Response to Second Round of Review Comments

2024-11-23

Dear Editor,

thank you for giving us the opportunity to resubmit our revised manuscript titled *"A Sentinel-2 Machine Learning Dataset for Tree Species Classification in Germany"* to ESSD. We genuinely appreciate the time and effort you and the reviewers have devoted for offering valuable feedback on our manuscript.

We are grateful to the reviewers for their insightful comments and suggestions, which have significantly improved the quality of our work. One of the major points raised by reviewer 2 concerned the method we used to assign labels in the training dataset, leading to an engaging and productive discussion about the pros and cons of various labeling strategies.

To address this, we have revised portions of the paper by introducing a new terminology that we define as "tree-centric" and "pixel-centric" labeling strategies. In the updated manuscript, these terms are introduced in Section 2.4. This section was partly rewritten and now contains a short literature review, demonstrating the lack of a "standard approach" and highlighting the variety of strategies that have been published for assigning tree species labels from forest inventory or management data. We further explore the advantages and disadvantages of the tree-centric pixel extraction approach for dataset usage in Section 4.5 ("Tree-centric pixel extraction") and discuss the implications and limitations ("Dos" and "Don'ts") for map production in the newly written Section 4.7, titled "Considerations for map production". Furthermore, the text has been reviewed by a native speaker familiar with remote sensing.

As the first freely accessible dataset derived from the German NFI for tree species classification, we are confident it will be well-received by researchers in the remote sensing community. Thank you once again for your consideration, and we look forward to your response.

Below, we have formulated our responses to the individual points raised by the reviewers. The updated manuscript, as well as a document highlighting the differences, have been uploaded.

On behalf of all authors,

Max Freudenberg

1 Reviewer 1

1.1 Minor comments

1. L26-L29 “Machine learning, particularly deep learning [. . .]” As I understand it, deep learning is a form of machine learning that uses many data layers and artificial neural networks in classification tasks, often applied in image recognition. It is also a buzzword. Please add references to this sentence to studies in which deep learning was used for tree species classification or similar. Good to shortly explain the difference between deep learning and machine learning here.

   Response: We added three citations and a short explanatory sentence regarding deep learning (line 27).

2. L77 “but due” indicates a decline in forest area from the 32% in 2012, but a decline in growing stock does not necessarily mean a decline in forest area, for example thinning. Please rephrase.

   Response: We split the sentence into two to separate the different meanings.

3. L109 “The growing space “approximately corresponds to the crown projection area”(Riedel et al., 2017, pp. 39, author’s translation), so we use these terms interchangeably in the following” No this should not be used interchangeably because it is very confusing to the reader. “approximately corresponds” is not sufficient to use the terms interchangeably. The “growing space” here is not defined, neither are its units. Is it cubic meter? This should really be changed, be careful with the use of terms and their units.

   Response: We added a sentence defining the growing space. Additionally, we clarified that the growing space and the crown area are not identical, and that we use the growing space as a proxy for the crown area. In the absence of a model for its direct estimation, it is the best approximation. The remainder of the text uses the term ”crown area”.

4. L113-L114 still very unclearly written, not sure how trees were selected as visible. Please rephrase.

   Response: We rephrased the passage again and hope that it is clearer now.

5. L196 “Obviously, broadleaf trees exhibit a much stronger seasonal pattern”, not if they are evergreen broadleaf trees. . . please separate leaf shape (broadleaf or needleleaf) and phenology (evergreen and deciduous) more clearly. For example, holly (Ilex) is an evergreen broadleaf tree/shrub in Europe that does not likely show an obvious seasonal pattern in reflectance. On the other hand, Larch (Larix) is a common deciduous coniferous needleleaf tree. Suggested edit: “Obviously, deciduous broadleaf trees exhibit a much stronger seasonal pattern than the evergreen coniferous trees in our dataset.”

   Response: Changed as suggested, sorry that this issue appeared again.

2 Reviewer 2

2.1 Major comments

1. The authors claim that the reason for assigning pixel spectra to single trees instead of tree composition parameters is the inaccuracy of tree locations derived from Bitterlich sampling. First, most/all large area studies that classify tree species suffer from the same inaccuracies. Yet, they train and validate their models the way I described. Second, if the field data lacks precision then the reference data should reflect that, i.e., a Sentinel-2 pixel corresponds to mixtures of trees and not individual trees.

   Response: The inaccuracies of tree locations do not arise from the Bitterlich sampling method itself. They mainly stem from imprecise GNSS measurements of the plot center, which is used as the reference point for tree locations. We believe this is sufficiently explained in Section 2.5. Many studies that use field reference data at the individual tree level face similar challenges. However, we would like to clarify that the practices for obtaining species labels are not as standardized as claimed. We have now included a brief literature overview of the different approaches (lines 127-131). The publications we reviewed all employed different methods for generating training labels from field data. Some focused on dominant species only, while others considered species shares. Some used forest inventory points, while others relied on polygons derived from forest management data, which have their own limitations. Additionally, some studies sample individual pixels within polygons, others use fixed-area plots, or calculate reflectances at the level of individual tree crowns, similar to our approach. Given these variations, our choice of extracting reflectances from individual trees is just one of many possible options.

2. The authors write I had suggested to train a classifier that predicts a certain tree species composition, whereas their approach yields probabilities of tree species occurrence. I must clarify that labeling pixels according to their tree species composition was merely one example how the authors could retain the complexity of the inventory data. Another way would be to assign to each pixel a tree list or estimates of species-specific basal area or cover (from their estimated crown area). Most importantly, predicting a discrete class or a probability estimate is besides the point. The training data I suggested can also be used to predict tree species probabilities. The argument is about the support size and labels of the reference observations and not the choice of estimators or prediction algorithm. Important for the argument is, that you claim to have produced a reflectance database for individual trees using up to 20x20 m pixel sizes, whereas the reflectances at that scale are a mix of trees, tree species, and background reflectance. In homogenous, single species stands, this may not matter as much, though I will later make the argument that it is still better to label pixels instead of trees in that case. Mixed species stands are often eliminated from training data because the inventory data is not precise enough to link species to pixels in those instances. Your dataset sill contains mixed species pixels but labels them according to single species. From the perspective of model training, this introduces unnecessary noise.

   Response: We argue that there are basically two approaches for linking field and satellite data, which we now introduce in section 2.4. as "tree-centric" and "pixel-centric" (lines 133-137). The tree-centric approach we chose aims to extract the most probable reflectance values for a given tree crown, while the pixel-centric approach attempts to label a set of pixels using metrics derived from field data, such as the ones suggested by the reviewer. We openly discuss the disadvantages of the tree-centric approach in section 4.5 and address the related geolocation errors in section 4.1. To provide just two examples of why the pixel-centric approach is not necessarily superior to the tree-centric approach: first, many studies work with pure stands based on an arbitrary definition of purity. Some classify stands as pure when the majority species has a share of more than 50% (Xi et al. (2021): `https://ieeexplore.ieee.org/document/9495140`), while others require more than 80% (Verhulst et al. (2024): `https://www.mdpi.com/2072-4292/16/14/2653`). Second, in the case of Bitterlich sampling, the support area is undefined and must be estimated, e.g. based on the tree diameters and the basal area factor (Blickensdörfer et al. (2024): `https://www.sciencedirect.com/science/article/pii/S0034425724000804`). We argue that is a priori unclear whether the errors introduced by these simplifications and assumptions are smaller than the errors occurring for the tree-centric approach. This is a research question that will certainly be addressed in the future.

3. The effect of oversampling field plots: This is a data publication and not a scientific article with scientific hypothesis. So as a reviewer, I try to consider the consequences for other users, i.e., can the dataset be used for the intended purpose and is the application and its limitations clear? If a dataset is incorrectly used, it can do more damage than good to the community. So, what is the purpose of this dataset? Training a foundation model? Producing maps? Both are fine but come with different requirements. When producing maps, we want to use the estimated model errors to infer map errors, which requires a probability sample. For land cover mapping studies, it is not as big of an issue to separate training datasets from validation datasets. Here, reference land cover data is relatively easy to come by. However, reference data for mapping tree species can only be obtained in-situ. In the case of the presented training dataset, any model errors estimated from boostrapping or cross-validation will be difficult to interpret because of the oversampling of the inventory plot. Although the NFI subplots follow a probability sample, the created training data focusing on individual trees is not. As such, such errors are not unbiased estimates of map accuracy (unlike those reported in previous studies). Now, it is possible to put the responsibility to the map users, but due to the lack of reference data, it is unlikely that users will follow good mapping principles or know how to. At the very least, the data publication should make a recommendation or be clear about what user's shouldn't do.

   Response: We believe that the community is well-equipped to handle diverse datasets and to advance to a new state of the art, should someone release a dataset better suited for the given task. To clarify the intended use case of the dataset and its limitations, we have added a sentence to the abstract and two sentences to the introduction (lines 67-69). Additionally, we have introduced a new discussion section, 4.7, titled 'Considerations for Map Producers' (lines 307 ff), that addresses the "Dos" and "Don'ts". In this section, we emphasize that users should be cautious when judging map accuracy based on model accuracy. Instead, we recommend using zonal statistics, such as comparing species distributions (of visible trees) at the state level or against published estimates of tree species abundance from the NFI. Finally, we suggest using auxiliary data for definitive validation.

4. It is reasonable to request that the authors provide sufficient information on how to use or not use this dataset, particularly as the dataset does not follow standard best practices. I am not suggesting, that the authors envision all potential use cases, but it would be good to understand and communicate the

limitations of the data. In this regard, I would ask the authors to add the information about the effect of added noise on model accuracies in the text so that it is citable.

Response: The newly introduced section 4.7 addresses the limitations of the data. However, we chose not to include a new passage describing the effects of added noise on model accuracies. Such a discussion would necessitate a detailed explanation of the methods used to assess this effect, which falls outside the scope of this work.

2.2 Minor comments

1. L135: What are these new possibilities?

   Response: We removed the text passage in this position and now list the possibilities in section 4.5 (line 280). The new possibility is mainly that it allows to extract data for rare species and such, that only appear in mixed stands, albeit their statistics will be influenced by mixed pixels.

2. Response to comment 15: To be correct, the geolocation error should be below half the size of a pixel. Since the error estimate specifies a range, +/- 9.5 m are OK for a pixel size of 20 m not 10 m. Also, the GNSS error of 5 m of the forest inventory data is surprisingly low. Can you provide more information how you obtain this estimate or a publication? Does this estimate only apply to the differentially corrected plots?

   Response: We added a text passage to section 4.1, that states why the geolocation error of 9.4 m is likely an overestimation. Regarding the GNSS errors: these errors were determined by us, as described in section 2.5 and 3.5. We analyzed the GNSS errors of the NFI measurements by matching tree positions to ortho images and Figure 13 shows the result (11.2m at 95% confidence / 5m at 81% confidence), which does not significantly differ between corrected and uncorrected measurements. Generally, the GNSS measurements are averaged positions of 100 measurements over 100 seconds. In addition, 76.5% of the measurements were corrected differentially using terrestrial reference stations. The dataset includes the correction status, so that users can filter by this property.

3 Reviewer 3

3.1 Major comments

1. The authors cover many important topics regarding reference data for large area mapping of tree species and supply a valuable dataset for the research community. While the dataset will lack behind expectation with regard to scientific freedom and flexibility, its provision will provide researchers, educators and students with additional data for the investigation of important research questions related to climate change, forest preservation, forest management and biodiversity studies. While some aspects of the data set, such as preprocessing, could be up for discussion, its publication will serve as a baseline for future publications of state and federal data sets and hopefully motivate more government authorities to provide their inventory data to the public. The writing style is excellent throughout the paper with only a single recommendation from my side. The researchers worked thoroughly on assuring high data quality and investigated the data set at hand for important characteristics, such as distinguishability of species from spectral signatures and geolocation accuracy, something that is to be expected from future related research. In my opinion, a few aspects of possible data usage were missed in the study design and discussion but overall, the state of the art in the field of tree species mapping with multi spectral satellite imagery is presented correctly.

   Response: We thank the reviewer for the positive comment. We now give a hint about possible data usages beyond machine learning in the introduction (line 68).

3.2 Minor comments

1. 106-107: “if their crown is overlapped by less than 50% by the surrounding trees” would be clearer language IMO

   Response: We revised this passage and included your wording.

2. 2.3: Additional TSA-processing withing the FORCE framework is not possible and the opportunity to create a dataset of even higher quality is missed. This would undoubtedly lower the amount of available tree observations but might help classification approaches that are sensitive to noise from cloud shadows and fog.

Response: TSA processing requires deciding on many individual processing parameters, which would have to be tailored to specific user needs. Furthermore, it requires high amounts of processing time and disk space. In consequence, we decided against it, especially as nowadays classification methods exist that can work with non-equitemporal time series (transformer-style neural networks for example).

3. 134-135: calculating the area-weighted average of a pixels might be a big source of noise if, let's say, the other 75% of that pixel depict a substantially different type of land cover than the target tree species. Think of the spectral signature of a deciduous tree that is added to a coniferous evergreen and the undergrowth signal in winter observations. In my opinion, some sort of outlier detection should be put in lace to detect possible addition of noise.

   Response: Due to the angle count sampling design, that does not measure all trees within a given area, we do not have complete information about a plot's coverage. In consequence, the described effect can occur and we discuss it in section 4.5. Computing the area-weighted average, however, reduces the noise, as it tries to be as precise as possible, even under uncertain conditions. We refrained from making further assumptions regarding which trees to include / exclude from the dataset, as we already filtered the trees based on their probable visibility. However, the end user of the dataset is free to implement further quality filters.

4. 276: This might be due to Pinus' often very top-heavy crown in plantations that allows undergrowth to be more visible. In combination with Betulas characteristic bark, it is no wonder that the signal gets mixed up. There might be similar issues with stands including Larix or Fraxinus. 310: One additional idea for use of your dataset could be the investigation of mixed pixels. For large area mapping it would be great to know if any given mixture of species within a single pixel can be learned by a classifier.

   Response: We added a sentence explaining that these species combinations were chosen because they are often co-occurring (line 229).

5. 310: One additional idea for use of your dataset could be the investigation of mixed pixels. For large area mapping it would be great to know if any given mixture of species within a single pixel can be learned by a classifier.

   Response: Unfortunately, this is not possible with the presented dataset, as it is impossible to derive a species share for a given area based on the included data. Section 4.5 (line 294 ff) lists the angle count sampling as underlying reason for this problem.

3.3 Other comments

While I see possible issues with area-weighted pixel extraction, I do not agree with the criticism stated by Reviewer 2: FORCE uses the ImproPhe algorithm (Frantz 2016) that alters the pixel values of the 20m bands. To my understanding, duplicate values within the vicinity of a datapoint and thus spatial autocorrelation will be quite unlikely given the large size of the dataset. I can also support the author's claim, that duplicates (as well as random noise) within a certain threshold as stated in regard to the LAION 400M dataset are no issues for modern machine learning algorithms, especially neural networks, from my personal experience. The addition of random noise is a valid point of criticism. However, as long as European NFI rely on fixed position sample plots, this approach seems to be the only viable method to provide data to the research community.