Remote Sensing of Environment 319 (2025) 114585

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Accounting for spatial variability with geo-aware random forest: A case study for US major crop mapping

Yiqun Xie [a,*], Anh N. Nhu [a], Xiao-Peng Song [a], Xiaowei Jia [b], Sergii Skakun [a], Haijun Li [a], Zhihao Wang [a]

[a] University of Maryland, 7251 Preinkert Drive, College Park, 20742, MD, United States
[b] University of Pittsburgh, 210 S. Bouquet Street, Pittsburgh, 15260, PA, United States

ARTICLE INFO

Edited by Marie Weiss

Dataset link: https://github.com/ai-spatial/STAR, https://cloud.google.com/storage/docs/public-datasets/sentinel-2, https://search.earthdata.nasa.gov/, https://www.nass.usda.gov/Research_and_Science/Cropland/Release/index.php

Keywords:
Random forest
Geo-RF
Spatial variability
Remote sensing
Crop classification

ABSTRACT

Spatial variability has been one of the major challenges for large-area crop monitoring and classification with remote sensing. Recent works on deep learning have introduced spatial transformation methods to automatically partition a heterogeneous region into multiple homogeneous sub-regions during the training process. However, the framework is only designed for deep learning and is not available for other models, e.g., decision tree and random forest, which are frequently the models of choice in many crop mapping products. This paper develops a geo-aware random forest (Geo-RF) model to enable new capabilities to automatically recognize spatial variability during training, partition the space, and learn local models. Specifically, Geo-RF can capture spatial partitions with flexible shapes via an efficient bi-partitioning optimization algorithm. Geo-RF also automatically determines the number of partitions needed in a hierarchical manner via statistical tests and builds local RF models along the partitioning process to explicitly address spatial variability and improve classification quality. We used both synthetic and real-world data to evaluate the effectiveness of Geo-RF. First, through the controlled synthetic experiment, Geo-RF demonstrated the ability to capture the artificially-inserted true partition where a different relationship between the inputs and outputs is used. Second, we showed the improvements from Geo-RF using crop classification for five major crops over the contiguous US. The results demonstrated that Geo-RF is able to significantly improve classification performance in sub-regions that are otherwise compromised in a single RF model. For example, the partition around downstream Mississippi for soybean classification led to major improvements for about 0.10-0.25 in F1 scores in the area, and the score increased from 0.57 to 0.82 at certain locations. Similarly, for rice classification, the partition in Arkansas led to F1 scores increasing from 0.59 to 0.88 in local areas. In addition, we evaluated the models under different parameter settings, and the results showed that Geo-RF led to improvements over RF in the vast majority of scenarios (e.g., varying model complexity and training sizes). Computationally, Geo-RF took about one to three times more training time while its execution time during testing was similar to that of RF. Overall, Geo-RF showed the ability to automatically address spatial variability via partitioning optimization, which is an important skill for improving crop classification over heterogeneous geographic areas at large scale. Future research can explore the use of Geo-RF for other geographic regions and applications, interpretable methods to understand the data-driven partitioning, and new designs to further enhance the computational efficiency.

1. Introduction

Spatial variability (a.k.a., spatial heterogeneity) has been one of the major challenges for large-area crop monitoring and mapping with remote sensing. The variability indicates that the distributions of input features **X** (e.g., satellite-derived reflectance values in the optical domain and backscatter coefficient in the microwave domain), output targets **y** (e.g., crop types), and the functional relationships between **X** and **y** can all vary over space. In particular, variations in the functional relationships $\mathbf{X} \rightarrow \mathbf{y}$ over geographic regions make it extremely difficult to learn a single model to approximate those different relationships (Goodchild and Li, 2021; Xie et al., 2021a; Atluri et al., 2018). Technically, one of the causes of such variations is the existence of unobserved variables that are related to the output targets. For example, readings from satellite sensors such as Sentinel-2 often only carry partial or aggregated information of the entire physical process

* Corresponding author.
E-mail address: xie@umd.edu (Y. Xie).

https://doi.org/10.1016/j.rse.2024.114585
Received 14 May 2024; Received in revised form 21 November 2024; Accepted 19 December 2024
Available online 12 February 2025

from field crops to the sensors, and some variables that contribute to the process may not be observed or fully reflected, such as soil moisture, temperature, crop health status, applications of pesticides, etc. However, the values of such unobserved variables are most likely not a fixed constant over geographic regions (Goodchild and Li, 2021), leading to changes in the optimal functional relationships – as learned by a machine learning algorithm – between observed variables **X** and **y** in different regions. This intrinsic property of geographic data and phenomena violates the classic independence and identical distribution (i.i.d.) assumption in most machine learning algorithms. In practice, this means that a single model is used to represent all data samples despite their potential differences across locations. While this is not ideal, it is still the most common practice in machine learning applications for crop mapping due to the convenience and lack of automated methods to recognize and address the variability issue. This gap may significantly limit a model's ability to reach its best performance, and can easily generate maps with variable quality of performance across regions.

In related work, various research efforts have tried to bridge this gap for different types of application scenarios. Geographically-weighted regression (GWR) is a traditional approach to explicitly model spatial variability (Brunsdon et al., 1999; Fotheringham et al., 2017). It can be considered as a "spatial" special-case of local regression, i.e., a nonparametric model that learns a local model for each different location in the dataset. However, GWR is designed for real-valued linear inference and is not suitable for modeling complex nonlinear relationships between satellite signals and crop classes. There have also been recent extensions of GWR to support non-linear models, such as the geographically-weighted random forest (GWRF). Similarly, GWRF learns a different local random forest (RF) at each location in the dataset (Georganos et al., 2021; Luo et al., 2022; Grekousis et al., 2022). However, this exhaustive generation of local models leads to a very high computational cost. As a result, GWRF has only been applicable to very coarse-granularity problems, such as county-level regression (i.e., with tens of counties where each county is one data point), and is not computationally feasible for most satellite remote sensing tasks, where pixel-level predictions are needed. One reason behind the exhaustive local model design is the lack of methods to capture the spatial footprints of variability patterns, e.g., which locations share the same or very similar $\mathbf{X} \rightarrow \mathbf{y}$ relationships and which regions are different?

To address this issue, recent studies in deep learning have introduced a spatial transformation framework (Xie et al., 2021a, 2023). This framework automatically partitions a heterogeneous region into multiple homogeneous sub-regions during the training process using statistics calculated from the performance of the deep learning models over space. However, spatial transformation is designed for deep learning and it includes design decisions (e.g., network-layer-based parameter sharing) that are specifically tailored for deep neural networks. The approach was evaluated using several examples of applications, including land cover classification and human mobility estimation. The land cover classification was carried out in a 80 km×80 km study area in California, and the model has an average F1 score at about 0.7. This framework is currently not available for non-deep-learning models, e.g., decision tree and RF.

On the other hand, these traditional machine learning models are still important and frequently used in satellite data processing to generate agriculture-related geoinformation products at national or global scales. For example, the US Department of Agriculture (USDA) and Agriculture and Agri-Food Canada (AAFC) use decision tree models to create annual crop type maps, Cropland Data Layer (CDL) (Boryan et al., 2011; CDL, 2024) and Annual Crop Inventory (Fisette et al., 2013), respectively. The European Space Agency-funded system, the Sentinel-2 for Agriculture (Sen2-Agri) (Defourny et al., 2019), uses a RF classifier to map crop types based on multi-temporal Sentinel-2 and Landsat −8 images. These classifiers are also among the most popular choices in the Google Earth Engine platform (Gorelick et al., 2017; Phalke et al., 2020; Xuan et al., 2023). In addition, RF remains widely used in recent studies on crop mapping. For example, Blickensdörfer et al. (2022) explored time-series crop mapping using RF under conditions with strong inter-annual meteorological variability. Wang et al. (2019) and Zhang et al. (2022) explored the use of RF in scenarios with no direct field-level labels. Efforts have also utilized RF to develop high-resolution crop maps using Sentinel-1/2, Landsat −8 or multi-source imagery in different geographies, including Brazil (Pott et al., 2021), China (Li et al., 2023), Europe (d'Andrimont et al., 2021), South Africa (Mpakairi et al., 2023) and broader regions (Phalke et al., 2020).

From the methodological point of view, tabular data representations are still very common in the practical applications of remote sensing (e.g., each data point with its **X** being the spectral band values of a pixel, and its **y** being the crop type of the pixel). In this realm, tree-based models still often outperform deep learning models as demonstrated by various evaluations (Grinsztajn et al., 2022; Shwartz-Ziv and Armon, 2022). With that said, the scope of the present study is not to compare traditional machine learning and deep learning models for crop classification. The ranking of traditional machine learning and deep learning models remains an open research question and they may be complementary to each other as model choices in different scenarios. This work focuses on RF, which remains a common choice for pillar remote sensing products and related Earth Science research, and aims to provide an enhanced variability-aware learning framework for this important model.

This work aims to develop a geo-aware RF (Geo-RF) by incorporating RF into the spatial transformation framework, with design decisions tailored based on the characteristics of the tree-based ensemble model. Compared to RF, Geo-RF aims to provide new capabilities to automatically recognize spatial variability, partition the space, and learn local models to break the dilemma a single model faces between different heterogeneous regions. Our experiments used both synthetic data and US major crop classification data to evaluate the effectiveness of Geo-RF. We carried out a case study using crop classification in the contiguous US (CONUS). The input **X** consists of 10 spectral bands from Sentinel-2 over 33 timestamps in 2021 with 3 additional terrain attributes, and the sample locations are 1 km apart expanding over the entire CONUS area. The labels for five example major crops (corn, soybean, wheat, rice, and cotton) are collected from the US CDL. We also carried out an extensive sensitivity analysis to better evaluate and understand Geo-RF's behavior in different scenarios.

2. Methods

In this section, we present details of the technical methods used to construct Geo-RF, which builds on our previous deep-learning-based spatial transformation framework (Xie et al., 2021a). Specifically, Section 2.1 provides an overview of variability-aware learning. Section 2.2 introduces the hierarchical process for partitioning the space into any number of partitions – as needed – where each step performs a bi-partitioning of a given area. Section 2.3 describes methods to optimize the bi-partitioning based on the performance statistics obtained by the current RF model over different locations. Section 2.4 discusses the use of statistical tests to verify the variability between sub-partitions and its impact on the performance of RF models. Finally, Section 2.5 explains the training and selection of the local RF models for the partitions.

2.1. Overview of variability-aware learning and Geo-RF

First, we provide an overview of different strategies to account for spatial variability when using machine learning. Fig. 1 shows a comparison of three paradigms for variability-aware learning. The first paradigm in Fig. 1 (left) requires a pre-defined or known partitioning that separates locations with different functional relationships $\mathbf{X} \rightarrow \mathbf{y}$ into different partitionings. This could be a suitable option when the input data are collected from a few discrete locations that are far

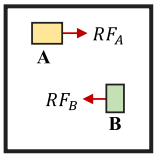
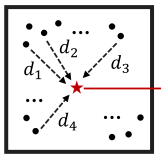
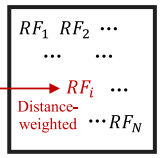
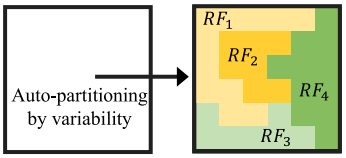
RF w. Predefined Partitions	Geographically-weighted RF	Geo-RF
Predefined partitions are required as inputs, which correctly separate different $\mathbf{X} \rightarrow \mathbf{y}$. A RF is learned for each partition.	A RF is learned for every location with distance-weighted samples. This is not suitable for data with a large number of different locations.	The space is automatically partitioned as necessary to separate regions with different $\mathbf{X} \rightarrow \mathbf{y}$ that are detected during training. A RF is learned for each partition.
 Predefined partitions	 Training points  RF at every location	 Predefined partitions

Fig. 1. An overview of different paradigms to address spatial variability. GWRF is often used for problems involving a small number of discrete locations, and is not computationally feasible for many remote sensing problems covering broad areas.

apart (Gupta et al., 2020, 2021) and is limited for general scenarios where the problem covers a contiguous large area (e.g., CONUS) and an optimal partitioning for different $\mathbf{X} \rightarrow \mathbf{y}$ is unknown. Fig. 1 (middle) shows the geographically-weighted paradigm, which is a non-parametric approach (Georganos et al., 2021; Luo et al., 2022; Grekousis et al., 2022). Instead of learning a fixed set of parameters for the entire region, it learns a local model at every location in the study area when making classifications/predictions, with the assumption that the similarity of functional relationships $\mathbf{X} \rightarrow \mathbf{y}$ depends only on geographical distances. This paradigm addresses a major drawback of the first paradigm by removing the requirement of an input partitioning of $\mathbf{X} \rightarrow \mathbf{y}$, which is often problem-dependent, data-dependent and unavailable in general use cases. However, a major limitation of this nonparametric approach is the computation, due to which the vast majority of its applications are for problems with a relatively small number of locations (Georganos et al., 2021). The computational cost will become largely infeasible for large-scale and high-resolution mapping tasks in remote sensing, including crop mapping, with millions or billions of spatial units for classification.

Finally, Fig. 1 (right) shows the spatial transformation paradigm adopted for this work on Geo-RF. The goal of this paradigm is to address the computational bottleneck posed by the exhaustive nonparametric models. Building on top of spatial transformation, Geo-RF is a top-down approach that will start with a single global model, and adaptively partition the space – only as needed – based on the estimated differences in functional relationships $\mathbf{X} \rightarrow \mathbf{y}$ during the training process. In this paradigm, local models are only learned for the partitions. As a simple example, if Geo-RF does not identify any significant heterogeneity in the functional relationships, it will just return a single RF model that is the same as the global RF model trained without using Geo-RF. This both avoids the need for a known partitioning and can effectively reduce unnecessary local models to make it computationally efficient and practical.

2.2. Hierarchical bi-partitioning of space

For general real-world scenarios, the number of spatial partitions needed to separate different functional relationships $\mathbf{X} \rightarrow \mathbf{y}$ is unknown. Thus, Geo-RF needs to have the ability to generate a flexible number of partitions as needed for a given problem. To achieve this goal, Fig. 2 shows a hierarchical process, where each step in the hierarchy carries out a bi-partitioning of the space. Starting from the entire input study area, the hierarchical bi-partitioning process continues at child-partitions until no further variability or heterogeneity is identified (determined by statistical tests in Section 2.4), allowing the number of partitions to grow as needed. Further, as shown in the middle row of Fig. 2, the RF model grows from a single global model to a collection of local RF models as the space gets partitioned. The last row of Fig. 2 shows the expected improvements in model performance as heterogeneous regions are partitioned and local models are built. Technically, the reason that a single RF is not able to get a better performance is due to the conflicts between regions with different functional relationships $\mathbf{X} \rightarrow \mathbf{y}$. These conflicts often lead to dilemmas for the single model, as reducing misclassification for samples at some locations may lead to more errors at others. The gradual improvements in model performance mean that the conflicts preventing the RF model from further improvements are addressed by the partitioning.

The two key building blocks left now are the methods to: (1) find an optimal bi-partitioning at each step; and (2) determine if the bi-partitioning really represents variability and can lead to meaningful improvement in model performance. Fig. 3 summarizes the key steps in the hierarchical partitioning workflow of Geo-RF, and the methods will be detailed in the following Sections 2.3 and 2.4.

2.3. Bi-partitioning optimization using statistics of RF performance

There are two essential components in identifying an optimal bi-partitioning of the space: (1) **A score function** to quantify and rank different candidate bi-partitioning schemes based on the potential differences between the functional relationships $\mathbf{X} \rightarrow \mathbf{y}$ of the two partitions; and (2) **A partitioning enumeration algorithm** to efficiently search and explore different candidate partitioning schemes, and find the optimal one based on the score function. Since the footprints of partitions can be arbitrary due to the underlying physical and social contexts, the search algorithm needs to have the ability to enumerate partitioning schemes with flexible shapes. For these two components, we apply our previous statistical framework developed for deep learning models (Xie et al., 2021a), and introduce it in the context of RF.

2.3.1. Score function based on a multivariate scan statistic

As the objective of the score function is to quantify whether there are potential differences between the functional relationships $\mathbf{X} \rightarrow \mathbf{y}$ of data points in the two partitions, we start with two hypotheses to facilitate the design of the function: (1) **Null hypothesis** H_0: The functional relationships $\mathbf{X} \rightarrow \mathbf{y}$ between the two (sub-)partitions are the same (i.e., no need for the partitioning to happen), and (2) **Alternative hypothesis** H_1: The functional relationships are different. Based on this, the multivariate scan statistic (MSS) is used to measure the differences. MSS (Neill et al., 2013; Xie et al., 2021b; Kulldorff et al., 2007), as an extension of the spatial scan statistic (Kulldorff, 1997), is a widely applied spatial statistical approach to separate regions with different patterns in event detection (e.g., disease surveillance). It identifies if there exists a spatial region with a significantly higher rate of generating incidents or cases of certain events (e.g., disease, crime) compared to the rest.

On the input side, MSS takes binary values of different event types. For example, in disease monitoring, whether an individual has a certain

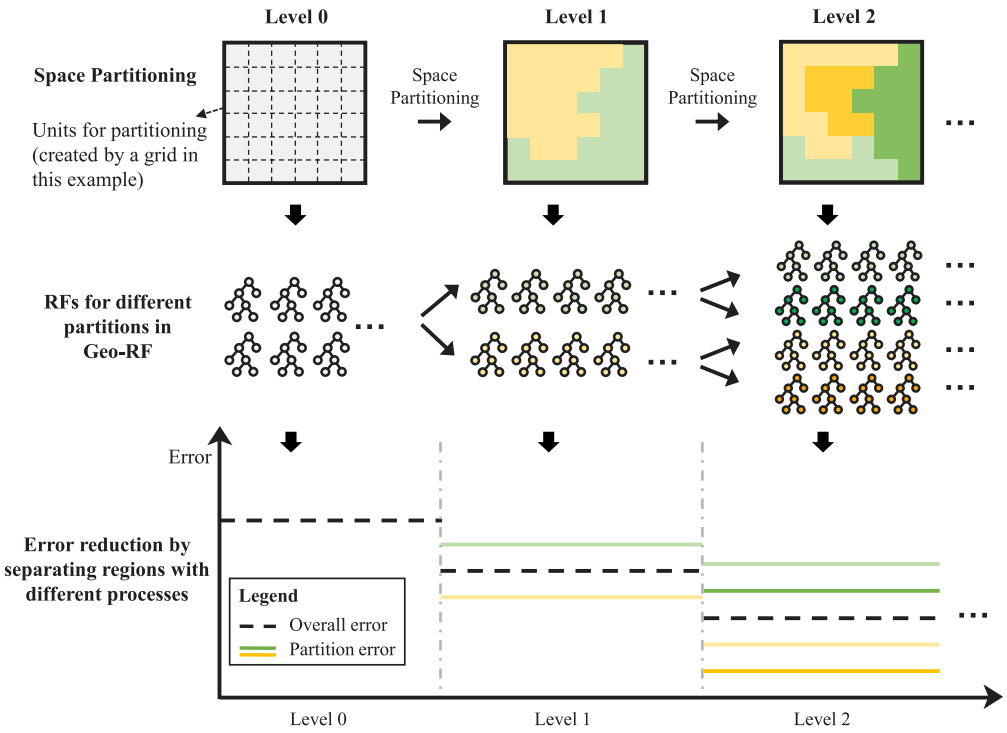


Fig. 2. A high-level illustration of the Geo-RF framework. Geo-RF uses a hierarchical partitioning process and the final number of partitions is determined through statistical tests. New RFs are created for each new partition to allow different functional relationships $\mathbf{X} \rightarrow \mathbf{y}$ to be learned for data following different processes. This further allows gradual enhancement of classification quality as illustrated in the last row. The minimum units used for space-partitioning can be specified by users, and by default Geo-RF uses a grid to create the units, which is described in Section 2.3.2.

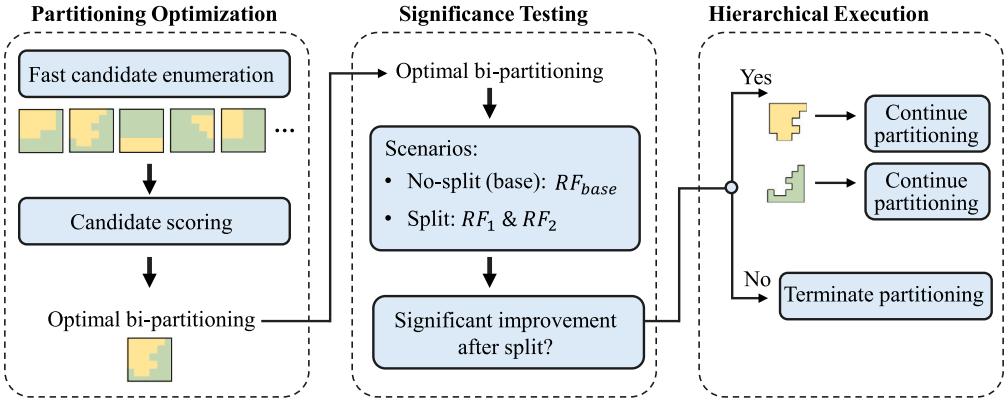


Fig. 3. The workflow chart describing the key steps of Geo-RF, including bi-partitioning optimization, significance testing and the hierarchical partitioning process to obtain a flexible number of partitions. The number of partitions needed is controlled by the significance testing procedure.

symptom is defined by a binary value. In the context of Geo-RF, the event is represented by prediction errors, i.e., whether the model made a mistake classifying a sample point. The prediction errors are estimated using samples from a validation set, which is formed by a subset of samples separated out from the training set. Specifically, we held out 20% of the training samples for the validation. To avoid confusion, there is no overlap between this validation set and the test set, where the test set is reserved for the final performance evaluation and completely excluded from the training process. Denote RF^{pre} as the RF model trained using all data points in the current area (i.e., prior to bi-partitioning). A data point is marked as "1" if RF^{pre} makes an incorrect classification at the point, and otherwise "0". If all data points from the current area follow homogeneous or similar functional relationships $\mathbf{X} \rightarrow \mathbf{y}$, we expect the error distribution for each class – based on the predictions by the single model RF^{pre} – to be more homogeneous over the area. On the contrary, if the data points follow different functional relationships $\mathbf{X} \rightarrow \mathbf{y}$ in the area, we expect the error distribution predicted by this single model to be more heterogeneous. This allows us to establish a proxy of the functional relationships $\mathbf{X} \rightarrow \mathbf{y}$ using the classification performance of the RF model, which are otherwise not directly observable from the dataset itself with only $\mathbf{X}$ and $\mathbf{y}$.

Under this representation, denote $c_{k,m}$ and $b_{k,m}$ as the observed and expected (baseline) number of errors for class m at a spatial location s_k, respectively; where $m = 1, \ldots, M$ and locations will be defined by the spatial units used for partitioning (e.g., grid cells) which will be formally defined in Section 2.3.2. The expectation or baseline $b_{k,m}$ can be calculated by $b_{k,m} = C_m \cdot \frac{n_{k,m}}{N_m}$, where C_m is the total number of misclassified samples of class m, $n_{k,m}$ is the total number of data points of class m at location s_k, and N_m is the total number of data points in the current area. If $c_{k,m}$ is similar to $b_{k,m}$, then there is likely nothing unusual for the current area. Based on these, we can

obtain a concrete mathematical formulation of our null and alternative hypotheses H_0 and H_1 introduced earlier. Specifically, H_0 states that the error distribution is homogeneous over space, and H_1 states that there exists a sub-region S where the rate of generating errors is q_m times the expected rate under H_0, i.e., the expectation in S is $q_m \cdot b_{k,m}$. Note that this sub-region S, if truly exists, will be used to bi-partition the space into to two sub-regions.

MSS estimates the log likelihood ratio[1] of the two hypothesis H_1 and H_0 based on Eq. (1). The larger the likelihood ratio, the more likely the alternative hypothesis H_1 (i.e., variability) is true compared to the null hypothesis H_0 (i.e., no variability). In order to estimate the likelihoods, MSS adopts the standard Poisson model for discrete point process (Neill et al., 2013; Kulldorff et al., 2007), where the expectation of the Poisson distribution at each location s_k is $b_{k,m}$ under H_0. Based on this, we get the log likelihood ratio as follows:

$$\log LR(S) = \log \frac{\text{Likelihood}(H_1, S)}{\text{Likelihood}(H_0, S)} = \log \frac{\prod_{s_k \in S} \prod_{m=1}^{M} \Pr(c_{k,m} \sim \text{Poisson}(q_m \cdot b_{k,m}))}{\prod_{s_k \in S} \prod_{m=1}^{M} \Pr(c_{k,m} \sim \text{Poisson}(b_{k,m}))}$$
$$= \sum_{m=1}^{M} \left(C_{m,S} \log q_m + B_{m,S}(1 - q_m) \right) \tag{1}$$

where S is the sub-region used to bi-partition the current area S_{all} into S and $(S_{all} - S)$, $C_{m,S} = \sum_{s_k \in S} c_{k,m}$, and $B_{m,S} = \sum_{s_k \in S} b_{k,m}$. The derivation is provided in Appendix A.1 in the supplementary materials.

Next, the optimal estimation of the only unknown q_m in the likelihood can be obtained by maximizing the likelihood, i.e., the maximum likelihood estimate, with the solution:

$$q_m = \frac{C_{m,S}}{B_{m,S}} \tag{2}$$

With the score formulated, the next goal is to find the optimal sub-region S^* that maximizes the score. This optimal sub-region S^* represents the region where the largest divergence of error distribution to the expected distribution is observed statistically.

2.3.2. Searching for the optimal bi-partitioning

We leverage the fast subset scanning approach to find the optimal sub-region S^* for the bi-partitioning, where S^* is one partition and the rest forms the other. This method offers a fast-search algorithm to overcome the computational challenge that the number of sub-region candidates is exponential to the number of locations. In short, the approach will transform the search space from an exponential space to a linear space to substantially reduce the computational cost.

The one and only prerequisite of this method is to construct local groups of data points (e.g., pixels in a local grid cell as one group). These groups can be considered as the spatial units used for the partitioning and they are needed to calculate several basic statistics of errors to enable the fast optimization process. In Geo-RF, we overlay a grid on top of the study area, and all pixels within each grid cell form a local group. This means that region S^* will be formed by a set of grid cells, and the flexibility of the shape of S^* depends on the size of the grid cells. In this work, we use a cell size of $0.5° \times 0.5°$ and our grid for CONUS has a size of 50×116, which allows the Geo-RF to create fairly flexible shapes of partitions as shown later in the experiments in Section 4.2. Based on this grid, in Eq. (1), each location s_k now represents a grid cell. Fig. 2 (first row) shows an example visualization of the grid cells and partitions based on them.

With the grid cells s_k as local groups, we use the linear-time subset scanning (LTSS) property (Neill et al., 2013) to find the optimal sub-region S^*. Specifically, for Geo-RF's score function $\log LR(S)$ in Eq. (1), it has been shown that S^* must be formed by the cells s_k with the highest rankings by the following ranking function (Neill et al., 2013; Xie et al., 2021a):

$$\gamma(s_k) = \sum_{m=1}^{M} (c_{k,m} \log q_m + b_{k,m}(1 - q_m)) \tag{3}$$

While the number of highest ranking cells is not certain, given N cells there are at most N possible numbers to consider (i.e., top 1, top 2, ..., top N), so the search space is now linear to the number of cells. However, one remaining issue is that we do not really know the right q_m values to use here, as we need to know what the region S^* is in the first place to calculate the corresponding q_m values using Eq. (2). This means that to find S^* using Eq. (3) we need to know q_m, but to know q_m we need to first know S^*, which becomes an endless loop. To break the loop, we first make an initial seed/guess on the q_m values, and then use coordinate ascent to iteratively search for the q_m and S^*. Essentially, the S^* found using the initial q_m will be used to update q_m, and the updated q_m will in turn be used to update S^*. The process is repeated till convergence, and more details can be found in Appendix A.2.

2.3.3. Modifications for the background class

This section discusses the modifications we make available for Geo-RF based on the characteristics of crop mapping, which may or may not be necessary depending on specific application contexts. Specifically, the modifications aim to address potential issues introduced by the background class. For example, in soybean mapping, the vast majority of pixels will have the background non-soybean class. As a result, including the non-soybean class in Eq. (1) will substantially reduce the response from the log likelihood ratio to the error distribution of the soybean class (i.e., the main class of interest). In other words, if data points with the soybean class has a large number of errors, but the quantity is insignificant compared to the number of points with the non-soybean class, the log likelihood ratio – as a statistically aggregated score over all points – may not be able to capture any interesting spatial variability patterns about the model's ability to predict the soybean class. Thus, Geo-RF includes the flexibility of specifying the set of classes being included or excluded for the calculation of the log likelihood ratio.

2.3.4. Enhancing spatial contiguity of partitions

While most partitions are contiguous due to spatial autocorrelation, there often exist standing-alone locations or small irregularities along the boundaries of partitions, which tend to colocate with higher errors during test. We describe the phenomenon as "spatial overfitting", where certain locations are put to a different partition than most of their local neighbors due to noises or limited representativeness of the sample points for the other points at the location. To remove these non-contiguous small fragmentations, we add a contiguity-refinement module with local majority-voting to enhance the spatial smoothness of the partitioning. Specifically, for each location, the refinement algorithm looks at its local 3×3 neighborhood (i.e., cells in the grid), and assigns the majority partition among the cells (i.e., either S^* or non-S^*) as the partition of the location in the final partition map. By default, we repeat this contiguity-refinement step 3 times for the partitioning result. We also empirically analyze the effect of the number of repetitions in the experiments, which will be shown in Section 5.5.

2.4. Testing the impact of partitions on RF performance

An optimal bi-partitioning itself does not necessarily mean the functional relationships $\mathbf{X} \rightarrow \mathbf{y}$ are different in the two partitions, because an optimal S^* can be found in any data, with and without variability. Thus, the next important step is to evaluate the actual impact of the partitioning on the performance of the RF model. If significant

[1] The "log" here is commonly used – and necessary in most cases – to reduce the magnitude of the ratio to a tractable range without affecting the ranking (i.e., greater or smaller than) between different ratios.

improvements are observed after applying the bi-partitioning, we are more confident that there does exist discrepancy in $\mathbf{X} \rightarrow \mathbf{y}$ that was creating a bottleneck for the performance of the RF model. Otherwise, if the local RF models built for the partitions do not provide additional enhancements compared to the single RF model, that indicates the area is more likely to be homogeneous and no further partitioning is necessary. To evaluate this, we adopt the T-test to perform significance testing. In the case of Geo-RF, the T-test considers the following two scenarios: (1) A single RF model for the two partitions: This is the baseline scenario or the "control group" for the testing, where no treatment is applied to the RF model based on the bi-partitioning. In this case, the RF model remains the same as before. (2) Two separate RF models for the two partitions: This is the "experiment group" for the testing, where we train two different RF models based on the bi-partitioning, so each of the RF models can learn a different functional relationship $\mathbf{X} \rightarrow \mathbf{y}$ in case variability exists between the two partitions.

Denote the RF model for the baseline scenario as RF_{base}, and the two local RF models for the split scenario as RF_1 and RF_2, respectively. The model results of the two scenarios are obtained from the combined validation data points from the two partitions, which are from a subset of training data points as described in Section 2.3.1 and not test data points. For RF_{base}, we simply obtain the classifications from the model on the combined data points. For RF_1 and RF_2, we first separately obtain the classifications of RF_1 on partition-1 (e.g., S^*) and RF_2 on partition-2 (e.g., non-S^*), and then concatenate their results together. Since the data points involved in the comparison between the two scenarios are the same, we use the dependent T-test (a.k.a. paired T-test), which is specifically designed for this situation.[2] Furthermore, we are only interested in the case where the bi-partitioning can significantly improve the performance. In other words, it is insufficient for the two scenarios to be just significantly different as the case where the base model RF_{base} is better does not validate the bi-partitioning. Thus, we use the upper-tailed dependent T-test for the significance testing. Details of the test statistic and testing procedure are provided in Appendix A.3.

If local models built from the bi-partitioning pass the testing, then we accept the bi-partitioning and continue to run the bi-partitioning on the resulting child-partitions in the hierarchical process. The partitioning process terminates at a partition if the bi-partitioning does not pass the testing.

2.5. Local model building and selection

Each local RF model is trained using the data from its corresponding partition to allow different models to learn different $\mathbf{X} \rightarrow \mathbf{y}$ relationships. Unlike the deep learning version (Xie et al., 2021a) where shallow layer weights can be shared among different child-partitions that have the same parent, trees in RFs are learned independently with randomness and different trees do not know or depend on each other. Thus, for Geo-RF all local RF models are also kept independent to better align with the nature of the model.

Next, for the RF model selection, the general principle is quite intuitive: If a local model passes the significance testing, then that local model will replace its parent model and be responsible for classification tasks within its corresponding partition. Similarly, if its partition is further split through the hierarchical process, it may get replaced by local models trained for its child partitions.

With that said, there is an exception to this selection process. As in Geo-RF each RF model is trained independently using its own partition's data points, there is no direct "inherit-and-finetune" process that exists in the deep learning version, where weights from the parent deep learning model can be naturally used as pre-trained weights to be finetuned at a child partition. Thus, it is not necessary that a local RF model built for a child-partition will always lead to improvements over the RF model trained for its parent partition. For example, the local models RF_1 and RF_2 together may pass significance testing and have significant improvements over their parent model RF_{base}. However, it could be that all the improvements are from RF_1 whereas the performance by RF_2 might decrease compared to RF_{base} due to a smaller amount of data points available (i.e., there might be a trade-off between local patterns and the representativeness/amount of data points). Based on this, after each significance testing, we additionally evaluate if the improvements are observed for both local models or there is one model with reduced performance in its corresponding partition compared to RF_{base}. If it is the latter case, we keep using RF_{base} for that child-partition without the replacement. For example, if the models pass the testing, but RF_1 has an improvement and RF_2 has reduced performance, we move forward with RF_1 for partition-1 and RF_{base} for partition-2. In this case, RF_2 will be removed from the hierarchical process afterwards and all its roles will be replaced by RF_{base} (i.e., $RF_2 \leftarrow RF_{base}$).

3. Data

The dataset contains ~7.8 million samples in total from the CONUS area. As an overview, the inputs $\mathbf{X}$ to the model consist of: (1) 10 spectral bands from Sentinel-2 over 33 timestamps between January to November, 2021; and (2) 3 additional terrain attributes from NASADEM. The output $\mathbf{y}$ contains the crop labels from the CDL in 2021. The details of the input and output are provided in the following sections.

3.1. Sentinel-2 satellite imagery

The Sentinel-2 Multi-Spectral Instrument (MSI) mission includes visible bands, near-infrared bands, red-edge bands, and shortwave infrared bands, with high spatial resolutions up to 10 m. Following the workflow established by Li et al. (2023), we downloaded Sentinel 2 A and 2B Level-2 A Bottom of the Atmosphere reflectance images acquired between January to November in 2021, using the cloud cover threshold of 80%. Based on the Sentinel-2 scene classification map, we merged categories of cloud shadows, thin cirrus, snow, cloud with low, medium and high probability into cloudy pixels. In addition, we corrected the bidirectional reflectance distribution function (BRDF) effects using the c-factor method to derive nadir BRDF-adjusted reflectance (NBAR) images using the global spectral BRDF model parameters from Roy et al. (2017a,b). We then resampled the 20 m spectral bands into 10 m resolution using the nearest neighbor and created 10-day median composites, resulting in 33 compositing intervals. We also implemented the temporal linear interpolation to fill the data gaps following Griffiths et al. (2019). Furthermore, we divided the CONUS into 934 $1^\circ \times 1^\circ$ tiles in geographic projection with WGS84 datum with the spatial resolution of 0.0001° to approximately match a 10-m resolution pixel in Sentinel-2 and reprojected all Sentinel-2 images from Universal Transverse Mercator (UTM) system to the geographic projection. In total, 10 spectral band (i.e., blue, green, red, three red-edge bands, near-infrared, narrow near-infrared, and two shortwave infrared bands) over 33 compositing intervals were used.

3.2. Topographic data

We downloaded NASADEM covering the CONUS and derived elevation, slope and aspect as additional terrain variables for the Geo-RF training. NASADEM is a collection of Digital Elevation Model (DEM) and associated products derived from the Shuttle Radar Topography Mission (SRTM), incorporating with the Ice, Cloud, and Land Elevation Satellite (ICESat) and Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) (Buckley et al., 2020). Specifically, two

[2] As an example, when testing the effect of a treatment using the same pool of participants, a dependent T-test aims to compare the indicators before- and after-treatment on the participants.

datasets including the NASADEM_HGT for elevation, and the NASADEM_SC for slope and aspect were used. All the images were resampled from 1 arc second to 0.0001 ° spatial resolution.

3.3. Cropland data layer

CDL maps are crop-specific land cover products over the CONUS generated by the USDA's National Agricultural Statistics Service (NASS). Quality-wise, the CDL product has high accuracy for major crop types, and the accuracy in most areas is close to 95% according to the CDL metadata (USDA-NASS, 2023; Cai et al., 2018). Because of this, CDL has been a valuable dataset widely used in remote sensing research for algorithm testing and evaluation (Cai et al., 2018; Ma et al., 2024; Johnson and Mueller, 2021). In contrast, CDL does tend to have lower accuracy for other non-major crops. Thus, in this study we consider five major crops (i.e., corn, soybean, wheat, rice and cotton) and derived their labels from CDL 2021. We used a systematic sampling method with a spacing of 1 km to generate a CONUS sample, and extracted the crop cover information from CDL over the sample locations.

4. Results

4.1. Experiment settings

By default, we use a train-test split of 0.4–0.6, where 40% of samples are used for training and the rest 60% are used for testing. We also consider the following hyperparameters: (1) model complexity (by the maximum tree depth), (2) ensemble size (by the number of trees), (3) train-test splits, (4) the maximum spatial partitioning depth allowed, and (5) spatial contiguity. These parameter values will be varied (e.g., ~4-5 different values) and evaluated on all five crop classification tasks, resulting in a total of 110 sets of experiment results to offer a more comprehensive picture of the model performances under different settings. When varying one parameter, the default values for the other hyperparameters are: "unlimited" for the maximum tree depth, 100 for ensemble size, 0.4–0.6 for train-test split, 4 for the maximum spatial partitioning depth, and 3 (the number of repetitions for smoothing) for spatial contiguity.

The testing samples used for evaluation are fixed across the experiments using the above default train-test split setting (i.e., 60% of the data, which is about 4.68 million samples), except for the sensitivity analysis on different train-test splits in Section 5.3.1.

For the metrics we used the F1 score, which is the harmonic mean of precision (user accuracy) and recall (producer accuracy), i.e., $F1 = 2/(precision^{-1} + recall^{-1})$.

4.2. Learned geographic partitioning and improvements

Fig. 4 summarizes the geographical partitioning automatically generated from Geo-RF's partitioning process for soybean, corn, wheat, rice, and cotton, respectively. As explained in Section 2.3.2, in this work we use a grid with cells of size $0.5° \times 0.5°$ as the local groups, and the cells are the smallest units for partitioning the space. In addition, Fig. 4 also includes the counts of pixels (i.e., sampled from CDL) for the corresponding crop in each $0.5° \times 0.5°$ grid cell to show the geographic distribution of the crop as a visual reference. Finally, the right column shows the performance enhancements brought by Geo-RF's variability-awareness, where the values are the improvements of F1 scores on the corresponding crop compared to RF on test samples at each $0.5° \times 0.5°$ grid cell. In this column, the cells with little production of corresponding crops (i.e., cell values smaller than 100 in the distribution shown in the middle column) are masked out for the visualization to make it easier to see the meaningful patterns. Otherwise, it is very easy to see large improvements (or in some occasions reductions) that are caused by just a few samples, which are less interesting. For example, a $0.5° \times 0.5°$ cell with 10 true positive samples of corn for testing could show a significant improvement if only one more sample is correctly classified. Fig. 5 also provides the histograms of the F1 score improvements for the masked cells for each of the crops.

According to Fig. 4 (a1), there are five geographic partitions generated for corn, excluding the non-CONUS area P_0. The first major observation is that the major production areas of corn (partitions P_2 to P_5) are separated from the background low-production areas (P_1). Statistically, this aligns with our expectation as the likelihood of a crop-like pixel belonging to corn can differ significantly in major and non-major production areas. Geo-RF's decision to separate these two areas indicates that there are likely non-corn areas (e.g., other cereal crops) in the non-major production area of corn whose spectral characteristics are similar to those of corn in its major production area, making it challenging for RF to distinguish between them. As a result, this separation can help mitigate potential confusion between these pixels with similar spectral characteristics and improve classification performance. Some potential factors creating such cross-region variability can also be attributed to differences in terrains and climates (e.g., mountainous terrain and arid climate in the Western US, and the Southeast's subtropical to tropical climate), and different likelihoods of different crops that appear similar in satellite images (e.g., due to farmers' preferences or local policies). Additionally, we observe that partitions P_3 and P_5 align with the Corn Belt region, where climatic and environmental conditions such as warm nights, deep/fertile soil with high moisture-holding capacity, and flat terrain are ideal for corn production (Corn Belt, 2024). Comparing the two, P_5 covers more of the external areas of the Corn Belt, where the proportion of non-corn types of crops (e.g., soybean) starts increasing. As we can see in Fig. 4(c1), the partitioning leads to about 0.1 to 0.2 improvement in F1 scores for local regions in P_5. Finally, P_4 in Fig. 4 (a1) covers regions with less favorable conditions for corn, such as rough terrains, drought and cold weather. In these areas, the features of corn pixels could overlap with the features of non-corn pixels in areas with ideal conditions. Moreover, considering that the quantity of positive corn samples in P_4 is relatively small according to Fig. 4(b1), combining it with areas such as P_3 and P_5 could make corn pixels in P_4 less likely to be detected. Thus, this partitioning leads to a large improvement in F1 scores (~0.2 to 0.3) as shown in Fig. 4(c1).

Similar patterns of partitioning can be seen in the other types of crops, where major production areas are first separated out, and then finer-scale partitions within major production areas are further captured. There are several interesting observations for soybeans. According to Fig. 4(a2), the partition P_7 around downstream Mississippi River and western Mississippi state is separated out, which leads to a major improvement in F1 scores for about 0.10 to 0.25. As a more concrete example, for the 0.25 increase, the F1 score went from 0.57 to 0.82 at certain locations. Compared to the other major production areas (e.g., Midwest US), P_7 around Mississippi has more exposure to conditions such as drought or excessive rainfall, and its soil properties have different characteristics. As a result, models that work better for the other areas tend to not work so well for the region, and vice versa. In addition, we can see the Midwest area is merged with the non-major production area in partition P_1 (i.e., the partitioning process did not see significantly distinct behaviors during training and kept them as one). This indicates that the signature for soybean is potentially stronger (e.g., less confusion with other types of crops), making it relatively easier to distinguish it from non-soybean pixels.

For wheat, one interesting observation is the separation of partition P_3 in the northern part. The region is unique compared to the other areas as the climate conditions are both colder and drier and the wheat in this region consists of a mixture of winter wheat and spring wheat. These differences in P_3 mean that the spectral characteristics of wheat pixels are likely different from those in the other areas, including its adjacent neighbors which are more dominated by spring wheat. Performance-wise, P_3 does benefit significantly from the partitioning

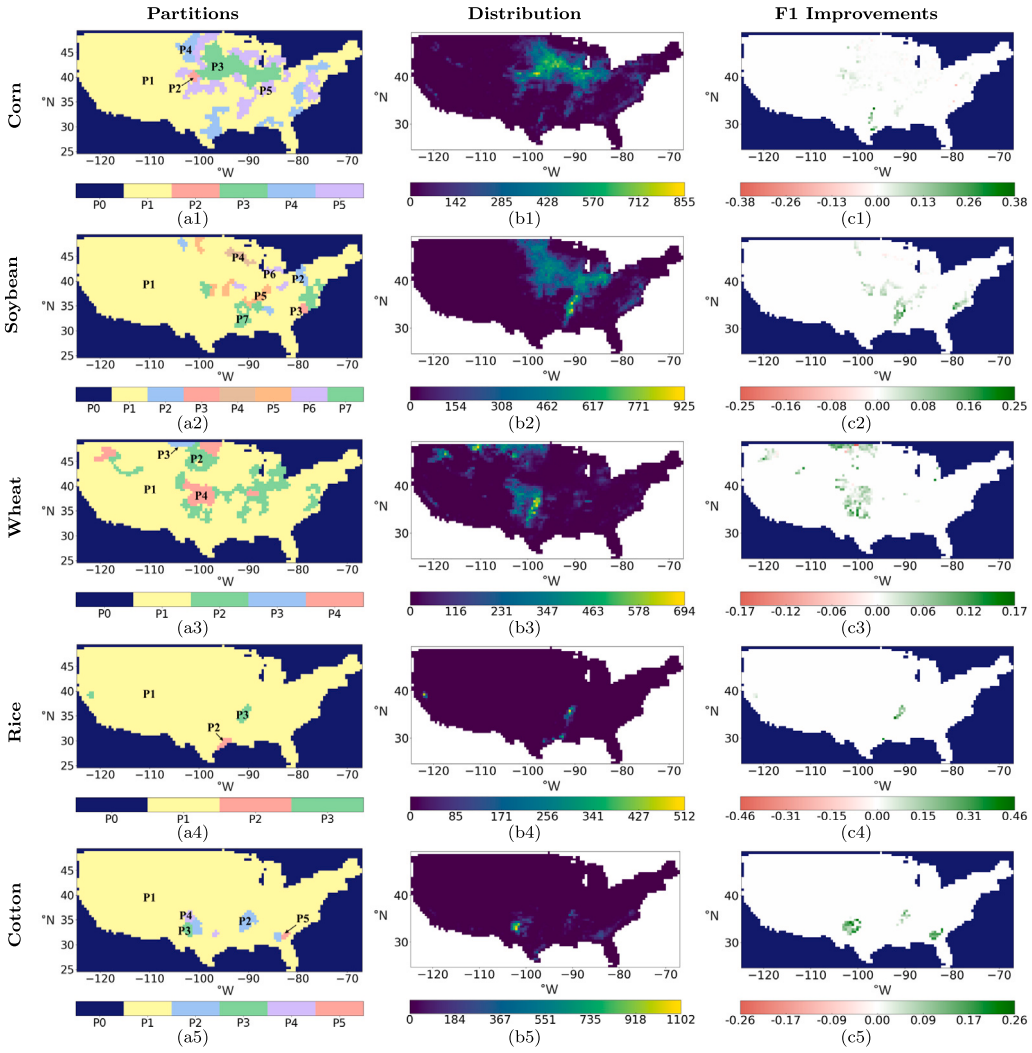


Fig. 4. Visualization of results. Column 1 shows automatically identified space partitioning for each crop, where the numbers represent partition IDs, and P_0 is the no-data area. Column 2 is the distribution of crop pixel counts, where the count value of each grid cell is the number of sampled CDL pixels corresponding to the crop in the $0.5° \times 0.5°$ cell. Column 3 shows the F1 score improvements by subtracting RF's F1 scores from Geo-RF's F1 scores across grid cells. Each row represents one of the five crops: (1) corn, (2) soybean, (3) wheat, (4) rice, and (5) cotton.

with its F1 score being increased by 0.1 to 0.15. In general, the separation also leads to major enhancements around the Kansas area.

For rice and cotton, we can see their production areas are relatively smaller compared to corn, soybean, and wheat. For rice, the first partitioning involves the separation of major rice-producing regions from the rest of CONUS (P_1). In subsequent partitioning, the rice region in the Gulf Coast (P_2) is further separated out. Compared to P_3, P_2's coastal subtropical climate with higher temperature and humidity levels can affect the physiological processes of plants, potentially influencing the spectral patterns. As shown in Fig. 4 (c4), the effectiveness of the partitioning is demonstrated by major F1 score improvements for about 0.30 in certain areas of P_2, and about 0.20–0.45 in many areas of P_3 (as examples, the F1 scores increased from 0.59 to 0.88, and from 0.33 to 0.80, at two grid cells, respectively). For cotton, interestingly Fig. 4 (a5) shows that the Northwest Texas cotton region is partitioned into 3 parts: P_2, P_3 and P_4, corresponding to three primary cotton regions in Texas: Rolling Plains, South Plains, and Panhandle, respectively (Texas A&M Agrilife Extension, 2024). The cross-region variability could be attributed to the differences in terrain, elevation, crop density, and relay cropping practices. Quantitatively, these partitions also result in major enhancements, demonstrated by a consistent increase of F1 scores of about 0.2, 0.2 and 0.25, respectively, for P_2, P_3, and P_4. Finally, P_5 covers an area in Georgia where soil properties differ from other regions by being more acidic, and management practices such as liming are commonly used to create optimal conditions for cotton growth. This region also benefited from the partitioning with a 0.18 improvement in F1 score as shown in Fig. 4 (c5).

Finally, it is worth-noting that the above descriptions about the partitions mainly aim to serve as potential interpretations of the decisions made by the data-driven partitioning optimization algorithm. In practice, the partitions may not necessarily align with the major or low production areas; regions with the same or different climate characteristics; etc. For example, if the machine learning model does not find difficulty in making correct predictions on samples from different climate regions (i.e., the $\mathbf{X} \rightarrow \mathbf{y}$ relationship can be approximated by a single model), then the regions will not be separated. Thus, enhancing the interpretability of the partitioning decisions could be a valuable future research direction to help better understand the causes of the partitions, which is analogous to interpreting the classification decisions made by the machine learning models.

4.3. Ability to capture controlled regions in synthetic data

This section aims to better validate the ability of Geo-RF to separate regions with different $\mathbf{X} \rightarrow \mathbf{y}$ relationships using semi-synthetic data

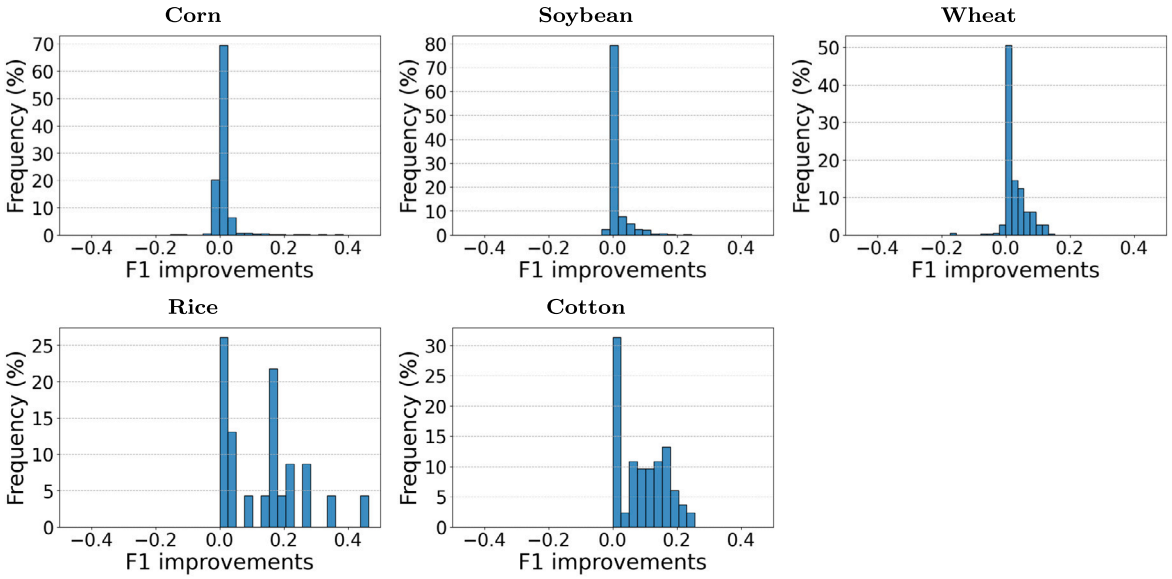


Fig. 5. Histograms showing the distribution of the F1 score improvements in the third column of Fig. 4 for the five types of crops. The frequency is represented by the proportion of cells in each bin. At this aggregated level over CONUS, the more significant improvements can be observed for wheat, rice and cotton.

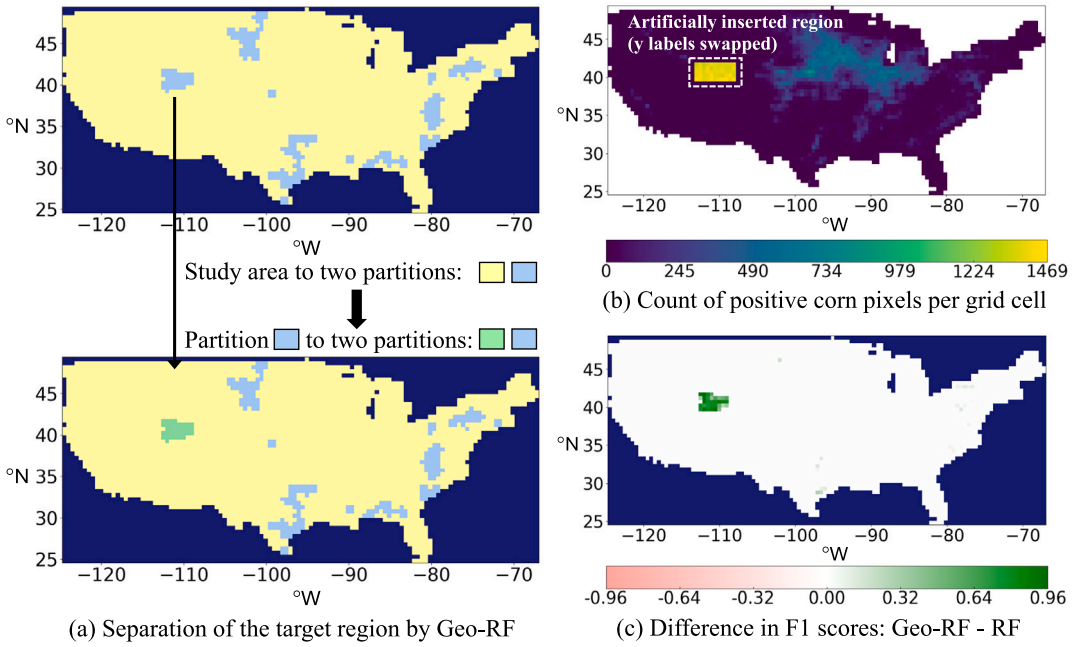


Fig. 6. (a) shows Geo-RF's automatic separation of the artificially-inserted region where the **X** → **y** relationship is made different by swapping **y** labels of points in the region while keeping **X** unchanged (i.e., changing "corn" to "non-corn", and vice versa). The two visualizations are the intermediate steps from Geo-RF's hierarchical partitioning process. We can see Geo-RF successfully separated the region from others in the bottom figure. (b) shows the distribution of positive corn pixels via counts aggregated over grid cells. This highlights the artificially-inserted region, which became the area with the most corn pixels. (c) shows the significant improvements of F1 scores (~0.75) with the variability-awareness offered by Geo-RF.

based off the original CONUS data. Specifically, we insert an artificial region into the corn classification task, where we enforce the points in the region to have a different **X** → **y** relationship than the outside points by swapping the **y** labels while keeping **X** unchanged (i.e., changing "corn" to "non-corn", and vice versa). In this way, features that previously correspond to non-corn in the region become indicators of corn. The goal is to see if Geo-RF is able to automatically capture this region during the partitioning process under this controlled setting. Fig. 6(b) shows the artificially-inserted region, where the previous dark area in Fig. 4(b1) now becomes an artificial major production area of corn. Fig. 6(a) shows the first two partitioning steps of a branch (i.e., Root → Partition "Blue" → Partition "Green") in the hierarchical partitioning process. We can see that Geo-RF successfully identified that the region has a different **X** → **y** relationship and separated it out at the second step. Fig. 6(c) shows the large F1 score improvements (~0.75) obtained with the variability-awareness of Geo-RF in the inserted region, compared with RF.

4.4. Summary of results with varying parameters

Figs. 7 to 11 show the results of Geo-RF and RF under different parameter settings, and the results correspond to each type of crop are shown in individual sub-figures. The X-axis shows the values of the

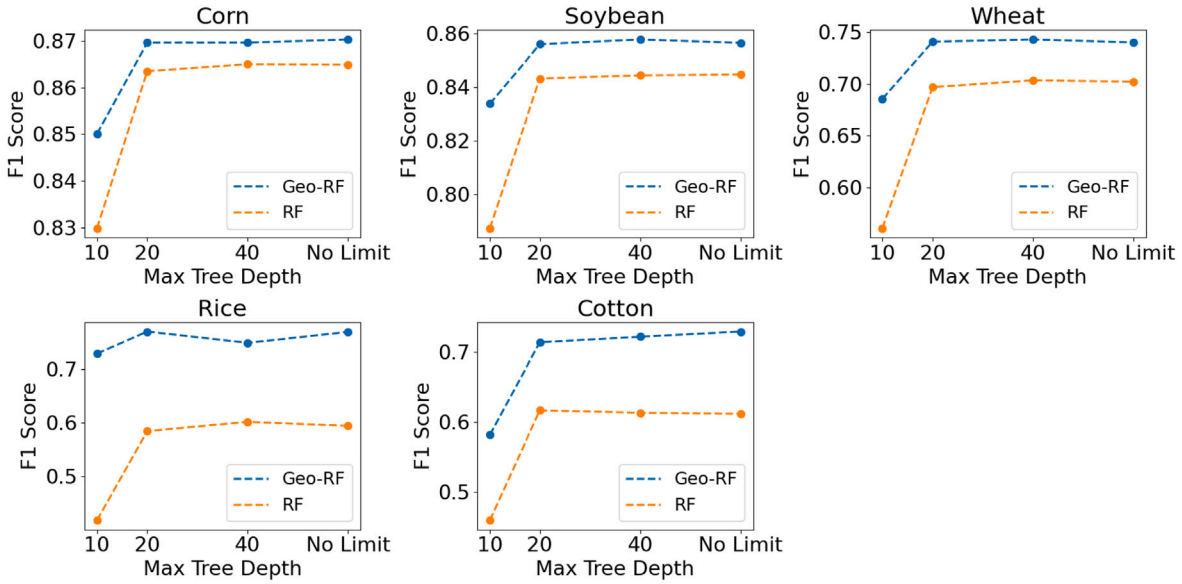


Fig. 7. Results by varying the maximum tree depth for different types of crops.

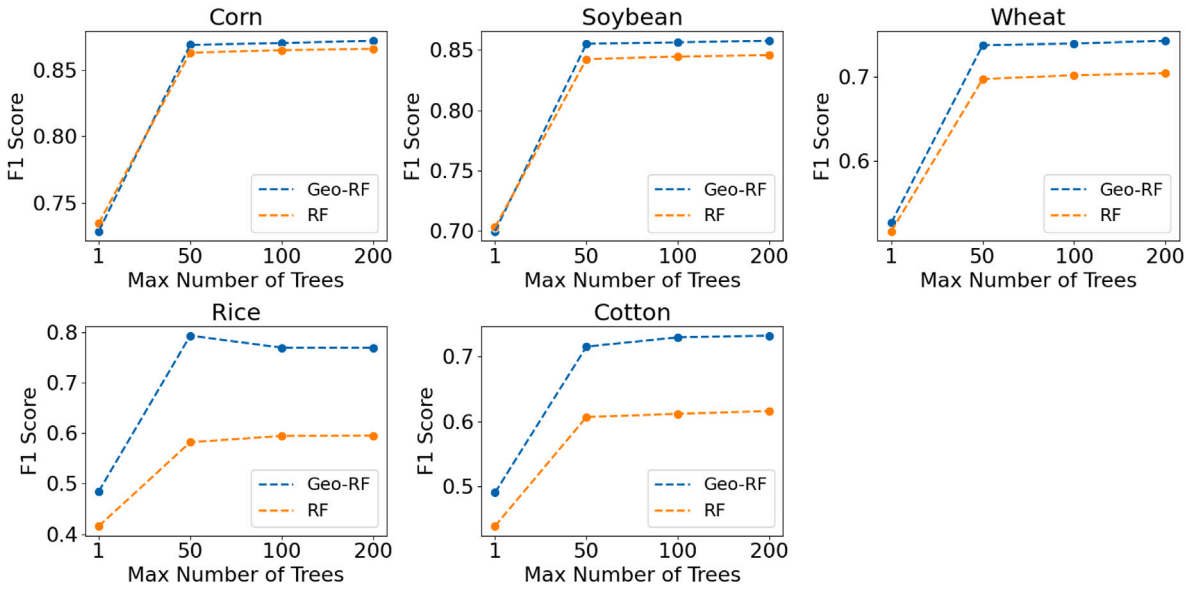


Fig. 8. Results by varying the maximum number of trees for different types of crops.

parameter being varied, and the Y-axis shows the F1-score. The detailed numbers are provided via the tables in the appendix. The general trend is that Geo-RF has higher F1 scores than those of RF in the vast majority of scenarios. The overall improvements are larger for rice and cotton, which have relatively smaller number of positive samples (i.e., samples belong to the crop type) compared to the other crops. The patterns are consistent across Figs. 7 to 11 with different varying parameters on tree depth, ensemble size, partition depth, train-test splits, and the number of regularization rounds. Here we mainly provide an overview of the results. In Section 5, we will discuss the effects of each parameter in details to better understand their choices in practice.

In addition, the results shown in the figures are the aggregated scores over all the test pixels in CONUS. As a result, while in some cases the overall improvements seem not as large (e.g., for corn and soybean in certain settings), the improvements in the local regions/states are significant as shown in the (c) column of Fig. 4. For example, as shown in Fig. 4 (c2), the F1 score of soybean classification near downstream Mississippi increased by about 0.10 to 0.25. The major improvements in local regions also align with the design of Geo-RF, which aims to address the drops in solution quality caused by differences across geographic regions. In other words, without variability-awareness, the model may have to compromise the solution quality in certain geographic regions when optimizing the others. It is worth-noting that ensuring high mapping quality across geographical regions is highly important in applications, and inaccuracies in certain areas could also lead to biases for downstream analysis and policy making. When looking at the aggregated numbers overall, these significant improvements in sub-regions (e.g., in the Mississippi area for soybean) are diluted. In the results, we visualize the improvements in Fig. 4 under the default setting and keep the aggregated numbers here to focus on the variations of model performances with different parameters. In Fig. 12, we also provide two examples of comparisons between Geo-RF and RF at local regions to better demonstrate the sub-region improvements. The sub-regions are highlighted by the boxes and the corresponding performance statistics on the test samples are presented in the bar charts. We can see that Geo-RF was able to significantly increase the

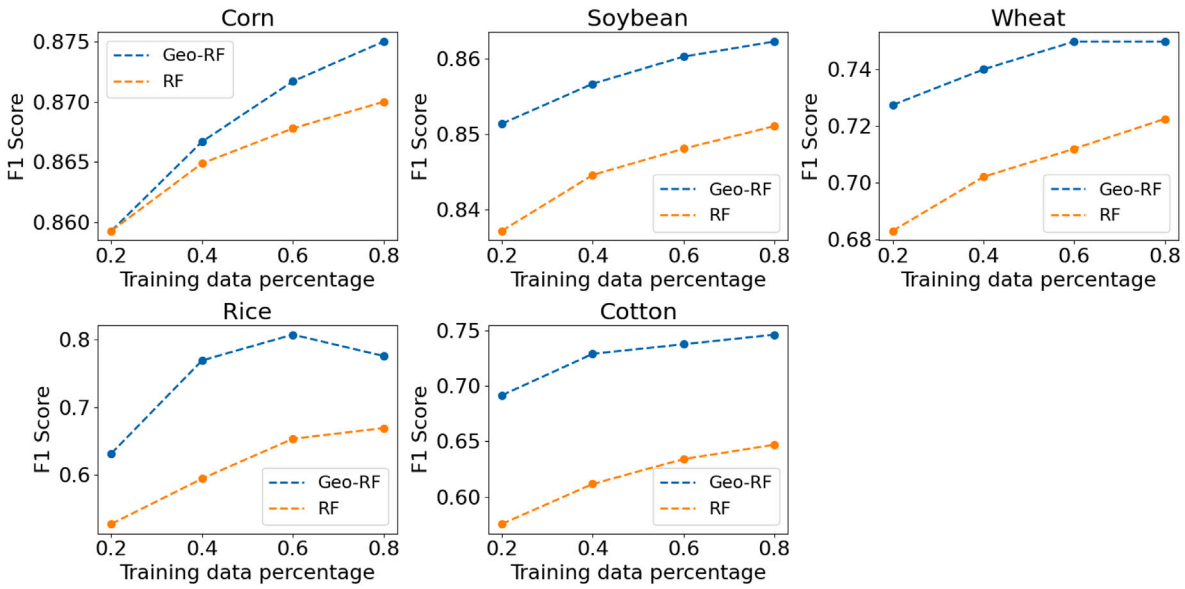


Fig. 9. Results by varying the train-test split for different types of crops. The amount of test data varies here and the percentage is one minus the training data percentage. This is the only set of experiments where the test data varies.

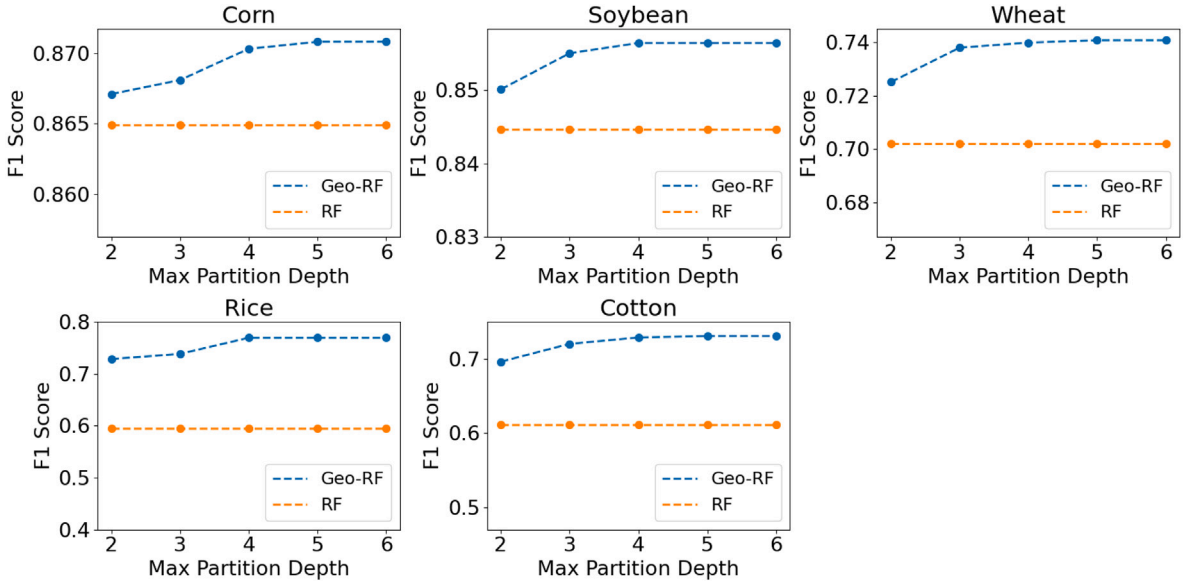


Fig. 10. Results by varying the maximum partitioning depth for different types of crops. This parameter is only applicable to Geo-RF so the results are fixed (flat) for RF in each sub-figure.

number of correct classifications of the crop samples and reduce the number of missed classifications. In particular, the F1 scores for the sub-regions increased from 0.70 to 0.81 for soybean and from 0.66 to 0.87 for rice, as shown in the bar charts.

5. Discussion

5.1. Effects of model complexity

In this sensitivity analysis, the model complexity is represented by the maximum tree depth allowed for each decision tree in the RF. According to Fig. 7, Geo-RF's most significant improvements over RF are observed in cotton and rice classification tasks. These two crops cover smaller areas but are distributed over broad geographic regions, making their overall performance statistics more affected by spatial variability. Specifically, the overall differences range from 0.175 to 0.31 for rice and 0.097 to 0.12 for cotton. Increasing the model complexity generally results in better performances for both Geo-RF and RF in all crop types. For example, the Geo-RF has a consistent increase of F1 score in corn (from 0.8501 to 0.8703), rice (from 0.7286 to 0.7691), and cotton (from 0.5818 to 0.7287). Thus, we recommend setting "No Limit" for the maximum tree depth. One interesting observation is that in general Geo-RF has a stronger ability to maintain a good performance with different tree depths (model complexity), and very often Geo-RF with a tree depth of 10 or 20 can reach similar or better performance compared to RF with an unlimited depth. In contrast, RF tends to have significantly reduced performance with smaller depth.

5.2. Effects of ensemble size

Fig. 8 shows the performance of Geo-RF and RF on different datasets with varying ensemble sizes, ranging from 1 to 200. We observe that

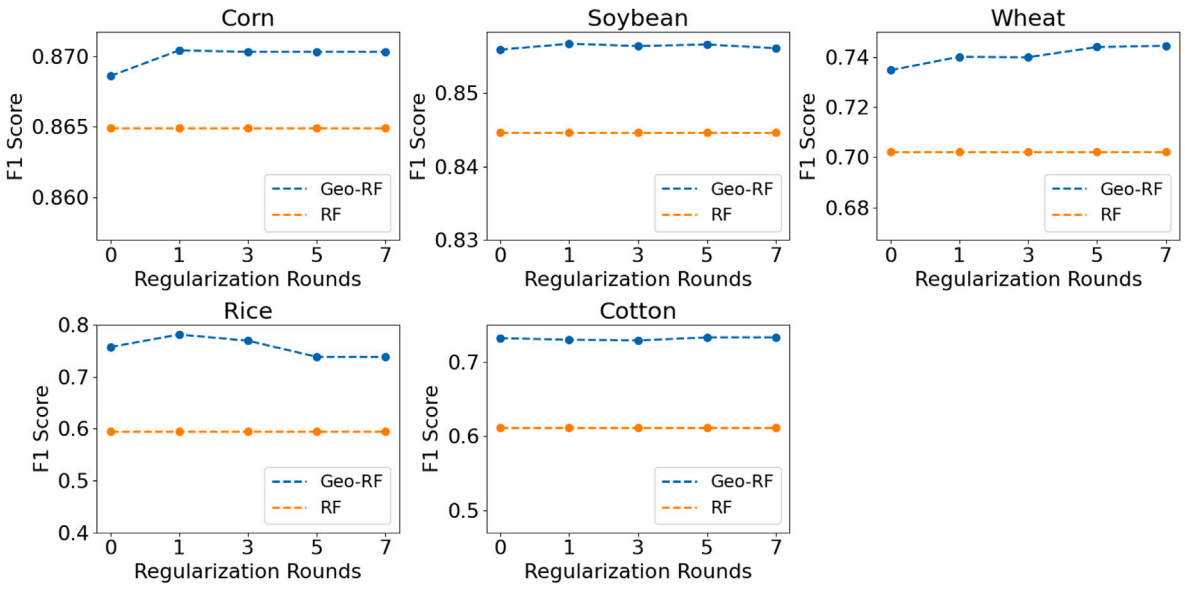
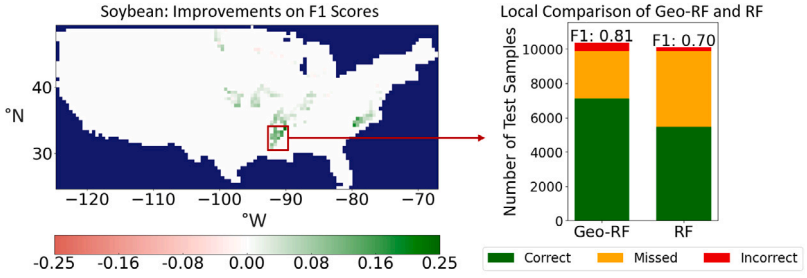
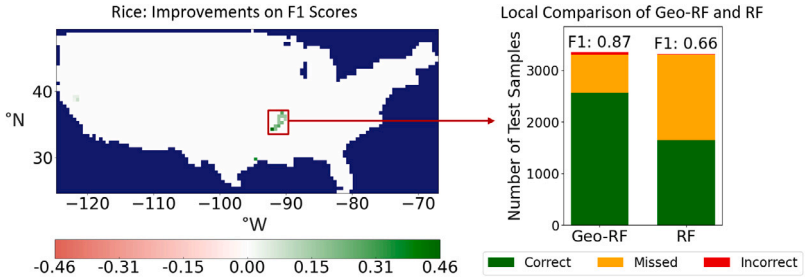


Fig. 11. Results by varying the number of regularization rounds in Geo-RF for different types of crops. This parameter is only applicable to Geo-RF so the results are fixed (flat) for RF in each sub-figure.



(a) Local performance statistics for soybean classification



(b) Local performance statistics for rice classification

Fig. 12. Examples of local performance statistics using (a) soybean and (b) rice. In each sub-figure, the left-side shows the sub-regions using boxes in maroon and the background is the F1 score improvements from the third column in Fig. 4. The bar charts show the corresponding performance statistics in the sub-regions. In the legend, "Correct" represents the number of test samples belonging to the crop that was correctly classified by the method, "Missed" represents the number of test samples belonging to the crop that the method failed to identify, and "Incorrect" means the number of test samples that the method classified as the crop but are not. The F1 scores calculated based on the statistics are also shown at the top of the bars. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

overall larger ensemble sizes result in better F1 scores and accuracy for both models in all crops. Larger ensemble sizes increase model diversity, reduce variance, and improve model robustness, allowing models to generalize better on test data. However, we can observe that the performance enhancement gradually becomes smaller as the number of trees increases, and the enhancements from 100 to 200 trees are in general very small. Thus, we recommend 100 trees considering the trade-off with computational time, and 100 is also the default setting in most libraries for RF. For all experiments, we observe that Geo-RF achieved a much better F1 score, and a higher accuracy, compared to RF, except on two occasions where the number of trees is 1 for corn and soybean. When only a single tree is created with bootstrapping, the variability captured is likely not representative and as a result reduces the performance slightly. However, this does not raise concern as in practice the ensemble size will be greater than 1 and here we only show this extreme setting to help understand the impact. When the number of trees is greater than 1, Geo-RF outperforms RF in all crops. Similar to the previous analysis on tree depth, the largest overall performance gaps between Geo-RF and RF are for rice and

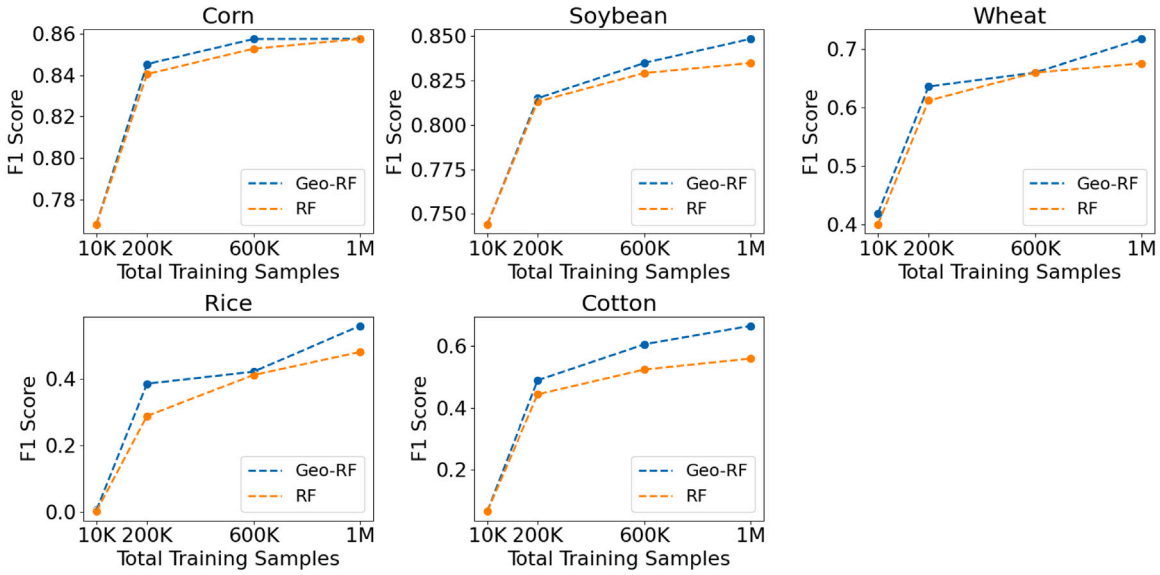


Fig. 13. Results for scenarios with limited numbers of training samples (K: thousand; M: million). The numbers of samples contain both training and testing samples, and the numbers of positive samples for the crops can be substantially smaller. For example, for cotton, the numbers of positive samples are 69; 1,419; 4,079; and 6,755 for the total numbers of samples at 10,000; 200,000; 600,000; and 1,000,000; respectively.

cotton, where the aggregated scores correlate more strongly with the improvements in local regions and states.

5.3. Effects of training data size

5.3.1. Varying train-test splits

We trained Geo-RF and RF on four training set ratios: 0.2, 0.4, 0.6, and 0.8, and measured the performance using the remaining data in each experiment.

Fig. 9 shows that as expected increasing the training set ratio generally leads to higher test F1 scores in both models. The improvements brought by a larger training set ratio were smaller for corn, soybean and wheat, potentially because their associated positive samples are larger and already representative at smaller ratios. In contrast, the effects on rice and cotton are much greater. We can see Geo-RF consistently achieved improvements over RF under different train-test ratios with the variability-awareness. The largest overall improvements were observed for rice and cotton, ranging from 0.0991 to 0.1748. This indicates Geo-RF's stronger ability in scenarios with limited observations.

5.3.2. Scenarios with limited samples

To better understand the model's performance in scenarios with limited observations, we further carried out the following evaluation cases with the following numbers of training samples: 10,000; 200,000; 600,000; and 1,000,000. The number 10,000 added at the beginning represents a very data-sparse scenario at the CONUS scale. It is also worth-noting that the numbers of samples listed above are the total numbers covering both positive and negative samples, and the numbers of positive samples (i.e., crop pixels) are substantially smaller. For example, for cotton, the numbers of positive samples are 69; 1,419; 4,079; and 6,755. For rice, the numbers are 25; 289; 821; and 1,357. At the CONUS scale, this means many states may only have a few or tens of examples of the crops, which are very difficult conditions for model training. For each of the other three crops that are relatively more produced, the number of positive samples ranges from a few hundred to several tens of thousands in total for training. In this evaluation, the test samples are fixed using the default 60% of data described in Section 4.1. The results are shown in Fig. 13. According to the figure, the general trends are similar, and Geo-RF shows improvements in most of the scenarios. This is important for Geo-RF, as having a very limited number of samples makes it more challenging for variability-aware learning, given the need to recognize and harness different functional relationships $\mathbf{X} \rightarrow \mathbf{y}$. From the results, we can see that Geo-RF is able to improve the performance or maintain the same level of performance compared with RF, when only very limited amounts of samples are available. In particular, Geo-RF continues to show significant improvements for rice and cotton. For example, for cotton, Geo-RF's performance with 600,000 samples (0.61) already surpassed RF's performance with 1 million samples (0.56) on the same set of the default 4.68 million testing samples.

In practice, if the number of samples is very limited but their functional relationships are highly different (i.e., the partitioning algorithm is able to statistically confirm it through significance testing with limited samples), then Geo-RF is expected to be able to separate the regions. However, if the number of samples is very limited and the differences between their functional relationships cannot be statistically confirmed, then technically it is not feasible for Geo-RF to separate the regions. This work focuses on the scenarios when training data is not a practical limitation, and future developments are needed for scenarios with very sparse training data. In addition, as a general context, the amount of training samples used for large-scale crop mapping (e.g., national or continental) is often on the larger-end of the numbers we considered here. While the labeling process via field surveys or local collaborations tends to require substantial effort, these activities are often carried out at scale to ensure data representativeness and the quality of the mapping products. For example, we used over 2 million samples for training the classification tree when creating the US national soybean map (Song et al., 2017), and used about 3 million training samples for mapping soybean expansion in South America (Song et al., 2021). Similarly, about 2.9 million samples were collected for crop mapping in three administrative regions in Ukraine (Gallego et al., 2014), and about 252,000 samples were collected in a smaller sub-region of it covering the Joint Experiment for Crop Assessment and Monitoring test sites as part of the GEOGLAM global crop monitoring initiative (Shelestov et al., 2017).

5.4. Effects of maximum allowed depth of partitioning

The maximum allowed depth of space-partitioning (i.e., from the root study area to the leaf partitions through the hierarchical process

of Geo-RF) controls the number of splits and thus the number of spatial partitions. This is independent of the maximum tree depth used for RF analyzed in Section 5.1. While a greater number of spatial partitions allows Geo-RF to have higher flexibility to capture heterogeneity with greater details, it is also more likely to lead to "spatial overfitting", meaning that the partitioning may be less likely to generalize to the test data. The maximum partitioning depth can be considered as an extra layer of control in addition to the significant testing introduced in Section 2.4. Fig. 10 shows that increasing maximum partition depths, in general, allows Geo-RF's F1 score to improve. This likely is the effect of significance testing, which helps prevent spatial overfitting by terminating the partitioning process if new partitions do not lead to statistically significant improvements of performance. Lastly, we observe that for all settings of maximum depth of partitioning, Geo-RF achieves higher F1 scores than the baseline RF for all crop types.

5.5. Effects of spatial contiguity regularization

As described in Section 2.3.4, the spatial regularizer aims to produce more spatially contiguous partitions by smoothing out local fragmentation. This is another measure used to reduce the chance of spatial overfitting in Geo-RF. To better understand the effects of spatial regularization, we train and analyze Geo-RF using different numbers of regularization rounds, ranging from 0 (no regularization) to 7. The results are shown in Fig. 11. Compared to the un-regularized version with "Regularization Rounds" = 0, we can see that Geo-RF's performance improves with the introduction of spatial regularization ("Regularization Rounds" = 1) in all crops other than cotton (from 0.732 to 0.730). For example, for rice, a single round of spatial regularization results in the best F1 score of 0.781, compared to the non-regularized model's score of 0.757. In general, we can see one round of spatial regularization is most stable and effective across different types of crops. It has the best performance in three out of the five crops, and in the other cases are within 0.005 of the optimal choice. While one-round can be a good choice, the other trend is that increasing the regularization rounds generally does not reduce performance by much (e.g., for corn the difference between 7 and 1 is only 0.0001), except for rice. The reason behind the different pattern for rice is likely due to the small size of its partitions, as shown in Fig. 4(a4). When the partitions are very small, smoothing could have larger impacts on the learned partitions and lead to major changes to the partitioning. For the other crops that have much larger partitions, more smoothing is overall beneficial.

5.6. Computational time

Here we compare the computational time of Geo-RF and RF. The expectations are: (1) Geo-RF intrinsically requires more time in the training phase, as Geo-RF uses RF as a base model before performing the partitioning and building local models. The additional training time for each of the new partitions should be smaller than the time needed for training the original model, as each model only considers a subset of the original data. (2) Once Geo-RF is trained, its execution time in inference should be similar to RF, as each sample will only go through one local RF based on its location.

Fig. 14 shows the execution time for training and testing for five types of crops with different amounts of training data. As the major differences will be for training, we considered different numbers of training samples: 200,000; 600,000; and 1,000,000. The testing set is kept the same as the default, i.e., 60% of the overall data with about 4.68 million samples. The experiments are carried out using AMD EPYC Processors with 32 cores. According to the results, the trends aligned well with the expectation: Most of the differences are observed during training, and the testing time costs are about the same for Geo-RF and RF. In terms of training, in general Geo-RF is about one to three times more expensive compared to RF. For example, the total training time for soybean with 1,000,000 samples is 230.6s for Geo-RF and 114.3s for RF, whereas the testing time on the testing data is 15.6s and 12.6s, respectively.

The computation for training can be further enhanced in future work for applications that are sensitive to training time. For example, a phased approach can be considered, where a subset of data can be used to generate the partitioning first to reduce training time spent on intermediate models for non-final partitions in the hierarchical partitioning process. Then the local models can be built based on the partitions using the complete dataset.

5.7. Broader applicability

Beyond crop mapping in CONUS, Geo-RF can be applied to different types of remote sensing data and use cases. Based on different application scenarios, certain settings may need to be considered and adjusted. For example, if the study area mostly consists of small and fragmented fields for crop mapping, a higher-degree of variability may exist at smaller scales (e.g., due to frequent changes of local environment conditions). In such scenarios, a finer grid would be more suitable for the partitioning to capture more fine-grained variability. If the study area consists of sub-regions with different degrees of variability, users may also use different grid cell sizes in different regions when defining unit groups. In addition, the settings for the spatial contiguity module such as the number of regularization rounds may also need to be adjusted to adapt to these regions. If the input study area itself is fragmented and has large empty gaps between the fields, users can specify the grid cells covered by the gaps as empty and the algorithm will automatically skip them during the partitioning. For convenience, the implementation of Geo-RF will be open-sourced on GitHub, and the URL is included in *Data Availability*. As discussed in Section 2.3.2, the main task users need to do as a preprocessing step is to define the unit groups for partitioning (e.g., using a grid) and the rest of the partitioning process, local model training, and prediction will be handled automatically.

5.8. Multi-class mapping

While Geo-RF can be directly used for multi-class classification, we used binary classification for each of the five types of crops to better control the factors that may otherwise affect the partitioning of space. For example, the number of samples from different types of crops could make the partitioning more responsive to crops with greater numbers of samples, while paying less attention to the other crops. This could make it more difficult to interpret the results and analyze the partitionings. Thus, in the experiments we chose the cleaner binary examples for better clarity. In practice, binary classification is also a common strategy used for large-scale crop mapping. For example, Li et al. (2023) created a maize and soybean map in China by combining two binary classifications (soy vs. non-soy, and maize vs. non-maize). Similarly, Potapov et al. (2022) used global Landsat data to map each thematic land cover and land use class and combined them into a multi-class global land cover map product. With that said, direct multi-class classification is an interesting and worth-exploring future direction and new designs can be developed to better separate the factors affecting the partitioning. For example, new constraints can be added to the partitioning optimization formulation to control the balance and representativeness over different classes.

6. Conclusions

This paper proposed a variability-aware Geo-RF that has the ability to automatically identify geographic regions with different functional relationships between input features (e.g., spectral band values) and target class labels to enhance the classification quality. A case study was carried out for major crop classification using over 7 million

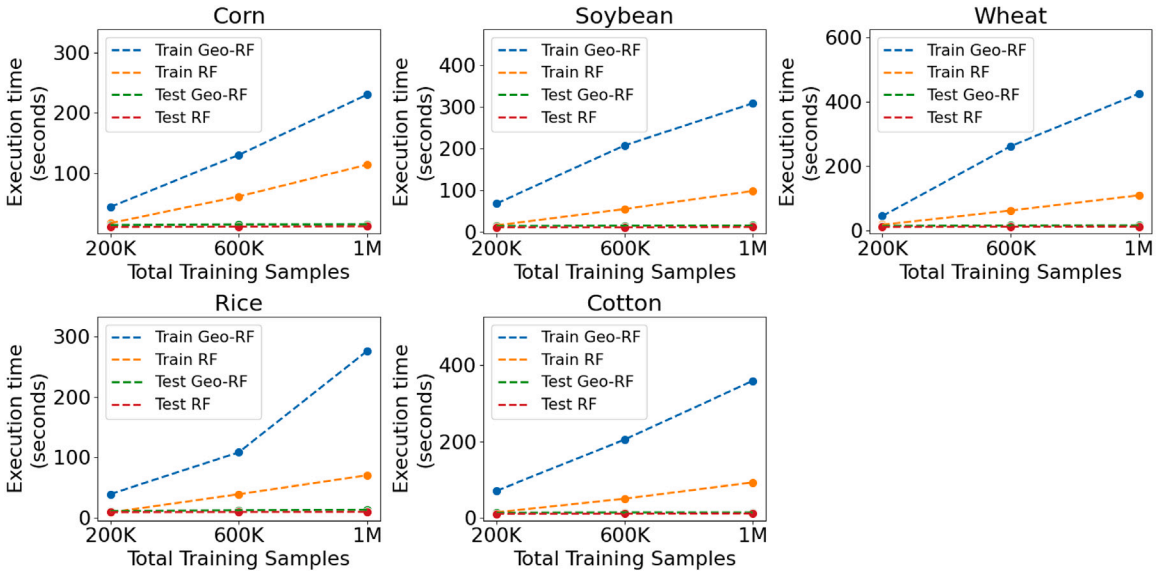


Fig. 14. Computational comparisons including both time for training and testing. The number of training samples is varied, and the test set is fixed.

samples from CONUS for five major crops, i.e., corn, soybean, wheat, rice, and cotton. Experiment analyses were carried out for the space-partitionings learned for different types of crops and the performances of the model under different conditions, as compared to RF without the variability-awareness. The results showed significant improvements of prediction quality in many partitions (e.g., states) where the physical conditions such as climate characteristics are different from the others. For example, the partition around downstream Mississippi for soybean classification improved the F1 scores in some local regions from 0.57 to 0.82, and the partition in Arkansas for rice classification led to F1 scores increasing from 0.59 to 0.88 at some locations. The sensitivity analyses also showed the effects of different hyperparameter choices, with recommendations including using unlimited tree depth. Geo-RF also showed stable performance compared with RF in scenarios with limited data. Finally, Geo-RF took about one to three times more time than RF during training, while the time for the testing phase remained nearly the same.

Future research directions will explore interpretable methods to understand the data-driven partitioning, new designs to further enhance the computational efficiency of the model for time-sensitive applications, new transfer learning or knowledge-guided extensions for scenarios with very sparse training data, new optimization formulation to better model partitioning over multiple classes, and comprehensive comparisons between different paradigms of learning models (e.g., tree-based models, deep learning and foundation models).

CRediT authorship contribution statement

Yiqun Xie: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Anh N. Nhu:** Writing – original draft, Visualization, Formal analysis. **Xiao-Peng Song:** Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Data curation. **Xiaowei Jia:** Writing – review & editing, Project administration, Methodology, Funding acquisition. **Sergii Skakun:** Writing – review & editing, Writing – original draft, Project administration, Investigation, Funding acquisition, Formal analysis. **Haijun Li:** Writing – original draft, Data curation. **Zhihao Wang:** Writing – review & editing, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This material is based upon work supported by the National Science Foundation, United States under Grant No. 2105133, 2126474, 2147195, 2425844, 2425845, and 2239175; NASA, United States under Grant No. 80NSSC22K1164, 80NSSC21K0314, and 80NSSC24K1061; Google's AI for Social Good Impact Scholars program; the DRI award and the Zaratan cluster at the University of Maryland, United States; and Pitt Momentum Funds award and CRC at the University of Pittsburgh, United States.

Supplementary Materials

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.rse.2024.114585.

Data availability

The implementation and evaluation of Geo-RF, including the code and the sampled dataset, are publicly available. The code will be available on GitHub at https://github.com/ai-spatial/STAR, including links to the datasets in machine-learning-ready formats. In addition, the source datasets used in this study are accessible online: (1) the Sentinel-2 data were downloaded from Google Cloud Platform (https://cloud.google.com/storage/docs/public-datasets/sentinel-2); (2) the NASA-DEM topographic data were acquired from the EarthData Search (https://search.earthdata.nasa.gov/); and (3) the CDL were obtained from the National Agricultural Statistics Service (NASS) of the US Department of Agriculture (USDA) (https://www.nass.usda.gov/Research_and_Science/Cropland/Release/index.php).

References

Atluri, G., Karpatne, A., Kumar, V., 2018. Spatio-temporal data mining: A survey of problems and methods. ACM Comput. Surv. 51 (4), 1–41.

Blickensdörfer, L., Schwieder, M., Pflugmacher, D., Nendel, C., Erasmi, S., Hostert, P., 2022. Mapping of crop types and crop sequences with combined time series of Sentinel-1, Sentinel-2 and Landsat 8 data for Germany. Remote Sens. Environ. 269, 112831.

Boryan, C., Yang, Z., Mueller, R., Craig, M., 2011. Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program. Geocarto Int. 26 (5), 341–358.

Brunsdon, C., Fotheringham, A.S., Charlton, M., 1999. Some notes on parametric significance tests for geographically weighted regression. J. Reg. Sci. 39 (3), 497–524.

Buckley, S., Agram, P., Belz, J., Crippen, R., Gurrola, E., Hensley, S., Kobrick, M., Lavalle, M., Martin, J., Neumann, M., et al., 2020. NASADEM: user guide. NASA J. PL: Pasadena, CA, USA 312.

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., Li, Z., 2018. A high-performance and in-season classification system of field-level crop types using time-series landsat data and a machine learning approach. Remote Sens. Environ. 210, 35–47.

CDL, 2024. Cropland data layer - USDA nass. https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php (Accessed: 05/09/2024).

Corn Belt, 2024. Corn belt. https://www.britannica.com/place/Corn-Belt.

d'Andrimont, R., Verhegghen, A., Lemoine, G., Kempeneers, P., Meroni, M., Van Der Velde, M., 2021. From parcel to continental scale–A first European crop type map based on Sentinel-1 and LUCAS Copernicus in-situ observations. Remote Sens. Environ. 266, 112708.

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Nicola, L., Rabaute, T., et al., 2019. Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. Remote Sens. Environ. 221, 551–568.

Fisette, T., Rollin, P., Aly, Z., Campbell, L., Daneshfar, B., Filyer, P., Smith, A., Davidson, A., Shang, J., Jarvis, I., 2013. AAFC annual crop inventory. In: 2013 Second International Conference on Agro-Geoinformatics (Agro-Geoinformatics). IEEE, pp. 270–274.

Fotheringham, A.S., Yang, W., Kang, W., 2017. Multiscale geographically weighted regression (MGWR). Ann. Am. Assoc. Geogr. 107 (6), 1247–1265.

Gallego, F.J., Kussul, N., Skakun, S., Kravchenko, O., Shelestov, A., Kussul, O., 2014. Efficiency assessment of using satellite data for crop area estimation in Ukraine. Int. J. Appl. Earth Obs. Geoinf. 29, 22–30.

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., Kalogirou, S., 2021. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int. 36 (2), 121–136.

Goodchild, M.F., Li, W., 2021. Replication across space and time must be weak in the social and environmental sciences. Proc. Natl. Acad. Sci. 118 (35).

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.

Grekousis, G., Feng, Z., Marakakis, I., Lu, Y., Wang, R., 2022. Ranking the importance of demographic, socioeconomic, and underlying health factors on US COVID-19 deaths: A geographical random forest approach. Heal. & Place 74, 102744.

Griffiths, P., Nendel, C., Hostert, P., 2019. Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. Remote Sens. Environ. 220, 135–151.

Grinsztajn, L., Oyallon, E., Varoquaux, G., 2022. Why do tree-based models still outperform deep learning on typical tabular data? Adv. Neural Inf. Process. Syst. 35, 507–520.

Gupta, J., Xie, Y., Shekhar, S., 2020. Towards spatial variability aware deep neural networks (SVANN): A summary of results. arXiv preprint arXiv:2011.08992.

Gupta, J., et al., 2021. Spatial variability aware deep neural networks (SVANN): A general approach. ACM Trans. Intell. Syst. Technol. (TIST).

Johnson, D.M., Mueller, R., 2021. Pre-and within-season crop type classification trained with archival land cover information. Remote Sens. Environ. 264, 112576.

Kulldorff, M., 1997. A spatial scan statistic. Comm. Statist. Theory Methods 26 (6), 1481–1496.

Kulldorff, M., et al., 2007. Multivariate scan statistics for disease surveillance. Stat. Med. 26 (8), 1824–1833.

Li, H., Song, X.-P., Hansen, M.C., Becker-Reshef, I., Adusei, B., Pickering, J., Wang, L., Wang, L., Lin, Z., Zalles, V., et al., 2023. Development of a 10-m resolution maize and soybean map over China: Matching satellite-based crop classification with sample-based area estimation. Remote Sens. Environ. 294, 113623.

Luo, Y., Yan, J., McClure, S.C., Li, F., 2022. Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model. Environ. Sci. Pollut. Res. 1–13.

Ma, Y., Chen, S., Ermon, S., Lobell, D.B., 2024. Transfer learning in environmental remote sensing. Remote Sens. Environ. 301, 113924.

Mpakairi, K.S., Dube, T., Sibanda, M., Mutanga, O., 2023. Fine-scale characterization of irrigated and rainfed croplands at national scale using multi-source data, random forest, and deep learning algorithms. ISPRS J. Photogramm. Remote Sens. 204, 117–130.

Neill, D.B., et al., 2013. Fast subset scan for multivariate event detection. Stat. Med. 32 (13), 2185–2208.

Phalke, A.R., Özdoğan, M., Thenkabail, P.S., Erickson, T., Gorelick, N., Yadav, K., Congalton, R.G., 2020. Mapping croplands of europe, middle east, russia, and central asia using landsat, random forest, and google earth engine. ISPRS J. Photogramm. Remote Sens. 167, 104–122.

Potapov, P., Hansen, M.C., Pickens, A., Hernandez-Serna, A., Tyukavina, A., Turubanova, S., Zalles, V., Li, X., Khan, A., Stolle, F., et al., 2022. The global 2000–2020 land cover and land use change dataset derived from the Landsat archive: first results. Front. Remote. Sens. 3, 856903.

Pott, L.P., Amado, T.J.C., Schwalbert, R.A., Corassa, G.M., Ciampitti, I.A., 2021. Satellite-based data fusion crop type classification and mapping in Rio Grande do Sul, Brazil. ISPRS J. Photogramm. Remote Sens. 176, 196–210.

Roy, D.P., Li, Z., Zhang, H.K., 2017a. Adjustment of Sentinel-2 multi-spectral instrument (MSI) Red-Edge band reflectance to Nadir BRDF adjusted reflectance (NBAR) and quantification of red-edge band BRDF effects. Remote Sens. 9 (12), 1325.

Roy, D.P., Li, J., Zhang, H.K., Yan, L., Huang, H., Li, Z., 2017b. Examination of Sentinel-2A multi-spectral instrument (MSI) reflectance anisotropy and the suitability of a general method to normalize MSI reflectance to nadir BRDF adjusted reflectance. Remote Sens. Environ. 199, 25–38.

Shelestov, A., Lavreniuk, M., Kussul, N., Novikov, A., Skakun, S., 2017. Exploring Google Earth Engine platform for big data processing: Classification of multi-temporal satellite imagery for crop mapping. Front. Earth Sci. 5, 232994.

Shwartz-Ziv, R., Armon, A., 2022. Tabular data: Deep learning is not all you need. Inf. Fusion 81, 84–90.

Song, X.-P., Hansen, M.C., Potapov, P., Adusei, B., Pickering, J., Adami, M., Lima, A., Zalles, V., Stehman, S.V., Di Bella, C.M., et al., 2021. Massive soybean expansion in South America since 2000 and implications for conservation. Nat. Sustain. 4 (9), 784–792.

Song, X.-P., Potapov, P.V., Krylov, A., King, L., Di Bella, C.M., Hudson, A., Khan, A., Adusei, B., Stehman, S.V., Hansen, M.C., 2017. National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey. Remote Sens. Environ. 190, 383–395.

Texas A&M Agrilife Extension, 2024. Cotton Production Regions of texas. Cotton Insect Management Guide. URL https://cottonbugs.tamu.edu/cotton-production-regions-of-texas/.

USDA-NASS, 2023. Cropland Data Layer Metadata. URL https://www.nass.usda.gov/Research_and_Science/Cropland/Release/ (Accessed: 2024-08-17).

Wang, S., Azzari, G., Lobell, D.B., 2019. Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. Remote Sens. Environ. 222, 303–317.

Xie, Y., He, E., Jia, X., Bao, H., Zhou, X., Ghosh, R., Ravirathinam, P., 2021a. A statistically-guided deep network transformation and moderation framework for data with spatial heterogeneity. In: 2021 IEEE International Conference on Data Mining. ICDM, IEEE, pp. 767–776.

Xie, Y., Jia, X., Chen, W., He, E., 2023. Heterogeneity-aware deep learning in space: Performance and fairness. In: Handbook of Geospatial Artificial Intelligence. CRC Press, pp. 151–176.

Xie, Y., Shekhar, S., Li, Y., 2021b. Statistically-robust clustering techniques for mapping spatial hotspots: A survey. ACM Comput. Surv.

Xuan, F., Dong, Y., Li, J., Li, X., Su, W., Huang, X., Huang, J., Xie, Z., Li, Z., Liu, H., et al., 2023. Mapping crop type in northeast China during 2013–2021 using automatic sampling and tile-based image classification. Int. J. Appl. Earth Obs. Geoinf. 117, 103178.

Zhang, C., Di, L., Lin, L., Li, H., Guo, L., Yang, Z., Eugene, G.Y., Di, Y., Yang, A., 2022. Towards automation of in-season crop type mapping using spatiotemporal crop information and remote sensing data. Agricult. Sys. 201, 103462.