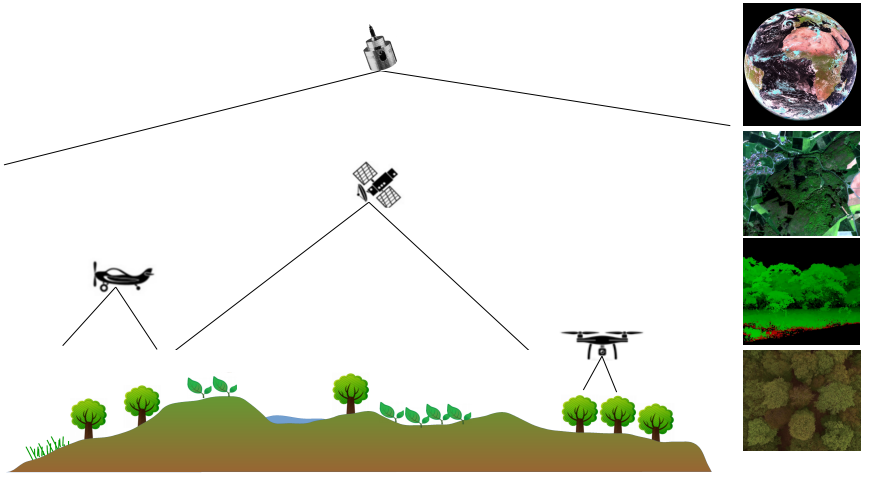
Machine learning for remote sensing applications

Hanna Meyer
Institute for Geoinformatics, WWU Münster

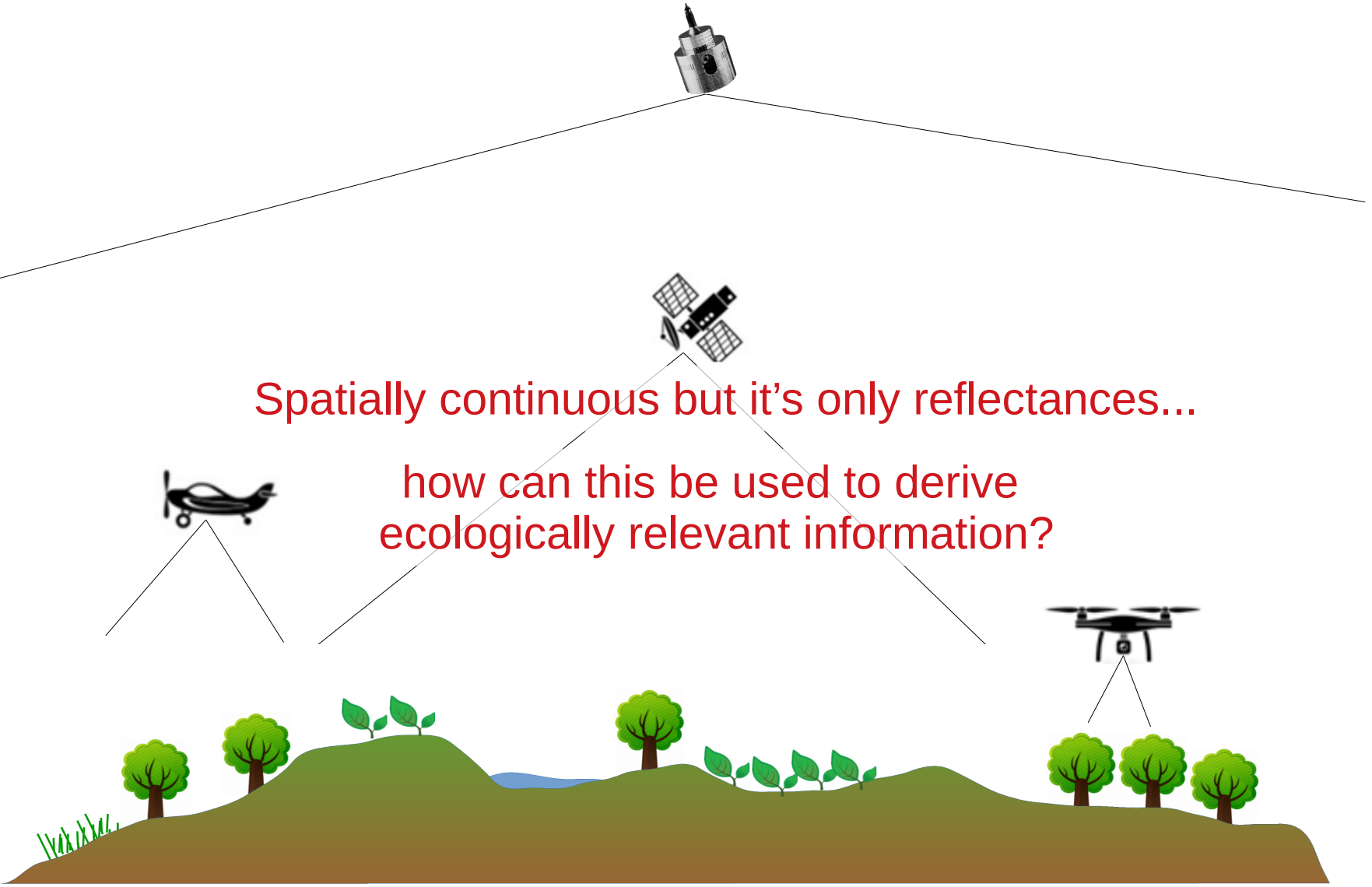


→ Slides and material: https://github.com/HannaMeyer/OpenGeoHub_2019

Problem: From field observations to maps of ecosystem variables

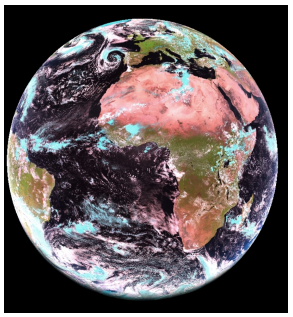
Nature 4.0 | Sensing Biodiversity

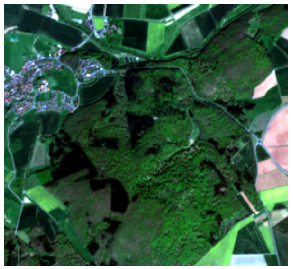
Remote Sensing of landscapes

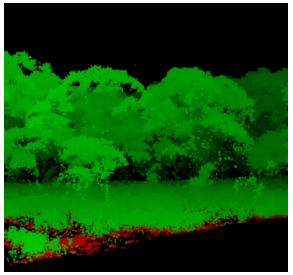


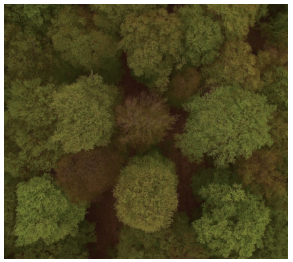
Spatially continuous but it’s only reflectances...

how can this be used to derive ecologically relevant information?







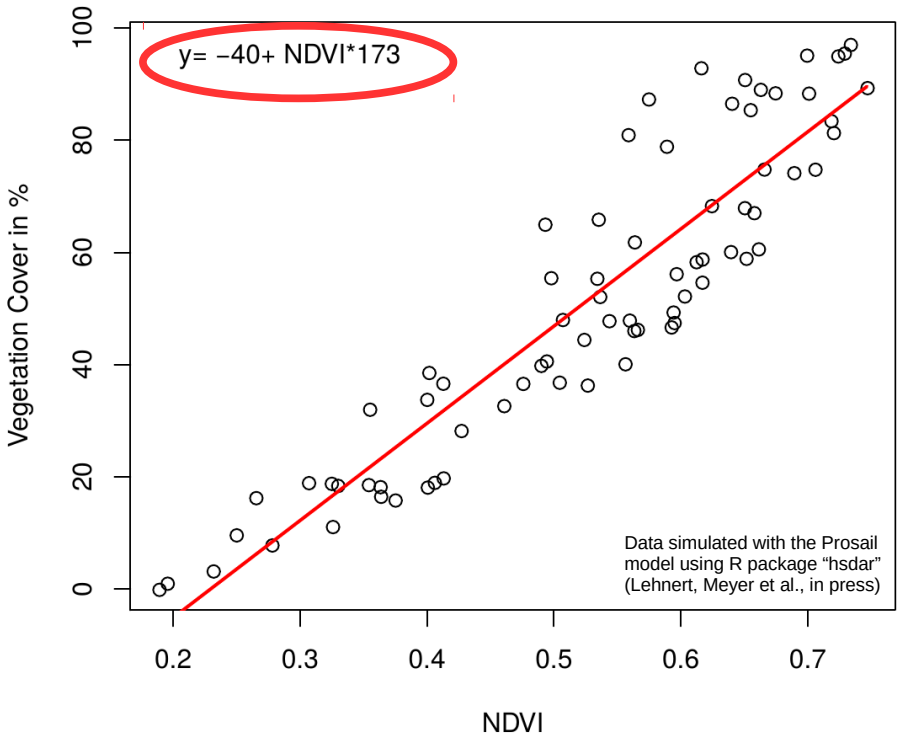


Direct and indirect sensing of the environment

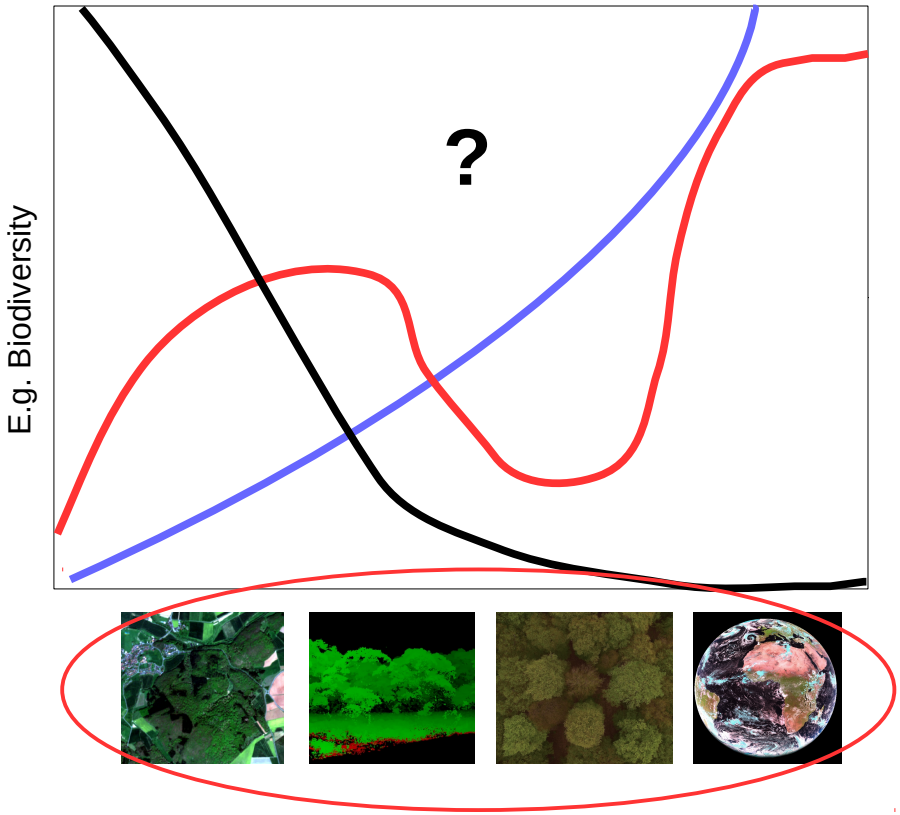
e.g. vegetation cover from satellite (VIS/NIR)

Typical ecological variables from satellite?

Field data

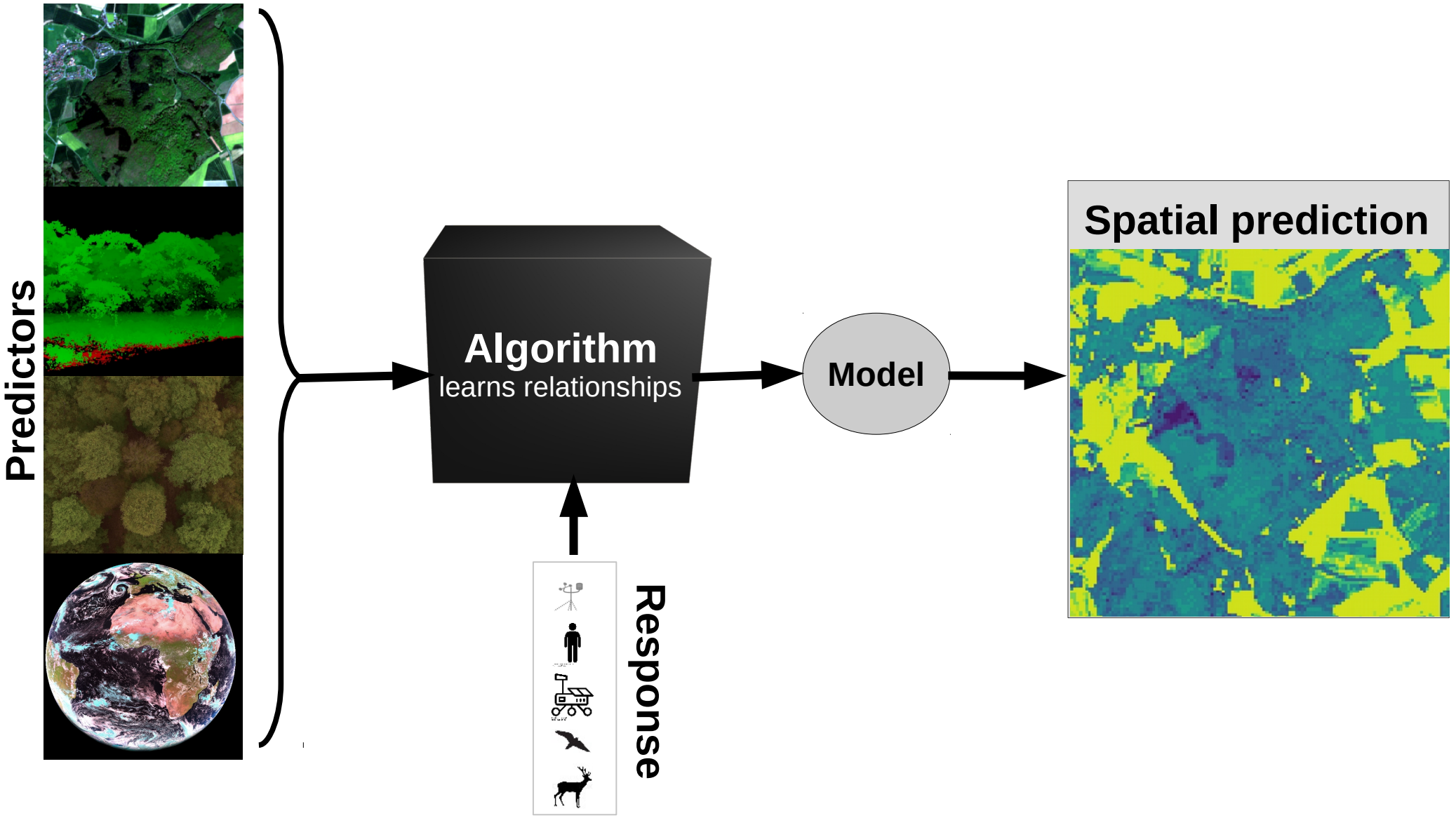


Remote sensing data



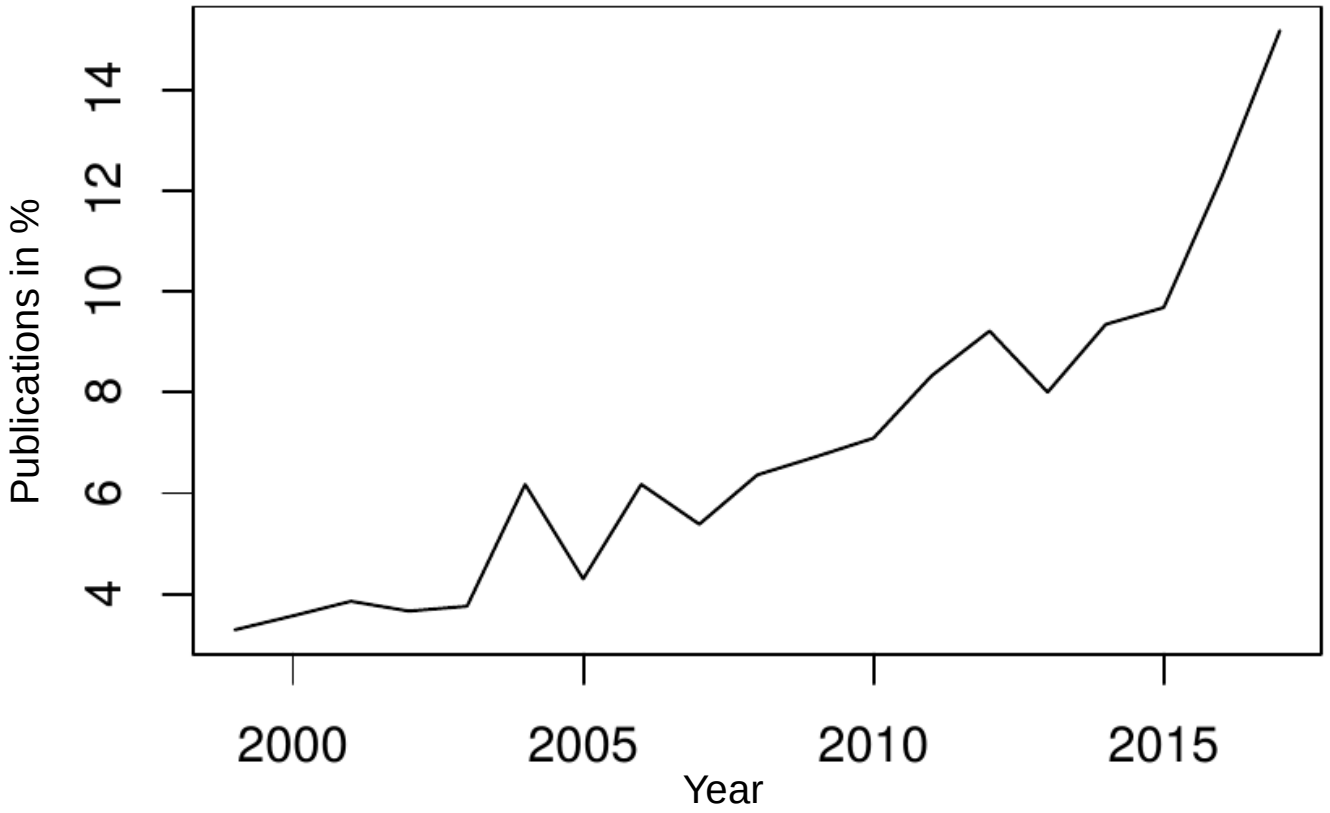
Models that can deal with complex nonlinear relationships are required!

Remote-sensing based monitoring of the environment: The machine learning way



Machine learning for environmental monitoring

Proportion of publications that use machine learning in environmental remote sensing

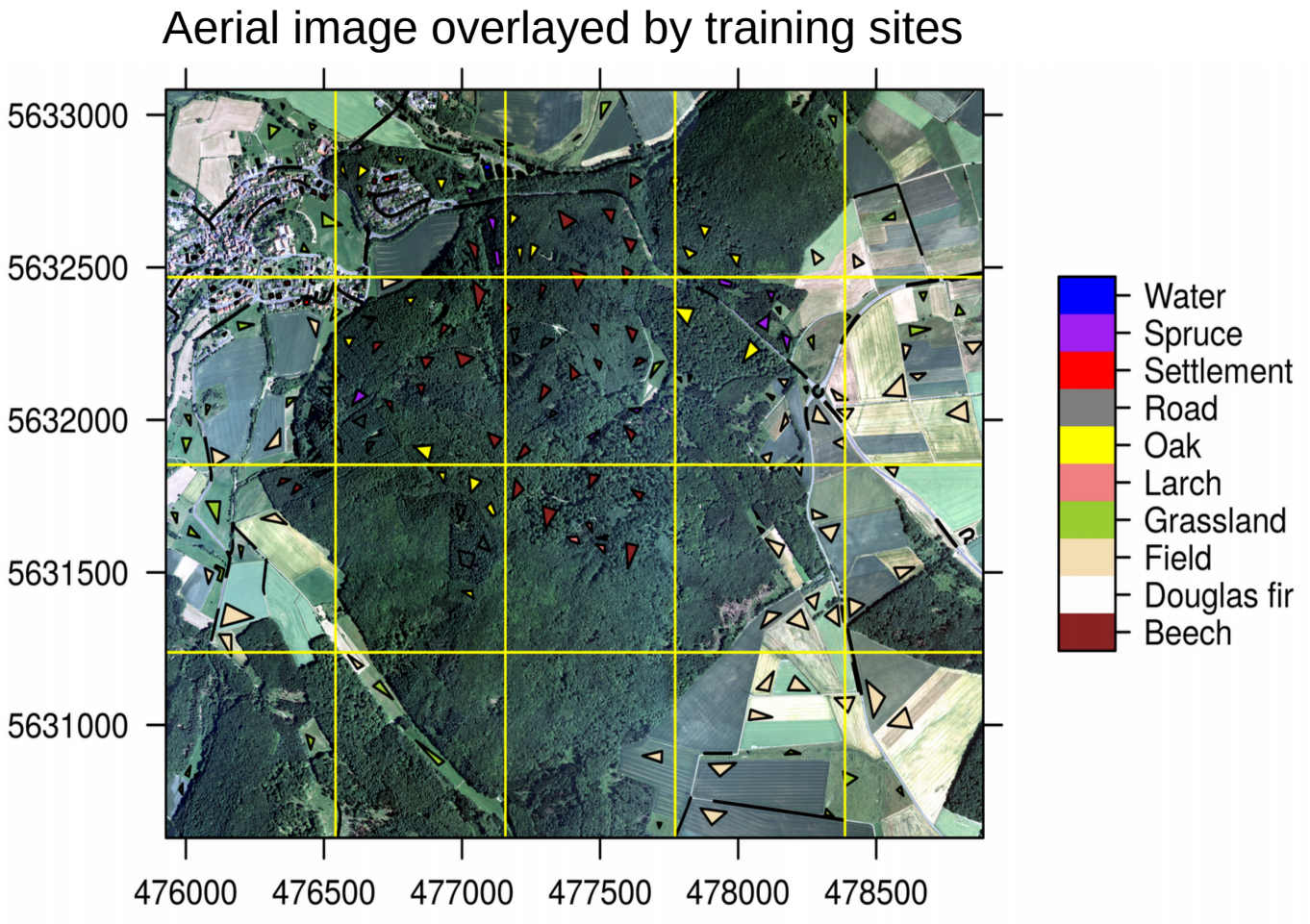


...but characteristics of spatial data are widely ignored

Can we really ignore them?

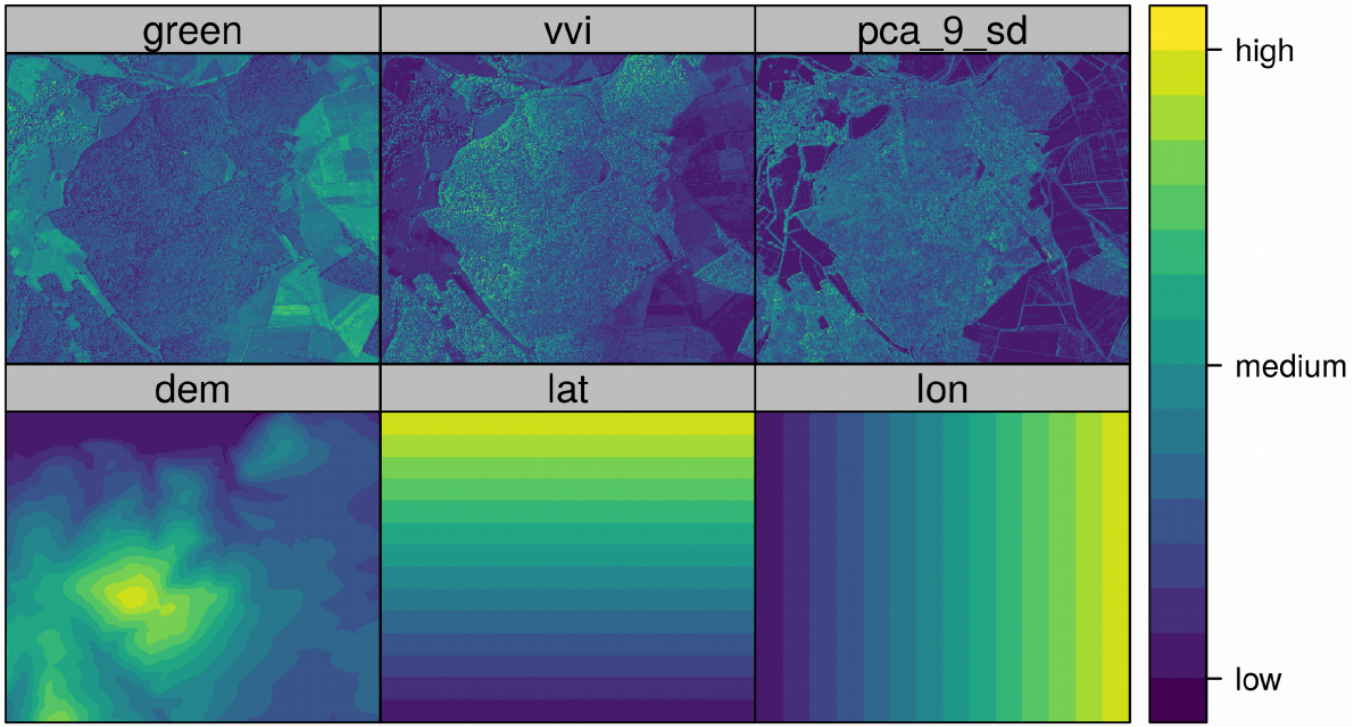
Data based on ScienceDirect search

Case Study: “classic” Land cover classification



Data and algorithm

- Response: Land cover from training polygons
- Predictors: Aerial image RGB, derived indices and texture, terrain, geolocation
- Random Forest algorithm



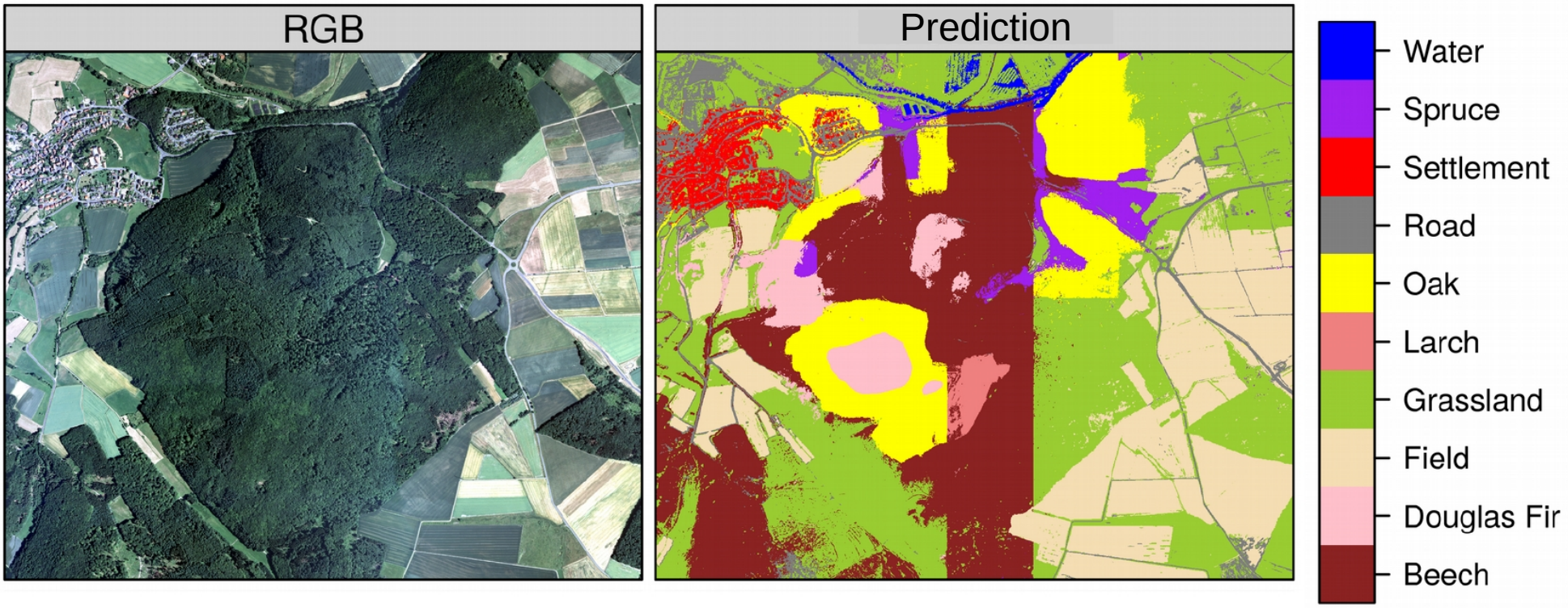
How well can we model land cover with this approach?

Assessment of spatial performance by default validation strategy

Variables	Validation	Accuracy	Kappa
all	random	>0.99	>0.99
all	spatial	**0.68**	**0.61**
selected by FFS spatial	spatial	**0.70**	**0.62**
selected by FFS spatial	random	0.78	0.82

Perfect prediction?

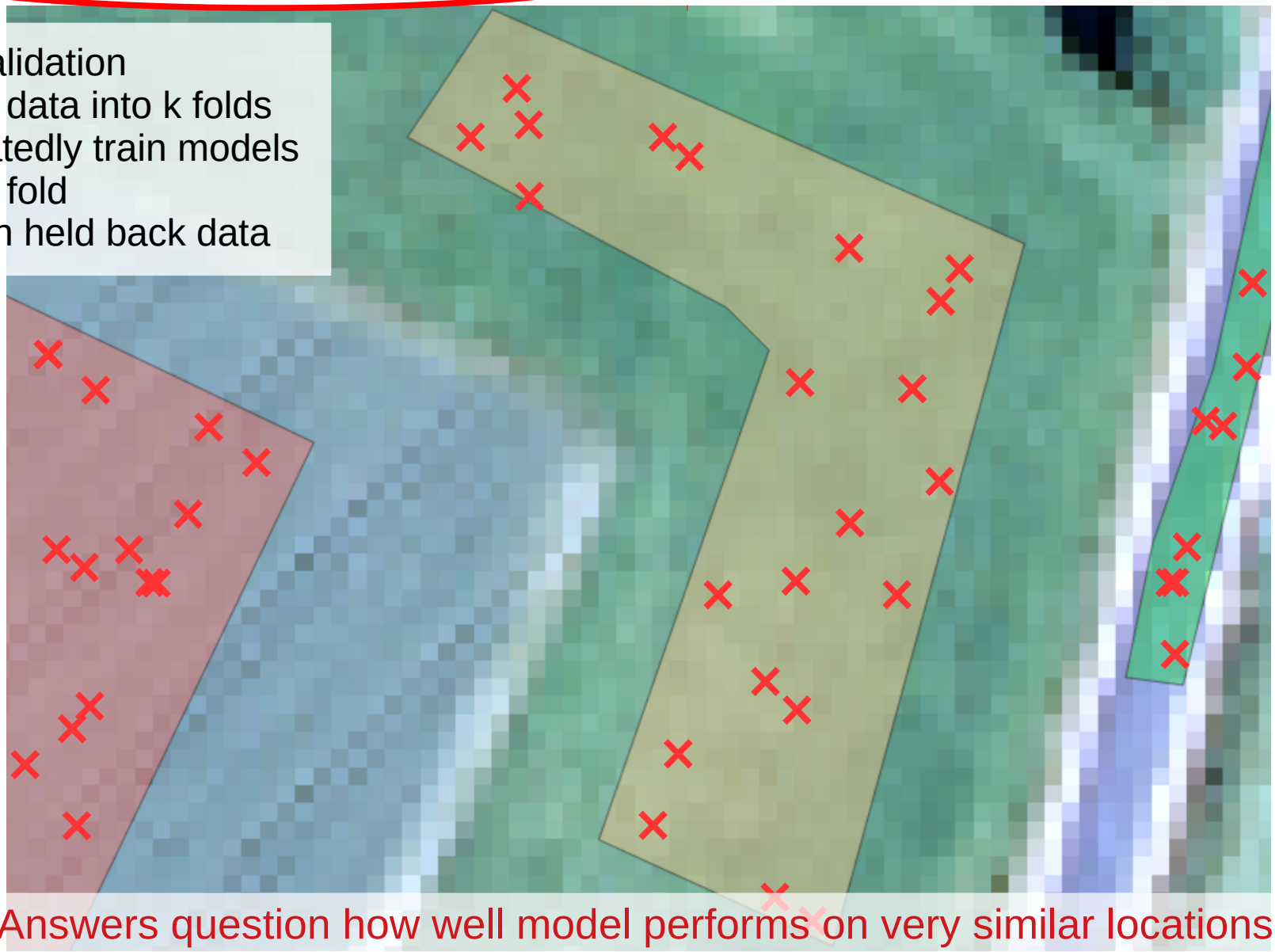
...but it doesn't look like a perfect prediction



But statistically it's a perfect model.
How is this possible?

Assessment of performance by default random cross-validation

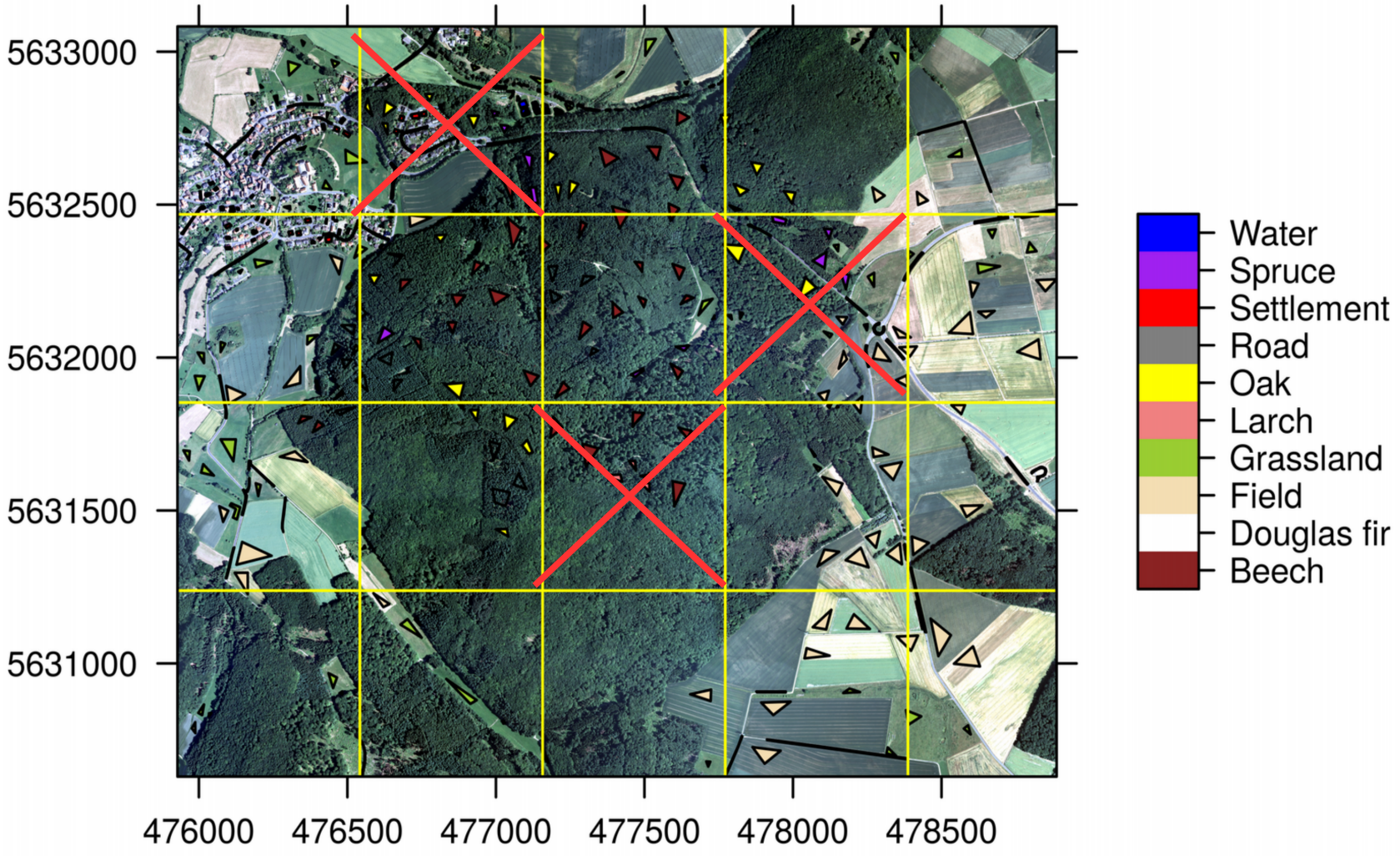
Cross-validation
- Divide data into k folds
- Repeatedly train models on k-1 fold
- Test on held back data



Answers question how well model performs on very similar locations

Assessment of spatial performance

- But the aim is to fill the gaps between sampling locations!
- Spatial cross-validation is required



Assessment of spatial performance

Variables	Validation	Accuracy	Kappa
all	random	>0.99	>0.99
all	spatial	**0.68**	**0.61**
selected by FFS spatial	spatial	**0.70**	**0.62**
selected by FFS spatial	random	0.78	0.82

Standard validation procedures lead to an overoptimistic view on prediction performance!

The relevance of spatial performance estimation is highly underestimated

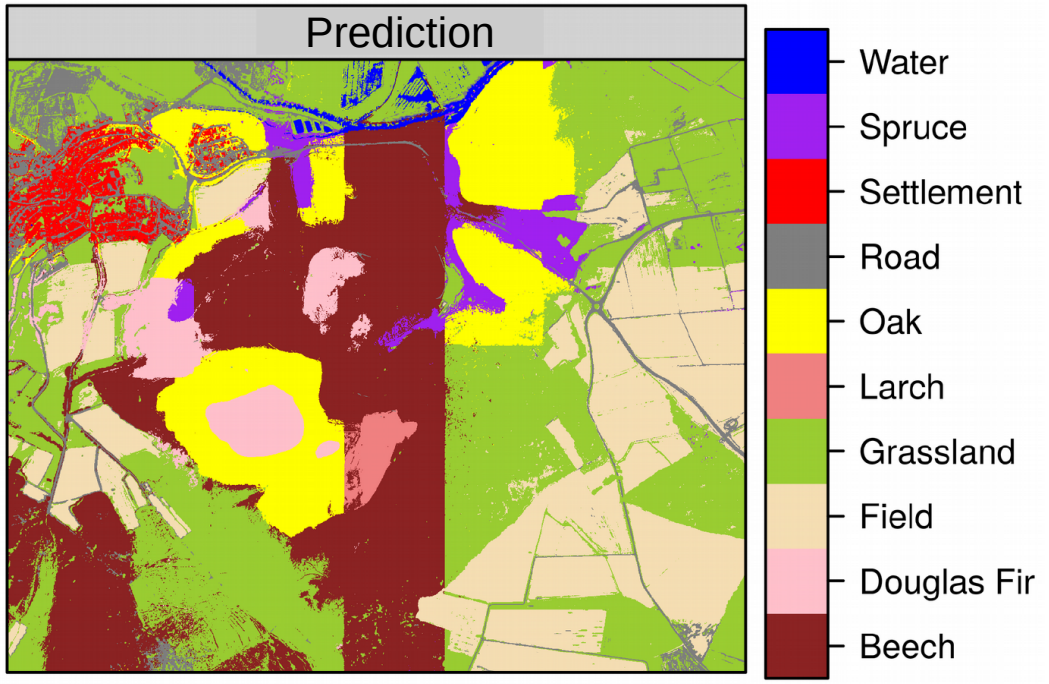
"*I am actually surprised to see the poor performance of your NN approach[...]. Typically with sufficient training data a NN approach can often reproduce the predicted variable very well even if the underlying reasons are unknown*"
(an editor from a journal with high impact in the remote sensing community)

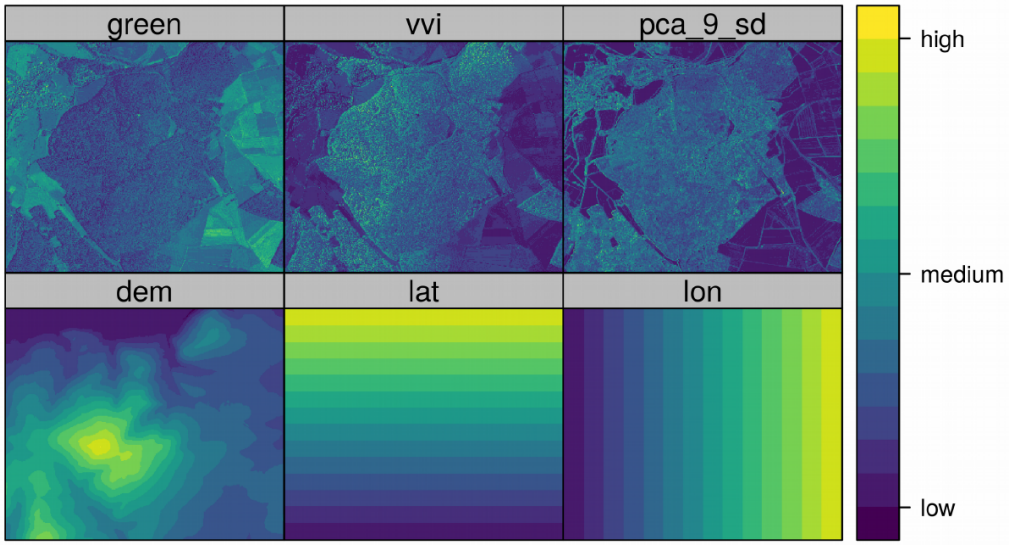
Data reproduction is not the same as data prediction!

Random cross-validation!

Spatial cross-validation!

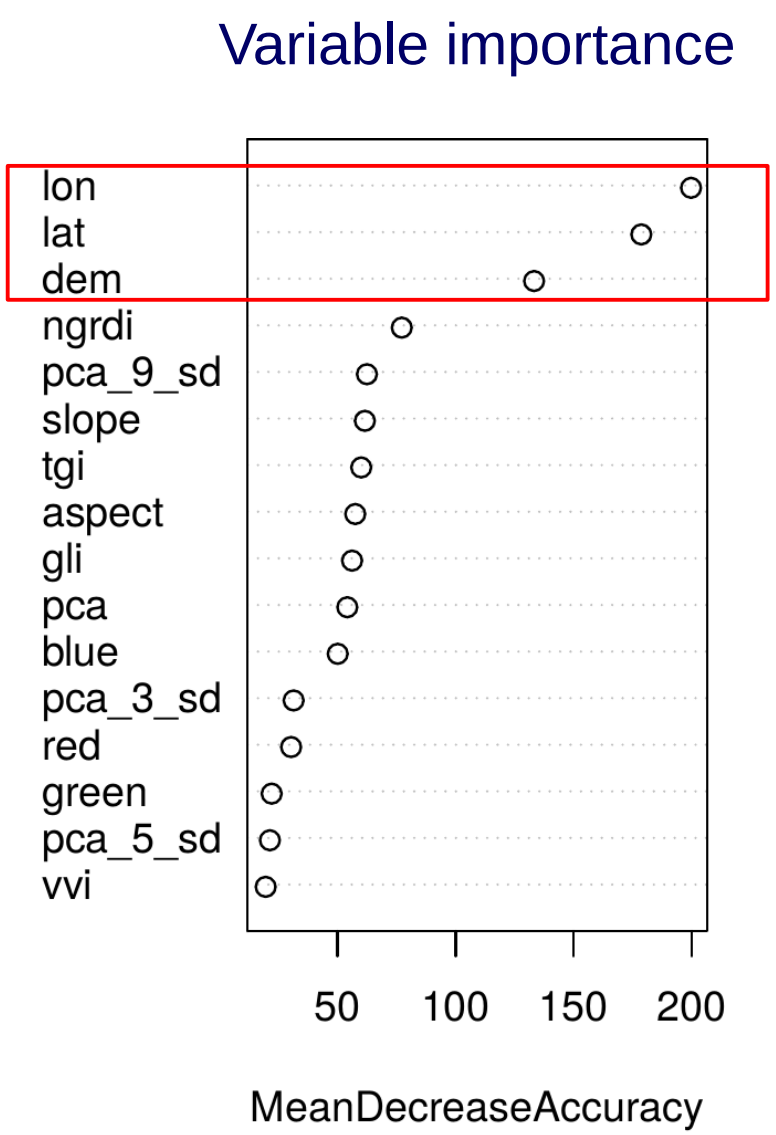
...but spatial performance needs to be improved





Do the spatially autocorrelated predictors lead to overfitting and prevent good spatial predictions?

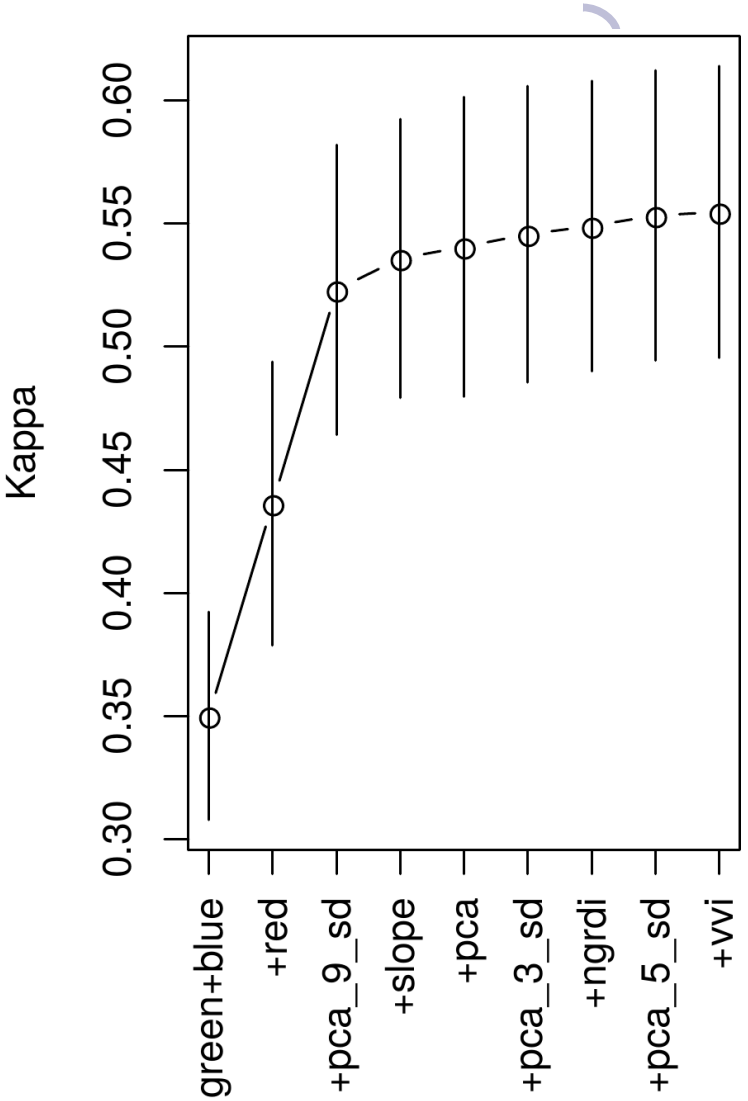
Misinterpretation of autocorrelated predictors?



- Removing variables that lead to overfitting should improve the results
- Spatial variable selection required

Spatial Variable Selection

```
for each resampling iteration do
    Partition the data into training
    Tune and train models using
    Predict on test data and calcu
end
Keep the best performing 2-variat
for each additional number of va
    for each remaining variable
        for each resampling itera
            Partition the data int
            Tune and train mode
            Predict on test data a
        end
    end
    if mean(error of model_i) > m
        break
    end
    Keep the best performing i-variable model (model_best)
end
```

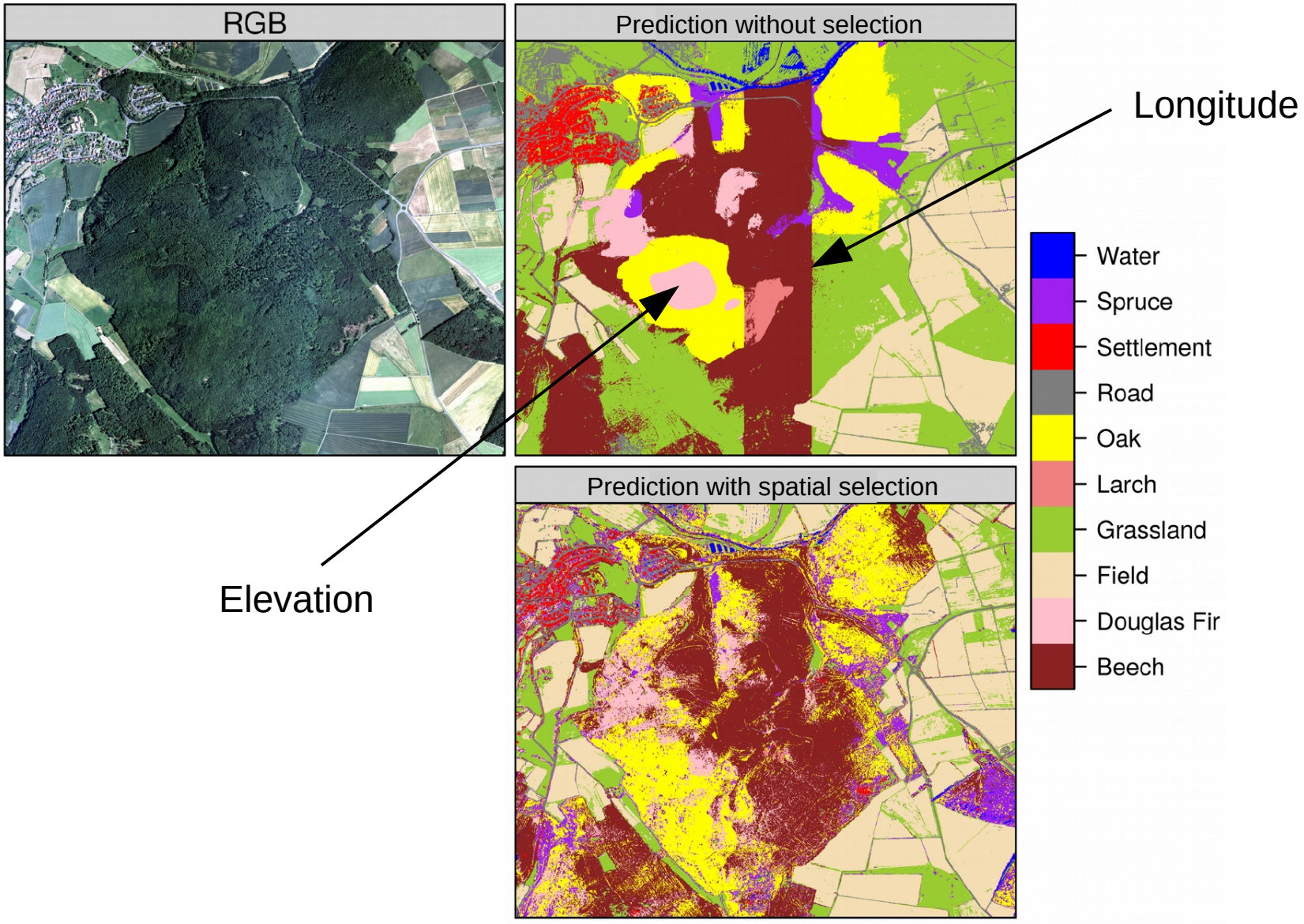


Which 2 variables lead to the best model?

Which further variables improve the model?

Function implemented in Meyer 2018: CAST: 'caret' Applications for Spatial-Temporal Models. R package version 0.2.0.

Improved performance by spatial variable selection



Statistical performance of the spatial model

Variables	Validation	Accuracy	Kappa
all	random	>0.99	>0.99
all	spatial	**0.68**	**0.61**
selected by FFS spatial	spatial	**0.70**	**0.62**
selected by FFS spatial	random	0.78	0.82

Conclusions

Spatial dependencies need to be considered!

- Standard validation procedures lead to an overoptimistic view on prediction performance
- Spatial validation is essential!
- Spatial dependencies cause misinterpretations and overfitting
- Spatial variable selection is required!

→ To answer ecological questions, we need to develop (and apply) methods not for data reproduction but for spatial **prediction!**

Thank you for your interest!

References & further reading

- Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T. (accepted): Importance of spatial predictor variable selection in machine learning applications - Moving from data reproduction to spatial prediction. Ecological Modelling. https://arxiv.org/abs/1908.07805
- Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T. (2018): Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environmental Modelling & Software, 101, 1-9.
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017): Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. doi:10.1111/ecog.02881.
- Schratz, P., Muenchow, J., Iturritx, E., Richter, J., Brenning, A. (2019): Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling 406: 109-120. https://doi.org/10.1016/j.ecolmodel.2019.06.002
- Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G.(2019): blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol Evol. 10: 225-232. https://doi.org/10.1111/2041-210X.13107