The authors would like to thank all reviewers for their time and the high-quality replies including thoughtful wording. We apologize for the delay regarding our reply and appreciate the patience on the journal side.

# Changelog

This is only a high-level changelog. Detailed replies can be found in the point-by-point replies.

* Added method spcv\_knndm to the package and manuscript
* Aligned all figures with respect to font and point sizes
* Removed CLUTO method from package and paper (as the method is not available anymore)
* Improved reproducibility script (instructions for installation of python based dependencies for plotly() calls)
* Literature research for uses of knndm method in publications
* Package: Added the option to show omitted points for spcv\_buffer
* Added Figure 1 showing an mlr3 ecosystem overview
* Sanitize capitalization in references
* Increased bundled R package version from v2.0.2 to v2.3.0
* Extended and renamed section “Resampling in hyperparameter tuning” to “Resampling in model optimization”

Reviewer 1

*> 1) I think it should be mentioned that the described strategies only become relevant when no probability sample of the area of investigation is available. This is the case for most studies but considering recent discussions (especially Wadoux et al., 2021) this should be clarified to avoid further misunderstandings. I think a sentence explaining that a probability sample and design based inference is the only way to assess the map accuracy would be required, and to explain that spatial CV can only help to approach it.*

Thanks for the suggestion. We are aware of Wadoux et al., 2021 and its approach.

As you outline correctly, the methods of Wadoux et al., 2021 are only applicable when a probability sample is available and the additional data can be set aside for model testing. However, this is typically not the case considering the cost of additional data acquisition.

To make this more clear, we have included the following sentence

“We would like to note that design-based approaches are also a viable approach in situations in which it is possible to obtain additional independently sampled test data as a probability sample [@wadoux2021].”

*> 2) The paper misses a very recent contribution https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13851 that is also implemented in the mentioned CAST package and presents a novel spatial leave-one-out resampling strategy. I understand that this strategy is not yet considered in mlr3spatiotempcv since it got published just very recently but I wonder whether it might be feasible to still account for this contribution.*

*I think that this would be a good addition especially relevant because the authors conclude that to choose a CV strategy, the strategy needs to “mimic the predictive situation in which the model will be applied operat*ionally”. This is explicitly the aim of this new method and not accounted for by the others.

Thanks for the suggestion to include a new method.

Indeed, this new method got published shortly after we handed in this article.

Meanwhile, the method has been implemented as a k-fold variant (CAST::knndm()).

We have added the method to the package in version 2.2.0 and added a section to the article following the same structure as for the other methods.

The overall idea is to add new methods to the package as they become available, while the article will always be a snapshot of available methods at the time of writing.

*> 3) “It is worth noting that a spatial cross-validation library named spacv has recently been developed for Python…” (page 3). This sentence sounds not well linked. I think something like “Parallel efforts are currently also made for Python, but for R such a library is still missing” would make it ore clear. Also, I think similar effort is currently done for tidymodels but this not mentioned here at all. I think for the reader it would be beneficial to point on the similar opportunities in other ecosystems as well and the fact that this is currently developed for other ecosystems does not make mlr3spatiotempcv redundant or irrelevant.*

Thanks for the suggestion to improve the mentioning of similar efforts.

We welcome this idea and always want to list all possible efforts, which can be seen by the existing list of other resources in the package README file.

We are a bit confused by your sentence "but for R such a library is still missing" as in this case it is rather Python and not R for which such libraries are missing.

Maybe you meant to say Python instead of R here?

We have updated the paragraph and mentioned spatialsample from the tidymodels framework, which has likely been heavily inspired by mlr3spatiotempcv, but never officially linked/mentioned mlr3spatiotempcv in any way.

Besides, we also included the Python package spatial-kfold in the list of other resources.

*> 4) “To reduce dependencies, some methods were re-implemented instead of importing them from the respective upstream packages” (page 6). This is relevant to know and the information should be added e.g. in Table 1.*

We have added a `\*` to all methods in Table1 which have been re-implemented instead of relying on their direct upstream functions.

*> 5) “Leave-location-and-time-out” (page 20): it should be mentioned here that this method was intended as a simple intuitive solution to handle spatially distributed sensor locations as training data for machine learning models where dependencies in space and time exist. LLO, LTO and LLTO should also clearly regarded as "custom\_cv" because they depend on user-defined space/time dimensions, e.g. catchments, sensor locations, even spatial blocks.*

LLO, LTO and LLTO are already labeled as subtype “custom” in Table 1. The custom\_cv mentioned is a dedicated method name from the mlr3 package and should not be confused with the subtype definition in Table 1 - see also its “Name” and “Sub-type” entries in Table 1.

*> 6) Fig. 2: I don’t understand why this is referred to as an example for presence/absence data since we don’t see presence/absence labels here and it’s not clear why this only applies to such data.*

Presence/absence data is considered the "default" spatial modeling case here, i.e. for users not familiar with species distribution modeling, and is only explicitly mentioned as the \pkg{blockCV} package - and its spatial\_block() function - target the species distribution community. This community sometimes works with presence/background data.

In the case of presence/absence the data does not have to be explicitly labeled.

The information that all data consists of presence/absence observations is enough to know that all binary values of the response variable are verified observations - which is not the case for presence/background data.

*> 7) The manuscript could better explain that the spatial leave-one-out strategies naturally lead to the exclusion of certain reference points in each iteration, hence some points are used neither for training nor for testing.*

We have added the following sentences to the introduction of the spatial leave-one-out section:

“This group of methods discards observations when applying buffers, meaning that these are not used for training or testing.

Depending on the fraction of data being discarded, this might result in a substantial reduction of the sample size and consequently in pessimistically biased performance estimates.”

*> 8) “Leave-one-disc-out resampling becomes spatial LOO CV for a radius of 0 m and when each observation is at a unique location.” (page 10) If I understand in correctly, it becomes a LOO CV without the need to call it a “spatial LOO CV” ?*

This is correct. We have removed the word “spatial” and reference to a radius of 0 m.

*> 9) When using single spatial clusters/blocks as held back data it is important to know how mlr3 handles the summary statistics. Are they calculated per fold and then averaged? Or calculated over all held-back data. Especially for R² this might make a huge difference since in many cases (looking e.g. at fig. 3) we would assume that the within-spatial-cluster variability is rather low (likely leading to small R² values)?*

mlr3 averages the statistics per fold and the aggregates using the mean by default (this is configurable).

To our knowledge, this is the default way of aggregating summary statistics in a nested CV setting, which is why we did not mention it explicitly (see e.g. Eq. (5.3) in James et al. 2021, Introduction to Statistical Learning with Applications in R, second edition, Springer). Yet to highlight this more after the explicit hint, we added the following hint to Figure 2:

“In \pkg{mlr3}, statistics are then calculated per fold and aggregated using the mean by default.”

*> 10) Page 25: Random CV is stated as over-pessimistic and I totally agree that this is likely the case here. The authors, however, also state that “Spatial CV, in contrast, provides a better/more accurate measure of a model’s ability to generalize from the training sample”. I also agree in most cases but find this statement a bit too general especially since it cannot be proven here. The paper misses to make aware that the chosen spatial CV approach might be over-pessimistic because it leaves huge parts of the study area out. Hence, the suitability also depends on the area to which the model is later applied and I think this could be better highlighted, also to account for the criticism in Wadoux et al (2021).*

Thanks for the suggestion to be more precise with the wording here.

We agree that the statement is lacking the possible disadvantages of using spatial CV, and CV in general. We would like to point out that this paragraph is about the interpretation of empirical results, which why we would like to refrain from referencing alternative approaches that were not considered in this limited case study.

We have updated the paragraph and added the following:

"Yet, it must be kept in mind that using spatial CV might also lead to a pessimistic estimate of the model's predictive performance if large parts of the study area are being left out during training, depending on the number of CV folds being used."

*> 11) Section on “ Resampling for hyperparameter tuning” (page 27) is highly relevant but this is only a selected part of the model tuning process. Why not mentioning the entire tuning procedure which also involves e.g. variable selection ?*

Thank you for this comment, which points to the fact that ‘tuning’ in machine learning can be defined very broadly, including feature selection.

Feature selection can, e.g., be conducted using wrapper methods or filter methods, and specific settings (such as filter thresholds) can themselves be treated as hyperparameters.

We have extended/rephrased the referenced section with more information related to feature selection and also added more references to aid the existing statements. Also we changed the title to “Resampling in model optimization” to make it more general because of the inclusion of wrapper-and filter based variable selection references. To keep the focus of the manuscript on resampling methods we decided to stay brief on the topic and refer to the linked references.

*> 12) “This contribution reviews the state-of-the-art in spatial and spatiotemporal cross-validation”.*

*Strictly speaking it doesn’t because a review of the state of the art would, in my opinion, also include an evaluation or discussion for specific cases.*

Thank you for this comment. We re-worded this to better reflect the intention of this manuscript:

“This contribution provides an overview of the state-of-the-art in spatial and spatiotemporal cross-validation techniques and their implementations in R.”

*> 13) “The results from this performance assessment would be over-optimistic, and perhaps badly so” (page 2): I think “badly” is not making it sufficiently clear. Suggestion: “and won’t reflect the ability of the model to make predictions for new fields”*

Thank you for the specific suggestion.

We have adapted it a bit and added the following:

"and will not reflect the model's true ability to make predictions for new fields."

*> 14) “The count in brackets after the package name represents the number of studies that were found having used this resampling technique until May 2021. For each method, up to three randomly selected references were added to the table.” It sounds odd to make a random selection here. Can’t they be selected e.g. according to scientific impact?*

We appreciate the suggestion. This is an issue that we did indeed discuss among the authors. We chose this admittedly unusual approach due to the increasing criticism toward conventional bibliometric impact assessments and the impracticality of reviewing all relevant publications.

More recent articles would also have lower citation counts.

We therefore decided that a random selection would be a fair choice, while being aware of the drawbacks of this approach.

*> 15) Very minor comment: “well-known case study sampling techniques for…” (page 22). I doubt that the case study itself is well known. I recommend to remove the “well-known”.*

We agree and removed it.