SI + CS, IBCV and IWCV, explicitly addressing covariate shift, outperform others significantly. The gap between IBCV and IWCV widens gradually with SAC range. IBCV is **46.78% less biased** than IWCV when $r = 8$, and **89.16% less biased** when $r = 12$.

Example Application We predicted whether a Hermit Warbler (HEWA), one of the most prevalent species in the Oregon 2020 dataset [13], was observed or not at a certain survey site based on the surrounding habitat, as represented by four vegetation indexes computed from remote sensing data [14, 15]. 1000 training points and 500 test points were randomly sampled from different regions of Oregon to create datasets for each of the four scenarios of Tab. 1 (Fig. 2). For these real datasets, we could not directly estimate risk due to having only one landscape per analysis. Instead, we used test error as a proxy and assessed the five machine learning models - Ridge classifier (Ridge), Linear SVM (LSVM), K-Nearest Neighbors (KNN), Random Forest (RF), and Naive Bayes (NB) - based on classification error rate, i.e., the proportion of misclassified samples in the test set.

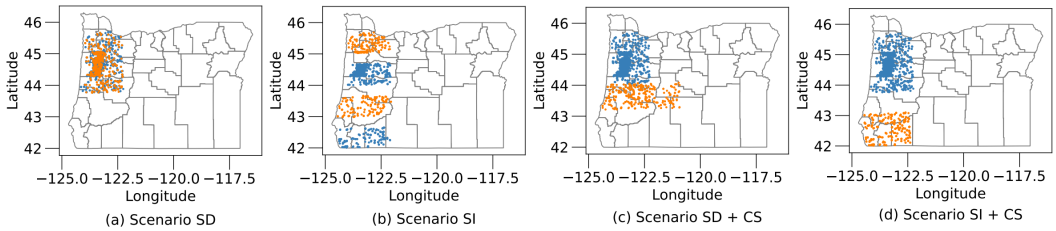


Figure 2: HEWA dataset: sampling strategies of training (blue) and test (orange) points in four scenarios. Subplots (a), (c) and (d) share the same training set. We fitted Matérn variogram functions with the lag class estimated by Scott's rule [16], and obtained the maximal range of the features was $r = 0.28$ degree for (a), (c), (d), while for (b) it was $r = 0.33$ degree. The nearest distances between training and test samples were $d = 0.00, 0.36, 0.00, 0.71$ degrees for (a), (b), (c), and (d), respectively. The p-values from the Cramér test were $p = 0.71, 0.04, 0.00, 0.00$ for (a), (b), (c) and (d), respectively. We set $\alpha = 0.01$. Therefore, the classification of the datasets align with Tab. 1.

From Tab. 3, the recommended CV methods for the four scenarios are almost consistent with those in the simulations. The only difference is that in Scenario SD, IWCV outperforms KFCV, though their estimates are quite close. Except for Scenario SD, KFCV usually severely underestimates model errors because when nearby points end up in different CV folds, spatial autocorrelation in features can transmit information across fold boundaries, leading to optimistic model error estimates. In Tab. 3, we set hyperparameters (block size and buffer size) to the maximum semivariogram range of all features. We also fine-tuned the hyperparameters and did a peak-to-peak comparison; in this case, the best CV method for each scenario remained the same (Appx. Tab. 4). That said, hyperparameters significantly influence CV estimates, and finding the best ones in practice can be challenging.

Table 3: HEWA: model classification error rates and 9-fold CV estimates thereof. The best estimates of test error (target) in each column of the same scenario block are in bold. The best CV methods which produce the closest estimates for most models are summarized alongside the scenario names.

Classifier	Ridge	LSVM	KNN	RF	NB	Ridge	LSVM	KNN	RF	NB
	Scenario SD: IWCV					**Scenario SI: BLCV**				
Test error	*0.1700*	*0.1720*	*0.1740*	*0.1740*	*0.1700*	*0.2320*	*0.2280*	*0.2440*	*0.2520*	*0.2440*
KFCV	0.1709	**0.1709**	0.1779	0.1910	0.1729	0.1890	0.1900	0.2180	0.2120	0.1950
IWCV	**0.1706**	0.1706	**0.1777**	**0.1907**	0.1727	0.1888	0.1898	0.2178	0.2118	0.1948
BLCV	0.1664	0.1678	0.1905	0.1964	0.1663	0.2007	**0.2047**	**0.2292**	**0.2607**	0.2043
BFCV	0.1783	0.1775	0.1909	0.1989	**0.1678**	**0.2173**	0.2876	0.2628	0.2692	**0.2110**
IBCV	0.1780	0.1773	0.1906	0.1986	0.1676	0.2170	0.2872	0.2625	0.2689	0.2107
	Scenario SD + CS: BFCV					**Scenario SI + CS: IBCV**				
Test error	*0.2140*	*0.2040*	*0.2080*	*0.1840*	*0.2160*	*0.2420*	*0.2540*	*0.2440*	*0.2540*	*0.2640*
KFCV	0.1709	0.1709	0.1779	**0.1910**	0.1729	0.1709	0.1709	0.1779	0.1910	0.1729
IWCV	**0.2400**	0.2430	0.2533	0.2706	0.2469	0.2239	0.2245	0.2370	0.2507	**0.2289**
BLCV	0.1644	0.1678	0.1905	0.1964	0.1663	0.1644	0.1678	0.1905	0.1964	0.1663
BFCV	0.1783	**0.1775**	**0.1909**	0.1989	0.1678	0.1783	0.1775	0.1909	0.1989	0.1678
IBCV	0.2526	0.2489	0.2649	0.2789	**0.2393**	**0.2370**	**0.2334**	**0.2484**	**0.2572**	0.2222

5 Conclusion

Recent studies have come to mixed conclusions on the best practice in evaluating models for geospatial problems, and our analysis yields points of both agreement and disagreement with the ongoing discussion. For example, Roberts et al. (2017) and Valavi et al. (2019) argue that spatial CV is less biased than non-spatial CV [6, 7]. Ploton et al. (2020) advocate for the adoption of spatial CV as the norm for spatially autocorrelated data [17]. Our results do provide evidence for spatial CV in some scenarios. However, Hoffimann et al. (2021) and Wadoux et al. (2021) show that spatial CV can yield notably pessimistic estimates [18, 19]. Our study corroborates this perspective, particularly in Scenario SD. Our agreements with these conflicting studies simply highlights our main message: *the best evaluation strategy for a geospatial problem depends on how the training set relates to the intended test set; specifically, the spatial and distributional relationships between features across CV folds should match those between the training and testing features.*

We see several directions for future work. The framework outlined in Tab. 1 served this study adequately, but to offer more precise recommendations to practitioners, we need more nuanced tools for characterizing geospatial problems, going beyond the current four discrete quadrants. While IBCV shows promise, further developments to aid hyperparameter selection, and improvements to the method itself, may be fruitful. Finally, we hope this paper not only serves practitioners interested in assessing models with geospatial data but also triggers the exploration of thoughtful ways for evaluating performance on other non-iid datasets.

6 Acknowledgements

We thank Tom Dietterich for comments on an early version of the manuscript, and four anonymous reviewers for comments that improved the paper. This research was supported in part by the National Science Foundation (NSF) under Grant No. III-2046678 (JW, RAH), the United States Department of Agriculture National Institute of Food and Agriculture (USDA-NIFA) award No. 2021-67021-35344 (AgAID AI Institute; JW, RAH), the National Aeronautics and Space Administration (NASA)

under Future Investigators in NASA Earth and Space Science and Technology (FINESST) Grant No. 80NSSC20K1664 (LMH), and the Bob and Phyllis Mace professorship (WDR).

References

[1] Jing Wang, Laurel Hopkins, Tyler Hallman, W. Douglas Robinson, and Rebecca Hutchinson. Cross-validation for geospatial data: Estimating generalization performance in geostatistical problems. *Transactions on Machine Learning Research*, 2023.

[2] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[3] Jose Garcia Moreno-Torres, Troy Raeder, Rocío Alaíz-Rodríguez, N. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognit.*, 45:521–530, 2012.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data mining, inference, and prediction*. Springer, 2001.

[5] Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association*, pages 1–22, 2023.

[6] 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.

[7] 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, 2018.

[8] Kévin Le Rest, David Pinaud, Pascal Monestiez, Joël Chadoeuf, and Vincent Bretagnolle. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global ecology and biogeography*, 23(7):811–820, 2014.

[9] 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.

[10] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, May 2007.

[11] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25:1324–1370, 2011.

[12] Theodore W Anderson. On the distribution of the two-sample cramer-von mises criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159, 1962.

[13] William D. Robinson, Tyler A. Hallman, and Jenna R. Curtis. Benchmarking the avian diversity of oregon in an era of rapid change. *Northwestern Naturalist*, 101:180 – 193, 2020.

[14] Eric P. Crist and Richard C. Cicone. A physically-based transformation of thematic mapper data—the tm tasseled cap. *IEEE Transactions on Geoscience and Remote Sensing*, GE-22(3):256–263, 1984.

[15] Laurel M Hopkins, Tyler A Hallman, John Kilbride, W Douglas Robinson, and Rebecca A Hutchinson. A comparison of remotely sensed environmental predictors for avian distributions. *Landscape Ecology*, 37(4):997–1016, 2022.

[16] Mirko Mälicke. Scikit-gstat 1.0: A scipy flavoured geostatistical variogram estimation toolbox written in python. *Geoscientific Model Development Discussions*, pages 1–43, 2021.

[17] 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 Pélissier Raphaël. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature communications*, 11, 2020.

[18] Júlio Hoffimann, Maciel Zortea, Breno De Carvalho, and Bianca Zadrozny. Geostatistical learning: Challenges and opportunities. *Frontiers in Applied Mathematics and Statistics*, 7, 2021.

[19] 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.

Appendix

Proof of Theorem 1 With the geospatial settings defined in Section 2, we further assume the training and test features have the same domain: $\mathbf{x_i}, \mathbf{x_j} \in \mathscr{X} \subset \mathbb{R}^m$, and the function $f : \mathbf{x} \to y$ is unchanged between training and test sets (i.e., no concept shift). Therefore, the domains of the response variables are also the same: $y_i, y_j \in \mathscr{Y} \subset \mathbb{R}$. The subscripts $k, -k$ denote the validation and training folds, respectively.

Proof. Since $p(X_{te}|X_{tr}) = p(X_k|X_{-k})$, we have

$$p(X_{te}, X_{tr})/p(X_{tr}) = p(X_k, X_{-k})/p(X_{-k}). \tag{1}$$

We multiply the LHS by $p(\mathbf{y_{te}}|X_{te})$ and the RHS by $p(\mathbf{y_k}|X_k)$. Since f is assumed constant, these quantities are equal, i.e., $p(\mathbf{y_{te}}|X_{te}) = p(\mathbf{y_{tr}}|X_{tr}) = p(\mathbf{y_k}|X_k)$.

We focus first on the LHS. Since $\mathbf{y}$ is conditionally independent of all other variables given its corresponding X, we can condition on additional variables, we have $p(\mathbf{y_{te}}|X_{te}) = p(\mathbf{y_{te}}|X_{te}, X_{tr}))$, and $p(\mathbf{y_{tr}}|X_{tr}) = p(\mathbf{y_{tr}}|X_{tr}, X_{te}, y_{te})$. Therefore, the condensed process is:

$$\begin{aligned}
&\frac{p(X_{te}, X_{tr}) \cdot p(\mathbf{y_{te}}|X_{te})}{p(X_{tr})} \\
&= \frac{p(X_{te}, X_{tr}) \cdot p(\mathbf{y_{te}}|X_{te}, X_{tr}) \cdot p(\mathbf{y_{tr}}|X_{tr}, X_{te}, \mathbf{y_{te}})}{p(X_{tr}) \cdot p(\mathbf{y_{tr}}|X_{tr})} \\
&= \frac{p(\mathbf{y_k}, X_k, \mathbf{y_{-k}}, X_{-k})}{p(\mathbf{y_{-k}}, X_{-k})} \\
&= p(T_{te}|T_{tr}).
\end{aligned}$$

Similarly, the result of multiplying RHS of Eqn. 1 with $p(\mathbf{y_k}|X_k)$ is $p(T_k|T_{-k})$. Note that T is short for T_{tr} when not emphasizing different predictive goals. Combining the manipulated LHS and RHS, we can conclude that

$$p(T_{te}|T_{tr}) = p(T_k|T_{-k}). \tag{2}$$

Now we are ready to show unbiasedness for the leave-one-out (LOO) setting. For this setting, Eqn. 2 is written as $p(T_j|T) = p(T_i|T_{-i})$, recalling that T_j is a single intended test instance outside of the full training set T (containing n training samples), T_i is a single-instance validation fold, and T_{-i} is the training fold (i.e., excluding T_i from T). As is typical, we assume that T_{-i} is distributed as T and of size n, ignoring the bias from the different sizes of T_{-i} and T; this gives $p(T_j|T) = p(T_j|T_{-i}) = p(T_i|T_{-i})$, which is needed for step (1) below. We use shorthand $\mathcal{L}_i$ for $\mathcal{L}(y_i, \hat{y}_i(\mathbf{x_i}; T_{-i}))$ and $\mathcal{L}_j$ for $\mathcal{L}(y_j, \hat{y}_j(\mathbf{x_j}; T_{-i}))$. The error estimate of standard LOOCV is

$$\hat{R}^{(n)}_{LOOCV} \equiv \frac{1}{n} \textstyle\sum_{i=1}^{n} L(y_i, \hat{y}_i(\mathbf{x_i}; T_{-i})).$$

So we have

$$\begin{aligned}
\mathbb{E}_T[\hat{R}^{(n)}_{LOOCV}] &= \frac{1}{n} \textstyle\sum_{i=1}^{n} \mathbb{E}_{T_{-i}, T_i}[\mathcal{L}_i] \\
&= \frac{1}{n} \textstyle\sum_{i=1}^{n} \mathbb{E}_{T_{-i}} \left[\int_{\mathscr{Y}} \int_{\mathscr{X}} p(\mathbf{x_i}, y_i|T_{-i}) \mathcal{L}_i d\mathbf{x_i} dy_i \right] \\
&\overset{(1)}{=} \frac{1}{n} \textstyle\sum_{i=1}^{n} \mathbb{E}_{T_{-i}} \left[\int_{\mathscr{Y}} \int_{\mathscr{X}} p(\mathbf{x_j}, y_j|T_{-i}) \mathcal{L}_j d\mathbf{x_j} dy_j \right] \\
&= \frac{1}{n} \textstyle\sum_{i=1}^{n} \mathbb{E}_{T_{-i}, T_j}[\mathcal{L}_j] \\
&= \frac{1}{n} \textstyle\sum_{i=1}^{n} R^{(n-1)} \overset{n\to\infty}{\longrightarrow} \approx R^{(n)}
\end{aligned}$$

All of these claims also hold for KFCV, with more bookkeeping required to account for varying fold sizes. □

Proof of Proposition 2 With the same settings and assumptions of Theorem 1, we show that IBCV is asymptotically unbiased.

Proof. We demonstrate the claim for LOOIBCV; it also valid for KIBCV with more bookkeeping for the folds. The leave-one-out IBCV estimator is

$$\hat{R}^{(n)}_{LOOIBCV} \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{p_{te}(\mathbf{x_i})}{p_{tr}(\mathbf{x_i})} L(y_i, \hat{y}_i(\mathbf{x_i}; T_{-i-bf})),$$

where bf is the buffer region, T_{-i-bf} is the training fold, the supscript (n) denotes the size of training set, and $\frac{p_{te}(\mathbf{x_i})}{p_{tr}(\mathbf{x_i})}$ is the density ratio of a validation sample T_i. Step (1) below holds because T_{-i-bf} and T_i are independent. Step (2) holds because T_{-i-bf} and T_j are independent. We use shorthand $\mathcal{L}_i$ for $\mathcal{L}(y_i, \hat{y}_i(\mathbf{x_i}; T_{-i-bf}))$ and $\mathcal{L}_j$ for $L(y_j, \hat{y}_j(\mathbf{x_j}; T_{-i-bf}))$.

$$\begin{aligned}
&\mathbb{E}_T[\hat{R}^{(n)}_{LOOIBCV}] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{T_{-i-bf}, T_i}\left[\frac{p_{te}(\mathbf{x_i})}{p_{tr}(\mathbf{x_i})}\mathcal{L}_i\right] \\
&\overset{(1)}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{T_{-i-bf}}\left[\int_{\mathscr{Y}}\int_{\mathscr{X}} \frac{p_{te}(\mathbf{x_i})}{p_{tr}(\mathbf{x_i})} p_{tr}(\mathbf{x_i}, y_i)\mathcal{L}_i d\mathbf{x_i} dy_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{T_{-i-bf}}\left[\int_{\mathscr{Y}}\int_{\mathscr{X}} p_{te}(\mathbf{x_j}) p_{te}(y_j|\mathbf{x_j})\mathcal{L}_j d\mathbf{x_j} dy_j\right] \\
&\overset{(2)}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{T_{-i-bf}, T_j}\left[\mathcal{L}_j\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} R^{(n-1-n_{bf})} \overset{n\to\infty}{\longrightarrow} \approx R^{(n)}.
\end{aligned}$$

□

Algorithm 1 LOOIBCV

Input: training set associated with the geocoordinates (lat, long) and density ratio (w) of each training point: $\{\mathbf{x_i}, y_i, lat_i, long_i, w_i\}_{i=1}^{n}$
Parameters: buffer size bf
Output: estimated error Err

1: **for** i = 1 to n **do**
2: Compute the distances from the validation point T_i to other training samples T_{-i};
3: Remove training samples with distance smaller than bf;
4: Fit a model $\hat{f}$ on the remaining training fold T_{-i-bf};
5: Compute density ratio weighted loss on the validation fold T_i:
$Err_i = w_i \cdot \mathcal{L}(y_i, \hat{y}_i(\mathbf{x_i}; T_{-i-bf}))$.
6: **end for**
7: **Return** the estimated error: $Err = \frac{1}{n}\sum_{i=1}^{n} Err_i$.

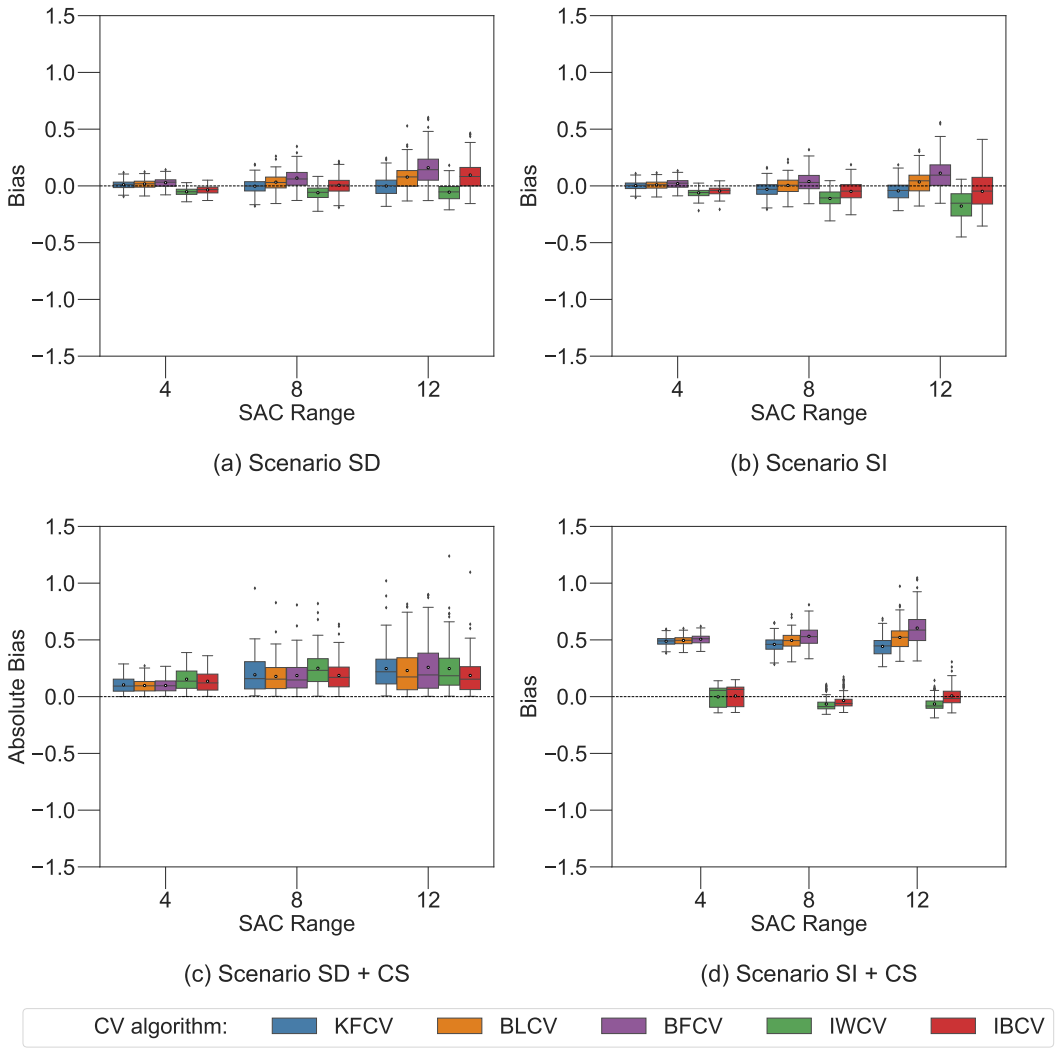


Figure 3: Biases of CV estimates in scenarios with various characteristics: spatially dependent (SD), spatially independent (SI), spatially dependent with covariate shift (SD + CS), and spatially independent with covariate shift (SI + CS). Circles inside the boxes display the mean values of biases. The black dash lines illustrate no bias. (a), (b) and (d) show bias of the average CV estimate to the risk. Since the feature distributions change across simulations in Scenario SD + CS, (c) plots the absolute bias, i.e., the absolute value of CV estimate minus test error in each simulation.

Table 4: HEWA1000: model classification test error rates (targets) and 9-fold CV estimates thereof (best estimates in each column in bold). BLCV-range, BFCV-range, and IBCV-range set the tuning parameters *a priori* based on the maximum semivariogram range of all features. BLCV-best, BFCV-best and IBCV-best estimates are selected from the best ones, for a peak-to-peak comparison. A dash means that setting the hyperparameters based on the range gives the best value.

Model	*Test error (target)*	KFCV	IWCV	BLCV -range	BFCV -range	IBCV -range	BLCV -best	BFCV -best	IBCV -best
				SD					
Ridge	*0.1700*	0.1709	**0.1706**	0.1664	0.1783	0.1780	-	-	-
LSVM	*0.1720*	**0.1709**	0.1706	0.1678	0.1775	0.1773	-	-	-
KNN	*0.1740*	0.1779	**0.1777**	0.1905	0.1909	0.1906	-	-	-
RF	*0.1740*	0.1910	**0.1907**	0.1964	0.1989	0.1986	-	-	-
NB	*0.1700*	0.1729	0.1727	0.1663	**0.1678**	0.1676	-	-	-
				SI					
Ridge	*0.2320*	0.1890	0.1888	0.2007	**0.2173**	0.2170	**0.2357**	0.2411	0.2408
LSVM	*0.2280*	0.1900	0.1898	**0.2047**	0.2876	0.2872	0.2406	0.2396	**0.2394**
KNN	*0.2440*	0.2180	0.2178	**0.2292**	0.2628	0.2625	**0.2402**	0.2383	0.2381
RF	*0.2520*	0.2120	0.2118	**0.2607**	0.2692	0.2689	-	0.2365	0.2363
NB	*0.2440*	0.1950	0.1948	0.2043	**0.2110**	0.2107	0.2424	**0.2441**	0.2438
				SD + CS					
Ridge	*0.2140*	0.1709	**0.2400**	0.1644	0.1783	0.2526	0.1976	**0.1997**	-
LSVM	*0.2040*	0.1709	0.2430	0.1678	**0.1775**	0.2489	0.2055	**0.2035**	-
KNN	*0.2080*	0.1779	0.2533	0.1905	**0.1909**	0.2649	**0.2186**	-	-
RF	*0.1840*	**0.1910**	0.2706	0.1964	0.1989	0.2789	-	-	-
NB	*0.2160*	0.1729	0.2469	0.1663	0.1678	**0.2393**	0.1984	**0.2037**	-
				SI + CS					
Ridge	*0.2420*	0.1709	0.2239	0.1644	0.1783	**0.2370**	0.1976	0.1997	**0.2384**
LSVM	*0.2540*	0.1709	0.2245	0.1678	0.1775	**0.2334**	0.2055	0.2035	**0.2544**
KNN	*0.2440*	0.1779	0.2370	0.1905	0.1909	**0.2484**	**0.2425**	0.2406	-
RF	*0.2540*	0.1910	0.2507	0.1964	0.1989	**0.2572**	0.2322	0.2453	-
NB	*0.2640*	0.1729	**0.2289**	0.1663	0.1678	0.2222	0.1984	0.2037	**0.2546**