EVALUATION CHALLENGES FOR GEOSPATIAL ML

Esther Rolf
Harvard University

ABSTRACT

As geospatial machine learning models and maps derived from their predictions are increasingly used for downstream analyses in science and policy, it is imperative to evaluate their accuracy and applicability. Geospatial machine learning has key distinctions from other learning paradigms, and as such, the correct way to measure performance of spatial machine learning outputs has been a topic of debate. In this paper, I delineate unique challenges of model evaluation for geospatial machine learning with global or remotely sensed datasets, culminating in concrete takeaways to improve evaluations of geospatial model performance.

1 MOTIVATION

Geospatial machine learning (ML), for example with remotely sensed data, is being used across consequential domains, including public health (Nilsen et al., 2021; Draidi Areed et al., 2022) conservation (Sofaer et al., 2019), food security (Nakalembe, 2018), and wealth estimation (Jean et al., 2016; Chi et al., 2022). By both their use and their very nature, geospatial predictions have a purpose beyond model benchmarking; mapped data are to be read, scrutinized, and acted upon. Thus, it is critical to rigorously and comprehensively evaluate how well a predicted map represents the state of the world it is meant to reflect, or how well a spatial ML model performs across the many conditions in which it might be used.

Unique structures in remotely sensed and geospatial data complicate or even invalidate use of traditional ML evaluation procedures. Partially as a result of misunderstandings of these complications, the stated performance of several geospatial models and predictive maps has come into question (Fourcade et al., 2018; Ploton et al., 2020). This in turn has sparked disagreement on what the "right" evaluation procedure is. With respect to a certain set of spatial evaluation methods (described in §4.1), one is jointly presented with the arguments that "spatial cross-validation is essential in preventing overoptimistic model performance" (Meyer et al., 2019) and "spatial cross-validation methods have no theoretical underpinning and should not be used for assessing map accuracy" (Wadoux et al., 2021). That both statements can simultaneously hold reflects the importance of using a diverse set of evaluation methods tailored to the many ways in which a geospatial ML model might be used.

In this paper, I situate the challenges of geopsatial model evaluation in the perspective of an ML researcher, synthesizing prior work across ecology, geology, statistics, and machine learning. I aim in part to disentangle key factors that complicate effective evaluation of model and map performance. First and foremost, evaluation procedures should be designed to measure as closely as possible the quantity or phenomena they are intended to assess (§2). After the relevant performance measures are established, considerations can be made about what is feasible with the available data (§3). With all of this in mind, possible evaluation procedures (§4) can be compared and tailored to the task at hand. Recognizing the interaction of these distinct but related steps exposes opportunities to improve geospatial performance assessment, both in individual studies and more broadly (§5).

2 MAP ACCURACY AND MODEL PERFORMANCE: CONTRASTING VIEWS

Estimating accuracy indices and corresponding uncertainties of geospatial predictions is essential to reporting geospatial ML performance (§2.1), especially when prediction maps will be used for downstream analyses or policy decisions. At the same time, the potential value of a geospatial ML model likely extends beyond that of a single mapped output (§2.2). Delineating the (possibly many) facets of desired model and map use is key to measuring geospatial ML performance (§2.3).

2.1 MAP ACCURACY AS A POPULATION PARAMETER TO BE ESTIMATED

Establishing notation we will use throughout, let $\hat{y}(\ell)$ denote a model's predicted value at location ℓ, and $y(\ell)$ the reference, or "ground truth" value (which we assume can be measured). To calculate a **map accuracy index as a population parameter** for accuracy index F is to calculate $A(\mathcal{D}) = F(\{(\hat{y}(\ell), y(\ell))\}_{\ell \in \mathcal{D}})$ where $\mathcal{D}$ is the **target population** of map use (e.g. all (lat, lon) pairs in a global grid, or all administrative units in a set of countries). Examples of common F include root mean squared error, and area under the ROC curve, among many others (Maxwell et al., 2021).

Typically, one only has a limited set of values y for locations in an evaluation set $\ell \in \mathcal{S}_{\text{eval}}$ from which to compute a statistic $\hat{A}(\mathcal{S}_{\text{eval}})$ to estimate $A(\mathcal{D})$. Wadoux et al. (2021) discuss the value of using a design-independent probability sample for design-based estimation of A (in contrast, model-based estimation makes statistical assumptions about the data (Brus, 2021)). Here a **design-independent sample** is one collected independently of the model training process. A **probability sample** is one for which every location in $\mathcal{D}$ has a positive probability of appearing in $\mathcal{S}_{\text{eval}}$, and these probabilities are known for all $\ell \in \mathcal{S}_{\text{eval}}$ (see, e.g. Lohr (2021)). Wadoux et al. (2021) emphasize that when $\mathcal{S}_{\text{eval}}$ is a design independent probability sample from population $\mathcal{D}$, design-based inference can be used to estimate $A(\mathcal{D})$ with $\hat{A}(\mathcal{S}_{\text{eval}})$, *regardless of the prediction model or distribution of training data.*

Computing statistically valid estimates of map accuracy indices is clearly a key component of reporting overall geospatial ML model performance. It is often important to understand how accuracy and uncertainty in predictions vary across sub-populations $\mathcal{D}_{r_1}, \mathcal{D}_{r_2} \ldots \subset \mathcal{D}$ (such as administrative regions or climate zones (Meyer & Pebesma, 2022)). If **local accuracy** indexes $A(\mathcal{D}_{r_1}), A(\mathcal{D}_{r_2}) \ldots$ are low in certain sub-regions, this could expose concerns about fairness or model applicability.

2.2 MODEL PERFORMANCE EXTENDS BEYOND MAP ACCURACY

Increasingly, geospatial ML models are designed with the goal of being used *outside* of the regions where training labels are available. Models trained with globally available remotely sensed data might be used to "fill in" spatial gaps common to other data modalities (§3.2). The goals of **spatial generalization**, **spatial extrapolation** or **spatial domain adaption** can take different forms: e.g. applying a model trained with data from one region to a wholly new region, or using data from a few clusters or subregions to extend predictions across the entire region. When spatial generalizability is desired, performance should be assessed specifically with respect to this goal (§4).

While spatial generalization is a key component of performance for many geospatial models, it too is just one facet of geospatial model performance. Proposed uses of geospatial ML models and their outputs include estimation of natural or causal parameters (Proctor et al., 2023), and reducing autocorrelation of prediction residuals in-sample (Song & Kim, 2022). Other important facets of geospatial ML performance are model interpretability (Brenning, 2022) and usability, including the resources required to train, deploy and maintain models (Rolf et al., 2021).

2.3 CONTRASTING PERSPECTIVES ON PERFORMANCE ASSESSMENT

The differences between estimating map accuracy as a population parameter (§2.1) and assessing a model's performance in the conditions it is most likely to be used (§2.2) are central to one of the discrepancies introduced in §1. Meyer et al. (2019); Ploton et al. (2020); Meyer & Pebesma (2022) state concerns in light of numerous ecological studies applying non-spatial validation techniques with the explicit purpose of spatial generalization. They rightly caution that when data exhibit spatial correlation (§3.1), non-spatial validation methods will almost certainly over-estimate predictive performance in these use cases. Wadoux et al. (2021), in turn, argue that performance metrics from spatial validation methods will not necessarily tell you anything about A as a population parameter.

A second discrepancy between these two perspectives hinges on what data is assumed to be available (or collectable). While there are some major instances of probability samples being collected for evaluation of global-scale maps (Boschetti et al., 2016; Stehman et al., 2021), this is far from standard standard in geospatial ML studies (Maxwell et al., 2021). More often, datasets are created "by merging all data available from different sources" (Meyer & Pebesma, 2022). Whatever the intended use of a geospatial model, the availability of and structures within geopsatial and remotely sensed data must be contended with in order to reliably evaluate any sort of performance.

3 STRUCTURES AND PATTERNS IN SPATIAL AND REMOTELY SENSED DATA

Geospatial and remotely sensed data exhibit distinct structures. For example, the chosen extent and scale of a spatial prediction unit (ℓ in §2) has important implications for the design, use, and evaluation of geospatial ML models, evidenced by the phenomena of "modifiable areal unit problem" and "ecological fallacy" (Haining, 2009; Nikparvar & Thill, 2021; Yuan & McKee, 2022). Here, I focus on two key factors of geospatial data that affect the validity of geospatial ML evaluation methods.

3.1 SPATIAL STRUCTURES: (AUTO)CORRELATION AND COVARIATE SHIFT

One key phenomena exhibited by many geopsatial data is that values of a variable (e.g. tree canopy height) are often correlated across locations. Formally, for random process Z, the **spatial autocorrelation function** is defined as $R_{ZZ}(\ell_i, \ell_j) = \mathbb{E}[Z(\ell_i)Z(\ell_j)]/\sigma_i\sigma_j$, where σ_i, σ_j are the standard deviations associated with $Z(\ell_i), Z(\ell_j)$. For geospatial variables, we might expect $R_{ZZ}(\ell_i, \ell_j) > 0$ when ℓ_i and ℓ_j are closer together, namely that values of Z at nearby points tend to be closer in value. The degree of spatial autocorrelation in data can be assessed with statistics such as Moran's I and Geary's C, and semi-variogram analsyes (see, e.g. (Gaetan & Guyon, 2010)).

Spatial autocorrelations and correlations between predictor and label variables can be an important source of structure to leverage in geospatial ML models (Rolf et al., 2020; Klemmer & Neill, 2021), yet they also present challenges. Models can "over-rely" on spatial correlations in the data, leading to over-estimated accuracy despite poor spatial generalization performance. Overfitting to spatial relationships in training data is of particular concern in the when data distributions differ between training regions and regions of use. Such **covariate shifts** are common in geospatial data, e.g. across climate zones, or spectral shifts in satellite imagery (Tuia et al., 2016; Hoffimann et al., 2021).

Presence of spatial correlations or domain shift alone do not invalidate assessing map accuracy with a probability sample (§2.1). However, when evaluation is limited to existing data, issues of data availability and representivity can amplify the challenges of geospatial model evaluation.

3.2 AVAILABILITY, QUALITY, AND REPRESENTIVITY OF GEOSPATIAL EVALUATION DATA

Many geospatial datasets exhibit gaps in coverage or quality of data. Meyer & Pebesma (2022) evidence trends of geographic clustering around research sites primarily in a small number of countries, across three datasets used for global mapping in ecology. Oliver et al. (2021) find geographical bias in coverage of species distribution data aggregated from field observation, sensors measurements, and citizen science efforts. Burke et al. (2021) note that the frequency at which nationally representative data on agriculture, population, and economic factors are collected varies widely across the globe. While earth observation data such as satellite imagery have comparatively much higher coverage across time and space (Burke et al., 2021), coverage of commercial products has been shown to be biased toward regions of high commercial value (Dowman & Reuter, 2017).

Filling in data gaps is goal for which geospatial ML can be transformative (§2.2), yet these same gaps complicate model evaluation. When training data are clustered in small regions, this can affect our ability to train a high-performing model. When *evaluation data* are clustered in small regions, this affects our ability to evaluate geospatial ML model performance at all.

4 SPATIALLY-AWARE EVALUATION METHODS: A BRIEF OVERVIEW

In §3, we established that geospatial data generally exhibit spatial correlations and data gaps, even when target use areas $\mathcal{D}$ are small. It is well documented that calculating accuracy indices with non-spatial validation methods (e.g. standard k-fold cross-validation) will generally *over-estimate* performance in such settings. Spatially-aware evaluation methods can control the spatial distribution of training and validation set points to better simulate conditions of intended model use.

4.1 SPATIAL CROSS-VALIDATION METHODS

Several spatial cross-validation methods have been proposed that reduce spatial dependencies between train set points $\ell \in \mathcal{S}_{\text{train}}$ from evaluation set points $\ell \in \mathcal{S}_{\text{eval}}$. Spatial cross-validation methods

typically stratify training and evaluation instances by larger geographies (Roberts et al., 2017; Valavi et al., 2018) e.g. existing boundaries, spatial blocks, or automatically generated clusters. Buffered cross-validation methods (such as spatial leave-one-out (Le Rest et al., 2014), leave-pair out (Airola et al., 2019) and k-fold cross validation (Pohjankukka et al., 2017)) control the minimum distance from any training point to any evaluation point. In addition to evaluating model performance, spatial cross-validation has also been suggested as a way to to improve model selection and parameter estimation in geospatial ML (Meyer et al., 2019; Schratz et al., 2019; Roberts et al., 2017).

While separating $\ell \in \mathcal{S}_{\text{train}}$ from $\ell \in \mathcal{S}_{\text{eval}}$ can reduce the amount of correlation between training and evaluation data, a spatial split also induces a higher degree of spatial extrapolation to the learning setup and potentially reduces variation in the evaluation set labels. As a result, it is possible for spatial validation methods to systematically *under-report* performance, especially in interpolation regimes (Roberts et al., 2017; Wadoux et al., 2021). In a different flavor from the evaluation methods above, Milà et al. (2022) propose to match the distribution of nearest neighbor distances between train and evaluation set points to the corresponding distances between train set and target use area.

4.2 Other spatially-aware validation evaluation methods

When the intended use of a geospatial model is to generate predictions outside the training distribution, it is critical to test the model's ability to generalize across different conditions. For example, studies have varied the amount of spatial extrapolation required by changing parameters of the spatial validation setups in §4.1 , e.g. with buffered leave one out (Ploton et al., 2020; Brenning, 2022) and checkerboard designs (Roberts et al., 2017; Rolf et al., 2021). Jean et al. (2016) assess extrapolation ability across pairs of countries by iteratitvely training in one region and evaluating performance in another. Rolf et al. (2021) find that the distances at which a geospatial model has extrapolation power can differ substantially depending on the properties of the prediction variable.

It is always critical to put the reported performance of geospatial ML models in context. Visualizing the spatial distributions of predictions and error residuals can help expose overreliance on spatially correlated predictors (Meyer et al., 2019) and sub-regions with low local model performance. Comparing performance to that of a baseline model built entirely on spatial predictors can contextualize the value-add of a new geospatial model (Fourcade et al., 2018; Rolf et al., 2021).

5 Taking stock: Considerations and opportunities

Comprehensive reporting of performance is critical for geospatial ML methods, especially as stated gains in research progress make their way to maps and decisions of real-world consequence. Evaluating performance of geospatial models is especially challenging in the face of spatial correlations and limited availability or representivity of data. This means non-spatial data splits are generally unsuitable for geospatial model evaluation with most existing datasets. Spatially-aware validation methods are an important indicator of model performance including spatial generalization; however, they generally do not provide valid statistical estimates of prediction map accuracy. This brings us to end with three key opportunities for improving the landscape of geospatial ML evaluation:

Opportunity 1: Invest in evaluation data to measure map accuracy and overall performance of geospatial models. When remote annotations are appropriate, labeling tools (e.g. Robinson et al. (2022)) can facilitate the creation of probability-sampled evaluation datasets. Data collection and aggregation efforts can focus on filling existing geospatial data gaps (Paliyam et al., 2021) or simulating real-world prediction conditions like covariate or domain shift (Koh et al., 2021).

Opportunity 2: Invest in evaluation frameworks to precisely and transparently and report performance and valid uses of a geospatial ML model (à la "model cards" (Mitchell et al., 2019)). This includes improving spatial baselines, expanding methods for reporting uncertainty over space, and delineating "areas of applicability" for geospatial models, e.g. as in Meyer & Pebesma (2022).

Opportunity 3: If the available data and evaluation frameworks are insufficient, **explain the limitations of what types of performance can be evaluated**. Distinguish between performance measures that estimate a statistical parameter and those that indicate potential skill for a possible use case.

ACKNOWLEDGMENTS

Esther Rolf was supported by the Harvard Data Science Initiative (HDSI) and the Center for Research on Computation and Society (CRCS). Thank you to Konstantin Klemmer, Caleb Robinson, and Jessie Finocchiaro for feedback on earlier drafts of this work.

REFERENCES

Antti Airola, Jonne Pohjankukka, Johanna Torppa, Maarit Middleton, Vesa Nykänen, Jukka Heikkonen, and Tapio Pahikkala. The spatial leave-pair-out cross-validation method for reliable AUC estimation of spatial classifiers. *Data Mining and Knowledge Discovery*, 33(3):730–747, 2019.

Luigi Boschetti, Stephen V Stehman, and David P Roy. A stratified random sampling design in space and time for regional to global scale burned area product validation. *Remote sensing of environment*, 186:465–478, 2016.

Alexander Brenning. Spatial machine-learning model diagnostics: A model-agnostic distance-based approach. *International Journal of Geographical Information Science*, pp. 1–23, 2022.

Dick J Brus. Statistical approaches for spatial sample survey: Persistent misconceptions and new developments. *European Journal of Soil Science*, 72(2):686–703, 2021.

Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021.

Guanghua Chi, Han Fang, Sourav Chatterjee, and Joshua E Blumenstock. Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 119 (3):e2113658119, 2022.

Ian Dowman and Hannes I Reuter. Global geospatial data from earth observation: Status and issues. *International Journal of Digital Earth*, 10(4):328–341, 2017.

Wala Draidi Areed, Aiden Price, Kathryn Arnett, and Kerrie Mengersen. Spatial statistical machine learning models to assess the relationship between development vulnerabilities and educational factors in children in queensland, australia. *BMC Public Health*, 22(1):1–12, 2022.

Yoan Fourcade, Aurélien G Besnard, and Jean Secondi. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, 27(2):245–256, 2018.

Carlo Gaetan and Xavier Guyon. *Spatial statistics and modeling*, volume 90. Springer, 2010.

Robert Haining. The special nature of spatial data. *The SAGE Handbook of Spatial Analysis. Los Angeles: SAGE Publications*, pp. 5–23, 2009.

Júlio Hoffimann, Maciel Zortea, Breno De Carvalho, and Bianca Zadrozny. Geostatistical learning: Challenges and opportunities. *Frontiers in Applied Mathematics and Statistics*, 7:689393, 2021.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

Konstantin Klemmer and Daniel B Neill. Auxiliary-task learning for geographic data with autoregressive embeddings. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, pp. 141–144, 2021.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

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.

Sharon L Lohr. *Sampling: design and analysis*. Chapman and Hall/CRC, 2021.

Aaron E Maxwell, Timothy A Warner, and Luis Andrés Guillén. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—part 1: Literature review. *Remote Sensing*, 13(13):2450, 2021.

Hanna Meyer and Edzer Pebesma. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13(1):1–4, 2022.

Hanna Meyer, Christoph Reudenbach, Stephan Wöllauer, and Thomas Nauss. Importance of spatial predictor variable selection in machine learning applications–moving from data reproduction to spatial prediction. *Ecological Modelling*, 411:108815, 2019.

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*, 2022.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Catherine Nakalembe. Characterizing agricultural drought in the Karamoja subregion of Uganda with meteorological and satellite-based indices. *Natural Hazards*, 91(3):837–862, 2018.

Behnam Nikparvar and Jean-Claude Thill. Machine learning of spatial data. *ISPRS International Journal of Geo-Information*, 10(9):600, 2021.

Kristine Nilsen, Natalia Tejedor-Garavito, Douglas R Leasure, C Edson Utazi, Corrine W Ruktanonchai, Adelle S Wigley, Claire A Dooley, Zoe Matthews, and Andrew J Tatem. A review of geospatial methods for population estimation and their use in constructing reproductive, maternal, newborn, child and adolescent health service indicators. *BMC health services research*, 21 (1):1–10, 2021.

Ruth Y Oliver, Carsten Meyer, Ajay Ranipeta, Kevin Winner, and Walter Jetz. Global and national trends, gaps, and opportunities in documenting and monitoring species distributions. *PLoS Biology*, 19(8):e3001336, 2021.

Madhava Paliyam, Catherine Nakalembe, Kevin Liu, Richard Nyiawung, and Hannah Kerner. Street2sat: A machine learning pipeline for generating ground-truth geo-referenced labeled datasets from street-level images. In *Tackling Climate Change with Machine Learning Workshop at the International Conference on Machine Learning*, 2021.

Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature communications*, 11(1):1–11, 2020.

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.

Jonathan Proctor, Tamma Carleton, and Sandy Sum. Parameter recovery using remotely sensed variables. Working Paper 30861, National Bureau of Economic Research, January 2023. URL `http://www.nber.org/papers/w30861`.

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, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

Caleb Robinson, Anthony Ortiz, Hogeun Park, Nancy Lozano, Jon Kher Kaw, Tina Sederholm, Rahul Dodhia, and Juan M Lavista Ferres. Fast building segmentation from satellite imagery and few local labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1463–1471, 2022.

Esther Rolf, Michael I Jordan, and Benjamin Recht. Post-estimation smoothing: A simple baseline for learning with side information. In *International Conference on Artificial Intelligence and Statistics*, pp. 1759–1769. PMLR, 2020.

Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 12(1):1–11, 2021.

Patrick Schratz, Jannes Muenchow, Eugenia Iturritxa, Jakob Richter, and Alexander Brenning. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, 406:109–120, 2019.

Helen R Sofaer, Catherine S Jarnevich, Ian S Pearse, Regan L Smyth, Stephanie Auer, Gericke L Cook, Thomas C Edwards Jr, Gerald F Guala, Timothy G Howard, Jeffrey T Morisette, et al. Development and delivery of species distribution models to inform decision-making. *BioScience*, 69(7):544–557, 2019.

Insang Song and Daehyun Kim. Three common machine learning algorithms neither enhance prediction accuracy nor reduce spatial autocorrelation in residuals: An analysis of twenty-five socioeconomic data sets. *Geographical Analysis*, 2022.

Stephen V Stehman, Bruce W Pengra, Josephine A Horton, and Danika F Wellington. Validation of the us geological survey's land change monitoring, assessment and projection (lcmap) collection 1.0 annual land cover products 1985–2017. *Remote Sensing of Environment*, 265:112646, 2021.

Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE geoscience and remote sensing magazine*, 4(2):41–57, 2016.

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. *bioRxiv*, pp. 357798, 2018.

Alexandre MJ-C Wadoux, Gerard BM 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.

May Yuan and Arlo McKee. Embedding scale: New thinking of scale in machine learning and geographic representation. *Journal of Geographical Systems*, 24(3):501–524, 2022.