
Spatial Cross-Validation for Machine Learning Systems

Abygail Stegenga¹ Julia Wąsala^{1,2} Jan van Rijn¹ Mitra Baratchi¹

¹Leiden Institute for Advanced Computer Science (LIACS), Leiden University, Leiden, the Netherlands

²Space Research Organisation Netherlands (SRON), Leiden, the Netherlands

Abstract Geospatial data exhibits spatial autocorrelation (SAC), whereby observations at geographically close locations tend to have similar values. Traditional machine learning model evaluation methods, however, often fail to account for these effects, potentially resulting in information leakage between the training and test sets. In this work, a generalised approach is proposed to mitigate SAC for remote sensing image classification using blocked spatial cross-validation (BSCV). To this end, the autocorrelation range is established using global Moran’s I correlograms and entropograms. The size of the BSCV blocks is set equal to the resulting autocorrelation range, ensuring that blocks are independent to mitigate correlation between folds. In our experiments, we quantify the effects of cross-validation methods in the face of information leakage for EuroSAT, a remote sensing image classification benchmarking dataset. We compare BSCV against random cross-validation (RCV), showing that RCV exhibits consistent performance inflation, which is mitigated by BSCV. These insights demonstrate the general applicability and superiority of BSCV as an evaluation method for ML systems working with geospatial data. Consequently, the integration of BSCV in automated machine learning systems and the effect of spatially-aware sampling on pre-training foundation models are identified as rich avenues to build upon in future work.

1 Introduction

Model evaluation aims to approximate the true prediction error of a model as accurately as possible using unseen data at test time (Rodriguez et al., 2009). Machine learning (ML) is increasingly being integrated into real-world domains using geospatial datasets. Geospatial data, however, often exhibits spatial autocorrelation (SAC), where geographically close points have more similar values compared to points further away. This poses a challenge for conventional ML pipelines when applied to practical scenarios. In particular, SAC creates a substantial risk of data leakage between the training and test sets when working with geospatial data (Apicella et al., 2025). Since random sampling methods during model evaluation do not take into account transfer of information due to geographical proximity, the resulting spatially non-disjoint train-test splits will contain information overlap. As a result, the measured model performance can be inflated, leading to the preference of suboptimal models during selection (Roberts et al., 2017). Currently, the possibility of autocorrelation and its effects are largely ignored in general ML pipelines, resulting in consistently underestimated prediction errors for AI models in geospatial applications.

Spatial cross-validation (SCV) is a model evaluation method that accounts for SAC. SCV separates geographically close samples into blocks, splitting the area that a dataset covers into subregions of a certain size. By assigning all observations within a subregion to the same fold, the validation folds can once again be considered to be independent from the training folds. Despite the multitude of available approaches to spatial autocorrelation mitigation, a widely applicable SCV method is missing in the current literature, since any specific type of technique may be more or less suitable depending on the application and the dataset at hand. This work ultimately aims to provide a generalised strategy for dealing with SAC in the existing ML pipeline. We identify blocked spatial

cross-validation (BSCV) (Meyer et al., 2019; Ploton et al., 2020) as a broadly applicable and flexible approach, implementing it in a way that allows for fair model evaluation in remote sensing image classification. Moreover, we suggest the general application of entropograms and global Moran’s I indicator correlograms for identifying appropriate block sizes for such tasks. Furthermore, the strengths and weaknesses of different cross-validation strategies for general ML systems in remote sensing contexts remain underexplored (Wąsala et al., 2026). Since the norm for performance estimation in ML is currently RCV, this work simultaneously aims to provide a more complete picture for model error computation on geospatial datasets.

2 Methods

This work proposes a unified way to integrate spatial cross-validation, specifically BSCV, into the general ML pipeline. This process consists of 2 steps: 1) determining the autocorrelation range, and 2) constructing spatially disjoint folds based on a geographical grid. Finally, we specify how the true model error for a given dataset is approximated based on the application.

2.1 Determining autocorrelation range

Spatial cross-validation strategies are only necessary if a dataset exhibits spatial autocorrelation of the target variable among instances to begin with. Since the options for quantifying the autocorrelation range for categorical targets are relatively limited, we compare two complementary pre-existing options to detect SAC and determine its autocorrelation range: the entropogram (Zhang et al., 2023) and the global Moran’s I correlogram (Rosenberg, 2000). The block size in the blocked spatial cross-validation is set to be equal to the computed autocorrelation range. With the idea of proposing a general approach applicable to any arbitrary geospatial classification problem, we adopt these techniques in specific ways.

The entropogram expresses SAC in terms of mutual information over the entire dataset. The mutual information is computed for each pair of samples that are separated by a distance h and summed together. This process is repeated for various separation distances to finally quantify how much uncertainty is explained by the influence of samples progressively further away, which is referred to as the entropogram. Samples at distances beyond where the mutual information is known to be zero do not explain any additional uncertainty, and correspond to maximal SAC. This point is then established as the autocorrelation range.

Global Moran’s I is originally only quantifiable for numerical variables. As such, we convert the categorical values into binary values per label with an indicator function (Carr et al., 1985). In this process, a new dataset is created for each label by treating it as a binary label in a one-versus-all approach. The global Moran’s I statistic can then be computed for each of these binarised versions of the dataset, resulting in a correlogram for each class label. To compute the autocorrelation range, we obtain Moran’s I for various distances for each target label at hand. The autocorrelation range is then set to the distance where Moran’s I is zero.

2.2 Blocked spatial cross-validation

Having determined the autocorrelation range based on the entropogram and correlograms, spatial cross-validation is implemented. The SCV strategy adopted in this work is blocked K -fold cross-validation, otherwise known as blocked spatial cross-validation (BSCV) (Ploton et al., 2020; Meyer et al., 2019), due to its lack of prior assumptions about a dataset and its flexibility. We establish the blocks for K folds by gridding the area spanned by the dataset’s observations into B blocks of size $r \times r$, with r being the autocorrelation range. Next, these distinct blocks are combined into K folds. Considering that blocks will contain groups of samples with the same label due to SAC,

randomly assigning blocks to folds yields unbalanced folds, both in terms of fold size and class imbalance. Optimal block combinations are obtained through a random search, where various combinations of blocks are generated, and the best configuration is adopted. Here, the criterion for an optimal configuration is two-fold: 1) equal fold sizes, and 2) equal label balance per fold, similar to stratified random K -fold cross-validation. The resulting folds will likely never contain exactly the same number of instances, nor will their class distributions be equal to that of the training set. As such, we undersample the majority classes in all folds until the total number of instances in each fold is equal to the smallest fold. This process ensures that each fold now contains an equal number of samples and is distributed more similarly to both the training distribution and the other folds. Since the individual data points are separated by at least the autocorrelation range, blocks are now considered to be independently distributed.

2.3 Approximating model prediction error

To determine whether standard ML evaluation protocols are overoptimistic, we need to compare their resulting performance estimate to the true prediction error of the model. This true model error is inherently unknown, however. We empirically establish an independent and reliable proxy of the true error of the model based on the data at hand. We refer to this proxy as the representative performance. Whereas the test set performance is computed by the cross-validation methods to estimate the model error, the representative performance is considered to be the true model error and is computed on data points kept strictly separate for this purpose alone. To securely investigate the generalisation capability of the trained model to applications within the distribution of a given dataset, the samples used to compute the representative performance need to be geographically separate from the remaining dataset. The test set for the representative performance is created by arbitrarily selecting 20 representative points and adding all surrounding points within 5.56 km to simulate cities. Subsequently, any samples in the remaining dataset within a distance corresponding to the autocorrelation range are omitted to avoid any information leakage in the representative performance due to spatial dependencies. We train a model on samples in the remaining spatially disjoint dataset and compute the representative performance on the selected sample clusters. The cross-validation performance estimates are directly compared to the representative performance. The training and test sets used for the representative performance are subsampled to contain 9 and 1 fold’s worth of data, respectively. This process is repeated 10 times to yield a robust and fair estimate of the true model performance.

3 Experimental Setup

The benchmarking of the cross-validation approaches is carried out for the EuroSAT dataset (Helber et al., 2019). We specifically focus on image classification to tackle the complexity of categorical values in spatial cross-validation. Here, random cross-validation (abbreviated as RCV) is taken as a baseline for comparison. Since BSCV includes class balancing, stratified random cross-validation (SRCV) is also included to ensure fair comparison. All cross-validation strategies use $K = 10$ folds in all experiments of this work. The size of each of these folds is ensured to be identical across the approaches. The benchmarking is approached with a ResNet-50 classifier. We adopt a learning rate of $1e - 4$, batch size of 64, ADAM optimiser, and CrossEntropy loss function.

4 Results: Blocked spatial cross-validation better represents model performance

Both the entropogram (Figure 1a) and the correlogram (Figure 1b) clearly demonstrate the presence of strong positive spatial clustering, characterised by high mutual information and positive global Moran’s I values at small scales. The spatial autocorrelation decreases more rapidly in the correlogram compared to the entropogram since uncertainty between different classes is not taken

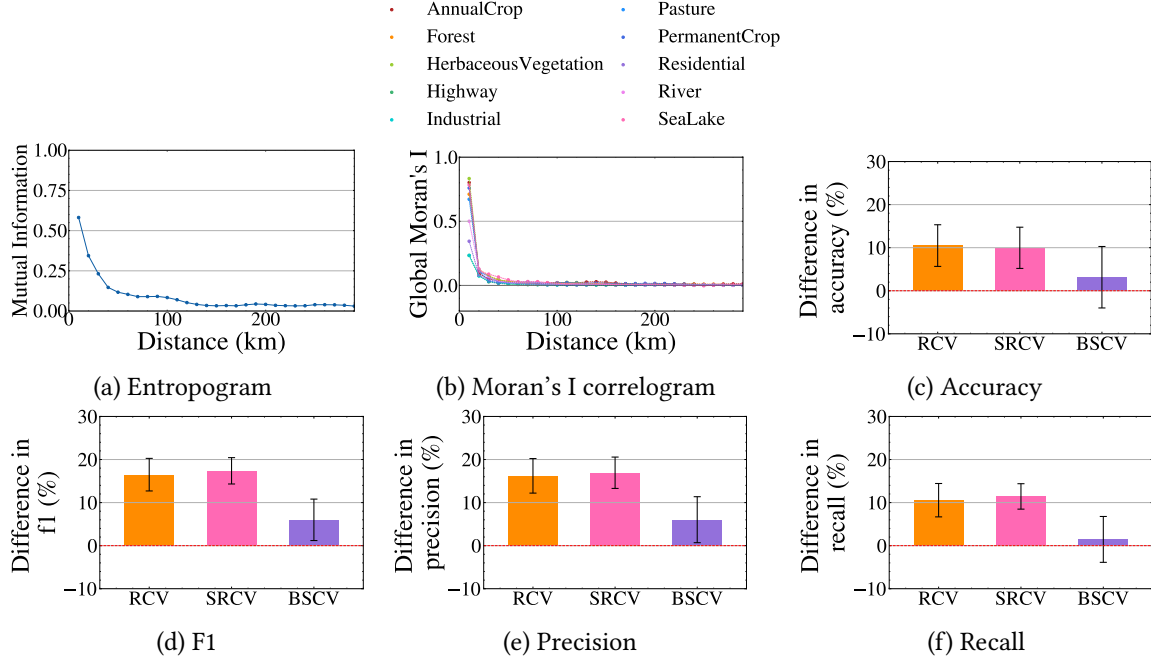


Figure 1: **a-b:** Autocorrelation as a function of geographical distance for the EuroSAT dataset displayed as a normalised entropogram and global Moran's I correlogram.
c-f: Performance inflation of a trained ResNet-50 on the EuroSAT dataset obtained using random cross-validation (RCV), stratified random cross-validation (SRCV) and blocked spatial cross-validation (BSCV). The difference in performance between the evaluation methods and representative performance is computed in terms of accuracy, F1-score, precision and recall. Error bars correspond to the standard deviation. BSCV uses a block size of $1.3^\circ/145$ km.

into account for Moran's I, demonstrating the superiority of the entropogram. The Moran's I correlograms, however, allow for insight into the autocorrelation landscape at class-level. Based on the entropogram's plateauing at a mutual information of zero from a distance of 145 km onwards, we establish this as the autocorrelation range.

Figures 1c - 1f illustrate the extent of the performance inflation. Positive values here indicate overoptimism, and negative values indicate overpessimism of an evaluation method compared to the representative performance. Across all metrics, BSCV is seen to be more indicative of the true model error. In fact, the representative performance (red dotted line) is seen to fall at most 2% outside the uncertainty boundaries for each metric, indicating that BSCV yields an accurate representation of the true model performance. In line with expectations, both RCV and SRCV show significant overoptimism, even up to 17% in terms of F1-score. As such, we conclude that random sampling-based evaluation protocols do not provide a realistic picture of a model's predictive capabilities on the EuroSAT dataset. Considering that the only difference between stratified random cross-validation and random cross-validation is the presence of class balancing across folds, a performance difference between the two effectively illustrates whether the overoptimism is partially due to class imbalance. The lack of difference in performance inflation between RCV and SRCV demonstrates here that class imbalance is not the root cause of the observed gap between traditional cross-validation and the true model performance. In fact, SRCV appears to be marginally more overconfident than the RCV variant. A potential explanation is that the fold balancing requires samples to be taken from dense regions of a specific label. As such, some additional spatial autocorrelation may be introduced during this process. The lack of effect observed in preliminary

experimentation with loss weights confirmed that the inflation is not due to class imbalance.

5 Conclusion

To conclude, the combination of these results shows that blocked spatial cross-validation is a more suitable performance evaluation method for representing interpolation applications based on the EuroSAT dataset. The reported performance of blocked spatial cross-validation is a good representation of the generalisability of the model, whereas traditional random cross-validation is considerably overoptimistic in the face of information leakage. Consequently, investigating the influence of different cross-validation methods on automated machine learning systems and sampling in pre-training foundation models is considered to be crucial.

References

Apicella, A., Isgro, F., and Prevete, R. (2025). Don't push the button! exploring data leakage risks in machine learning and transfer learning. *Artificial Intelligence Review*, 58(11):339.

Carr, J., Bailey, R., and Deng, E. (1985). Use of indicator variograms for an enhanced spatial analysis. *Journal of the International Association for Mathematical Geology*, 17(8):797–811.

Helber, P., Bischke, B., Dengel, A., and Borth, D. (2019). Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226.

Meyer, H., Reudenbach, C., Wöllauer, S., and Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications—moving from data reproduction to spatial prediction. *Ecological Modelling*, 411:108815.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature communications*, 11(1):4540.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Rodriguez, J. D., Perez, A., and Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):569–575.

Rosenberg, M. S. (2000). The bearing correlogram: a new method of analyzing directional spatial autocorrelation. *Geographical Analysis*, 32(3):267–278.

Wąsala, J., Maasakkers, J. D., Schneider, R., Hoos, H. H., Baratchi, M., et al. (2026). Automl to advance earth observation. *Journal of Data-centric Machine Learning Research*.

Zhang, W.-B., Ge, Y., Bai, H., Jin, Y., Stein, A., and Atkinson, P. M. (2023). Spatial association from the perspective of mutual information. *Annals of the American Association of Geographers*, 113(8):1960–1976.