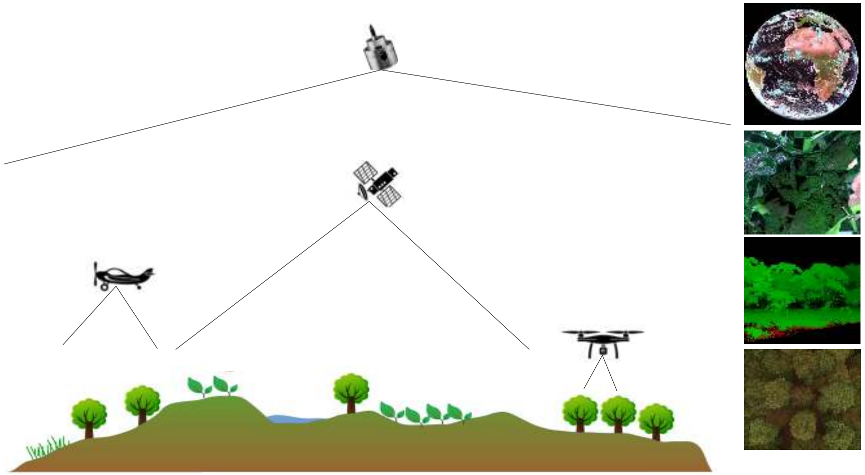
Predictive modelling of spatial (or spatio-temporal) environmental data -

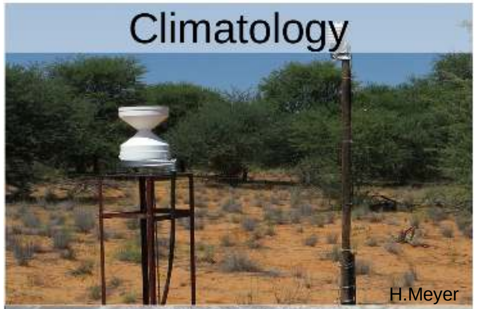
Moving from field observations to maps of ecosystem variables

Hanna Meyer

Institute for Geoinformatics, WWU Münster



Common research aims in environmental science


Climatology

H.Meyer

Monitoring of spatio-temporal rainfall dynamics

Soil Science

N.Meyer

Revealing spatial patterns of soil properties

Ecosystem science

H.Meyer

Explaining spatio-temporal patters of vegetation

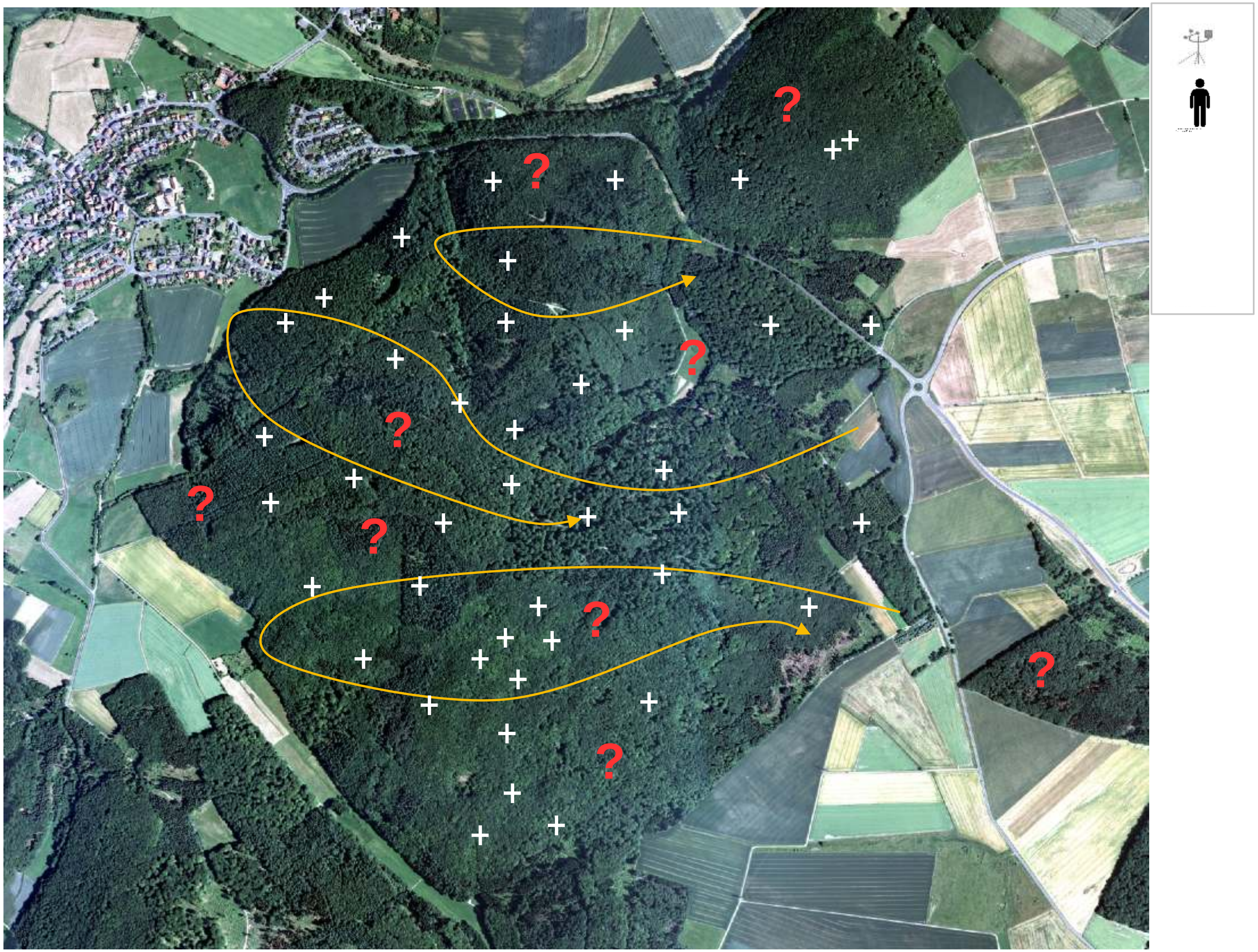
Animal ecology

H.Meyer

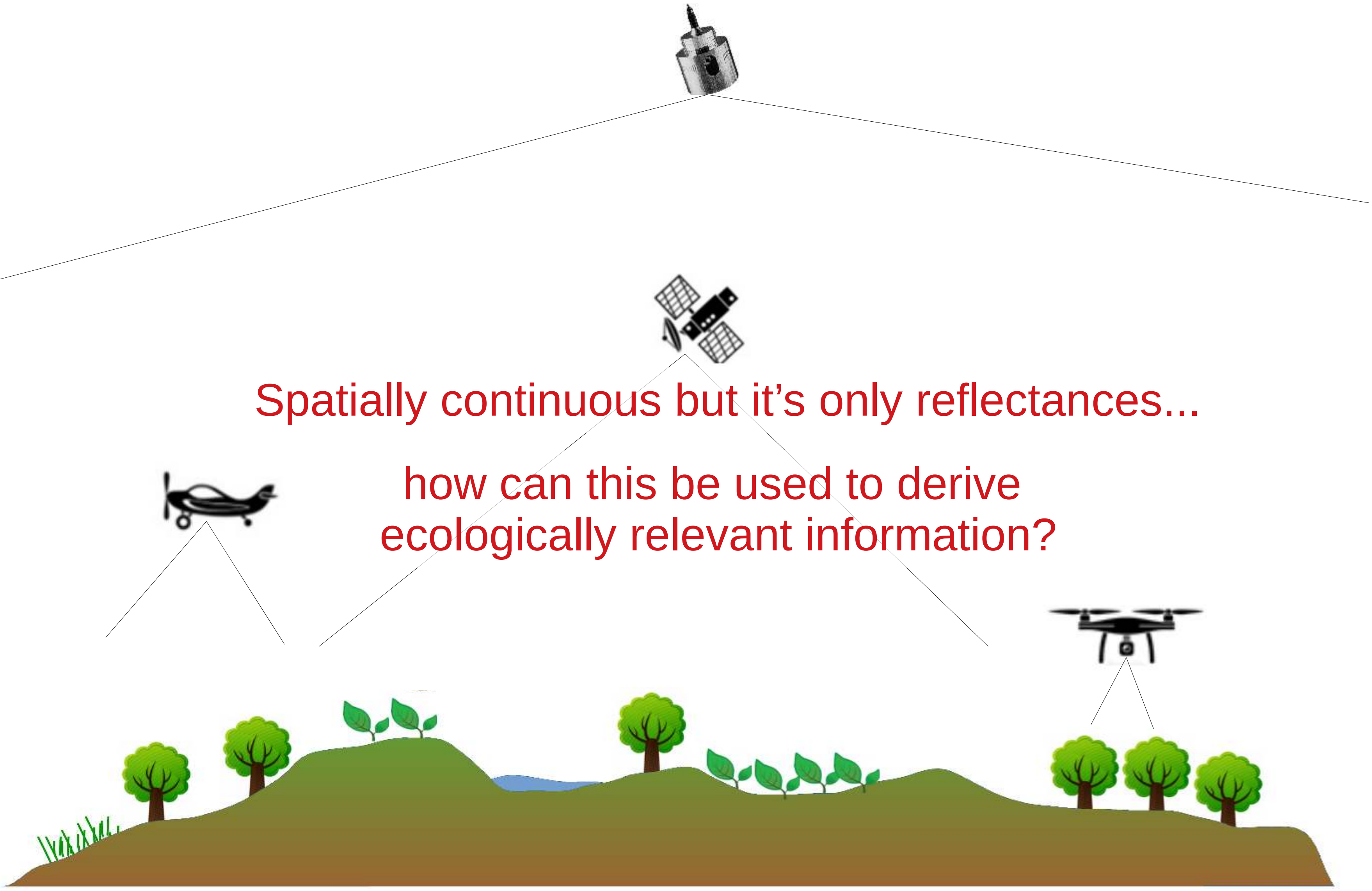
Studying distribution and dynamics of animal species

Problem: Moving from field observations to maps of ecosystem variables

Nature 4.0 | Sensing Biodiversity

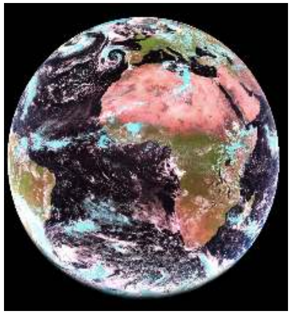


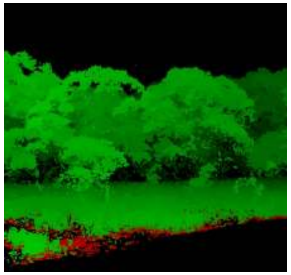
Remote Sensing of landscapes



Spatially continuous but it’s only reflectances...

how can this be used to derive
ecologically relevant information?

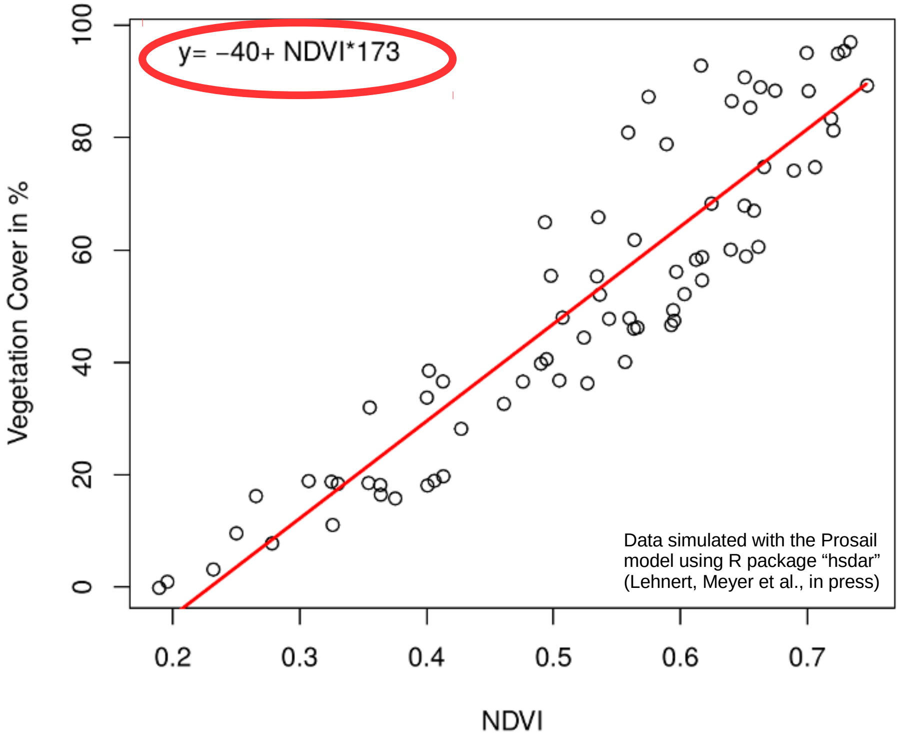




Direct and indirect sensing of the environment

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

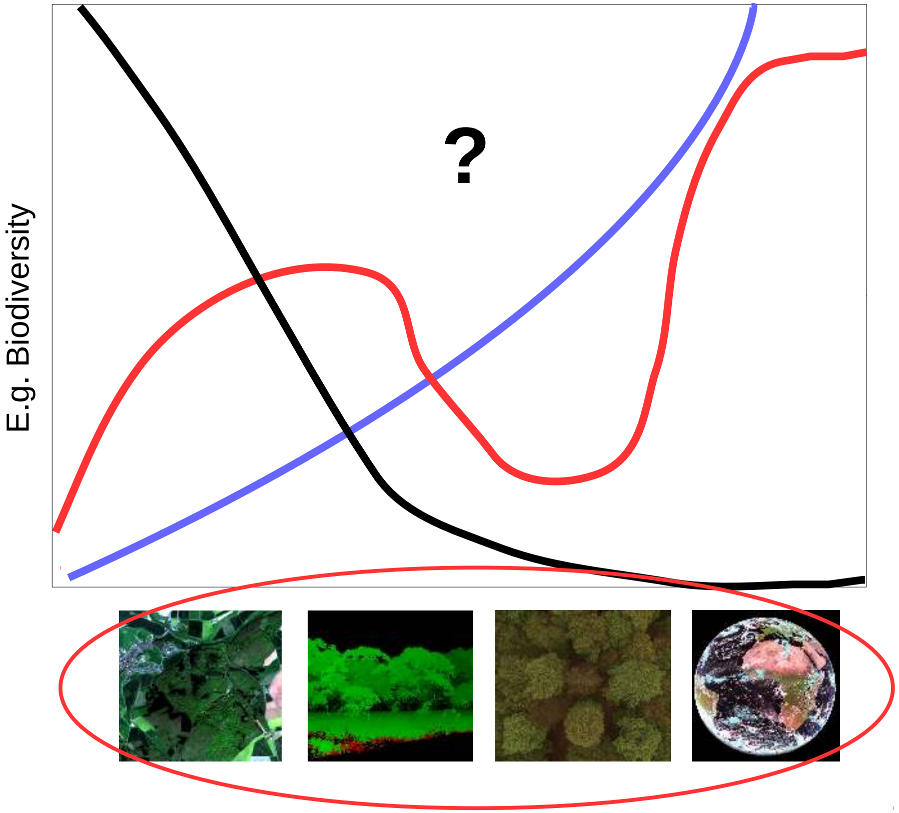
Field data



Data simulated with the Prosail model using R package “hsdar” (Lehnert, Meyer et al., in press)

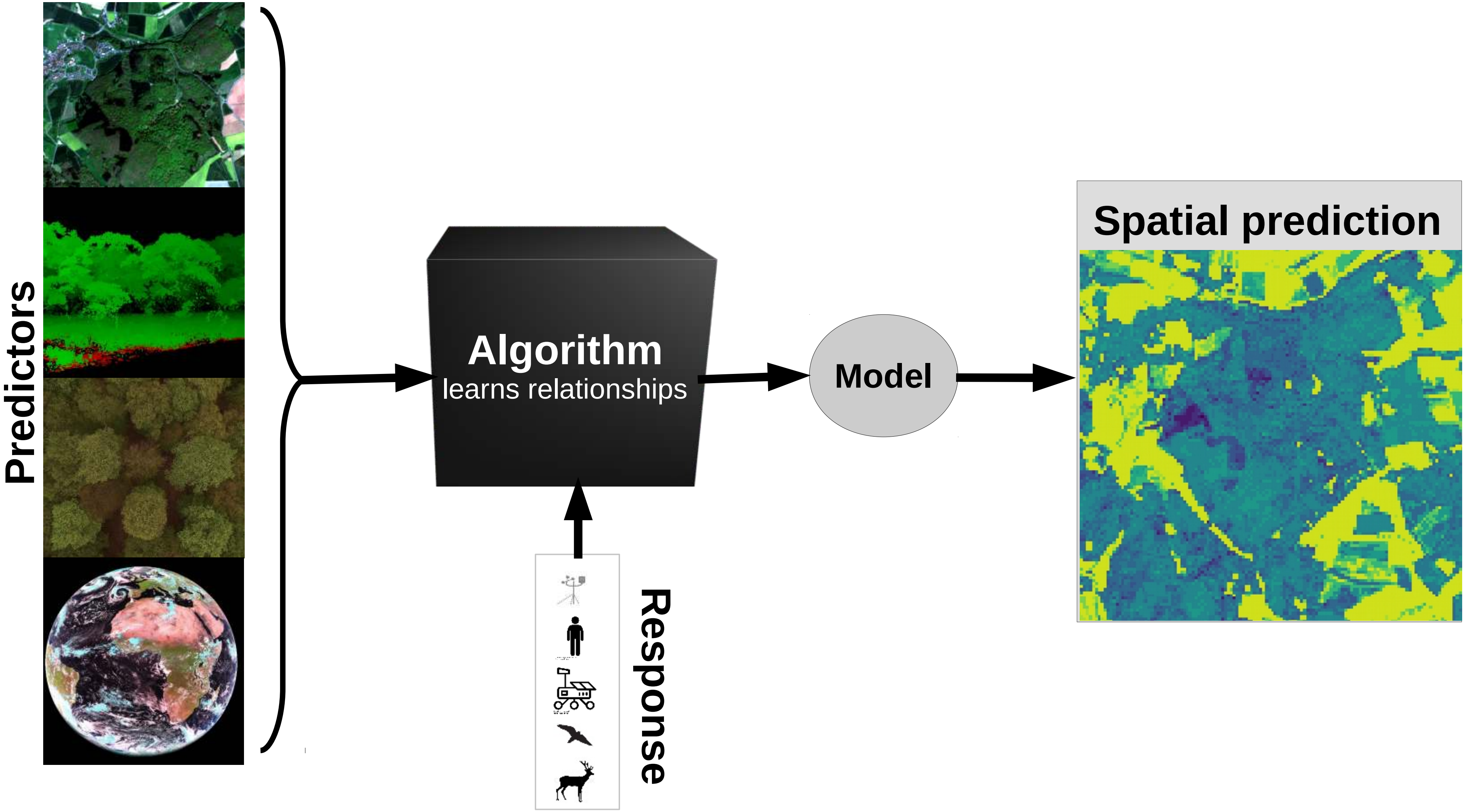
Remote sensing data

Typical ecological variables from satellite?



Models that can deal with complex nonlinear relationships are required!

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



Predictors

Algorithm

learns relationships

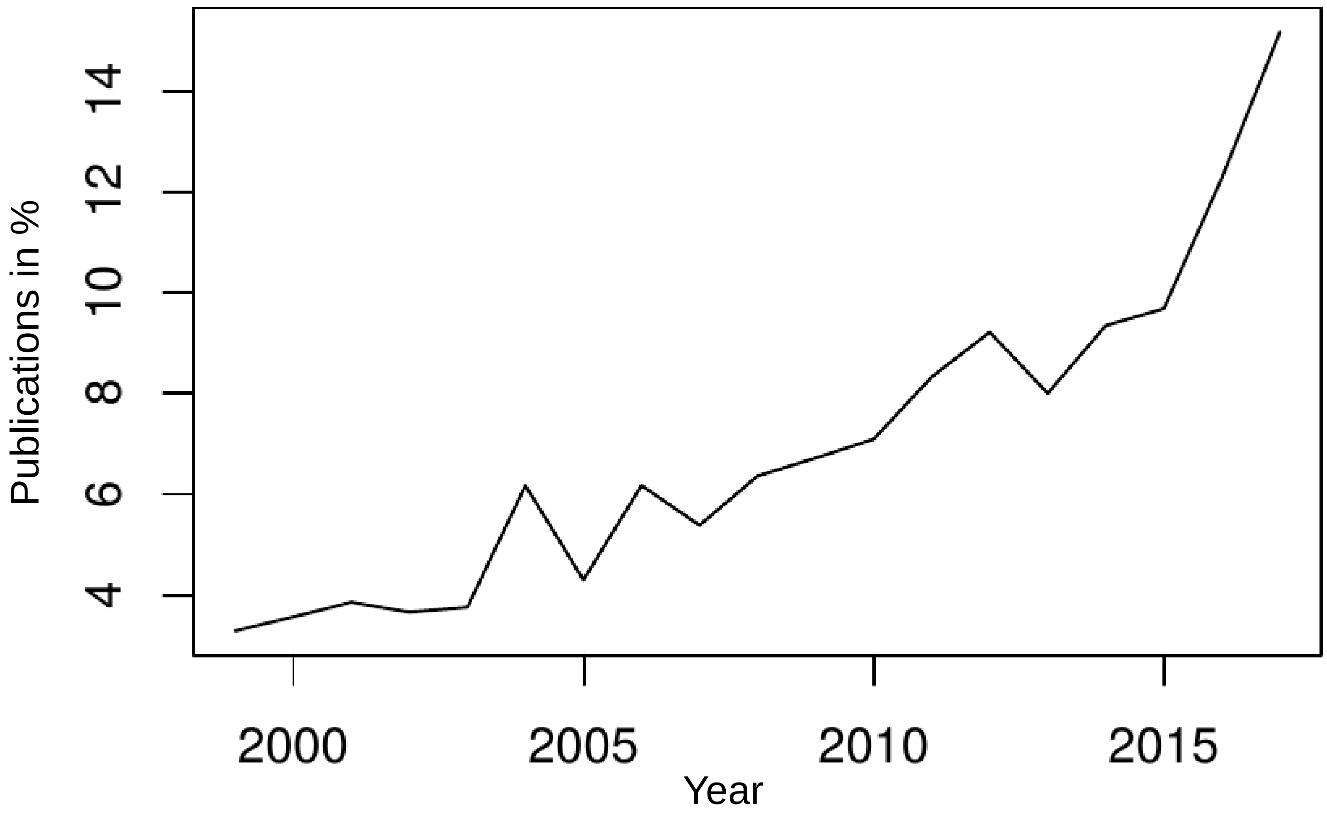
Model

Spatial prediction

Response

Machine learning for environmental monitoring

Proportion of publications that use machine learning in environmental remote sensing



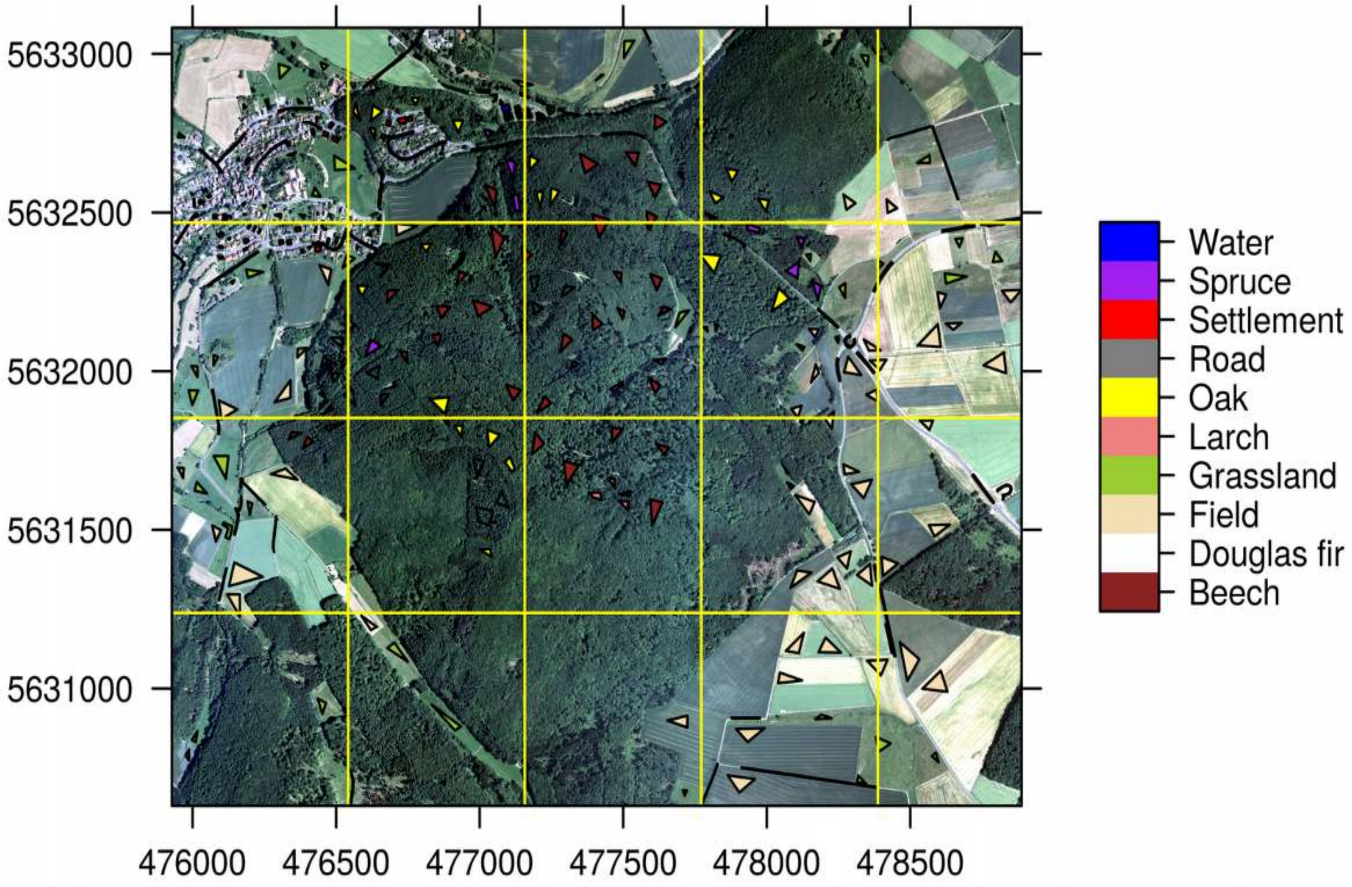
...but characteristics of spatial data are widely ignored

Can we do this?

Data based on ScienceDirect search

Case Study: “classic” Land cover classification

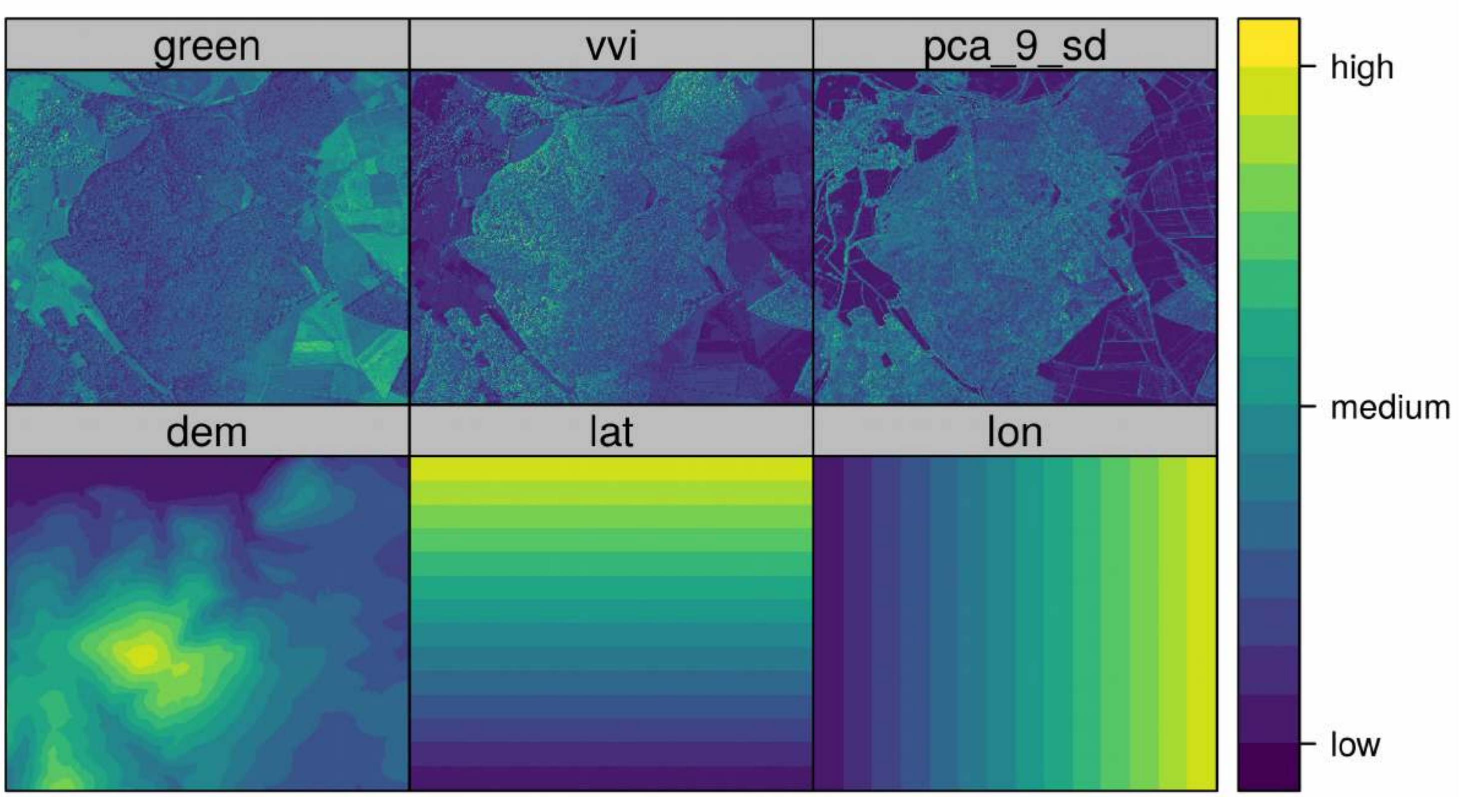
Aerial image overlayed by training sites



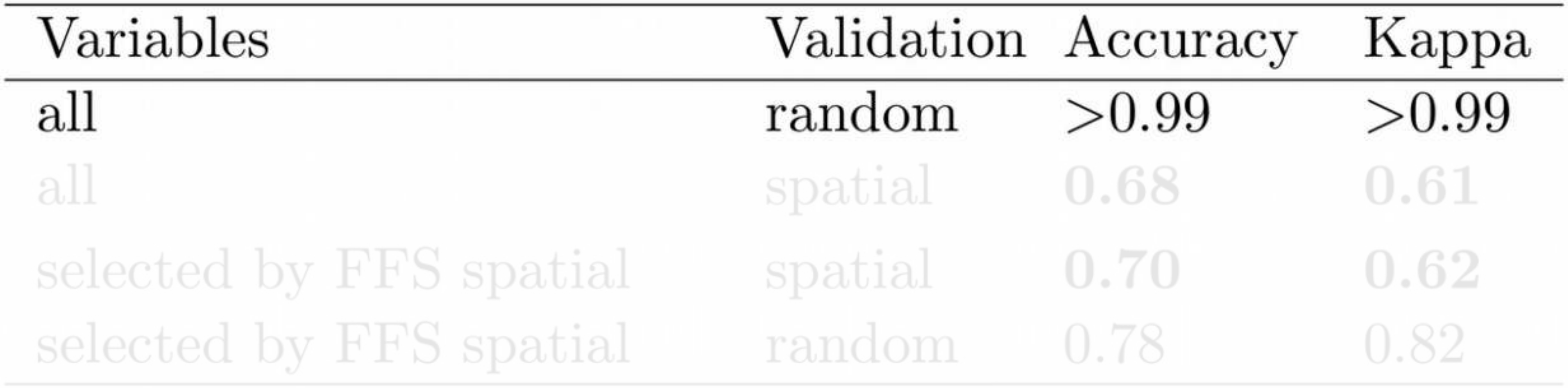
Meyer et al. (in review)

Data and algorithm

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



Assessment of spatial performance by default validation strategy

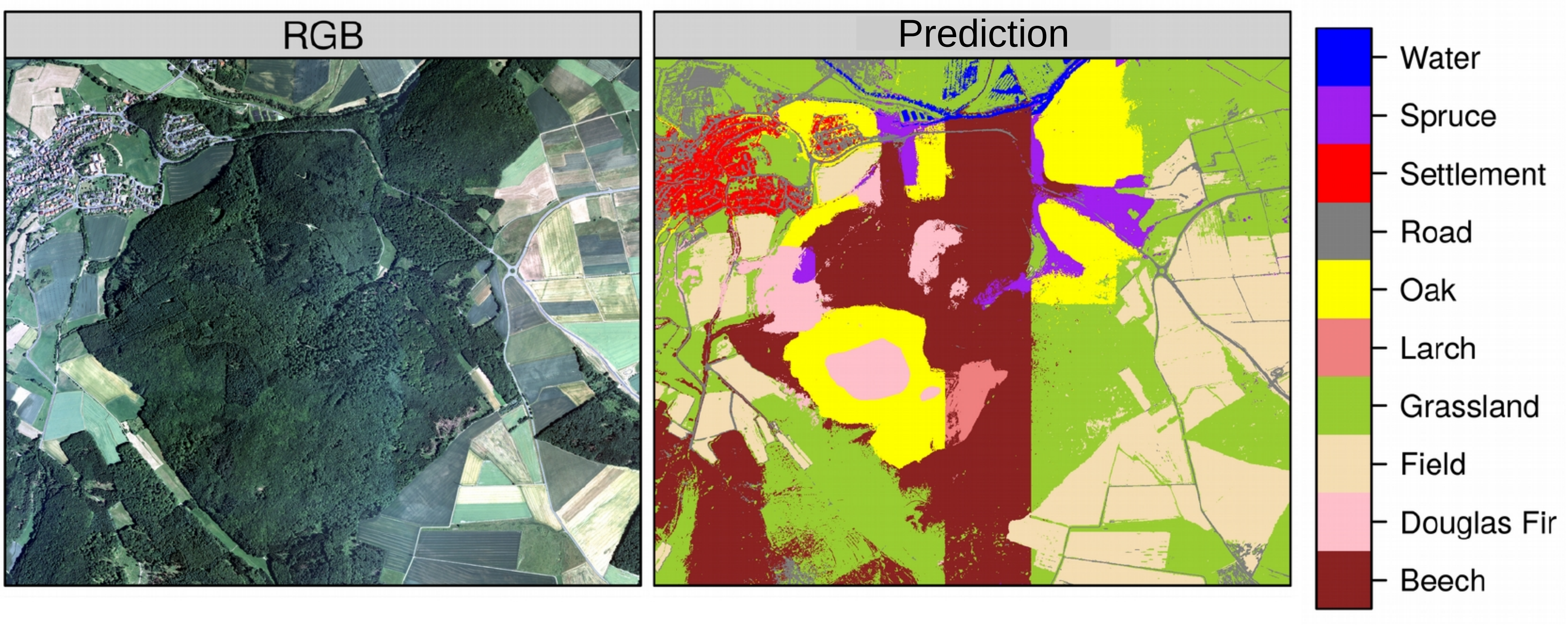


Variables	Validation	Accuracy	Kappa
all	random	>0.99	>0.99
all	spatial	**0.68**	**0.61**
selected by FFS spatial	spatial	**0.70**	**0.62**
selected by FFS spatial	random	0.78	0.82

Perfect prediction?

Meyer et al. (in review)

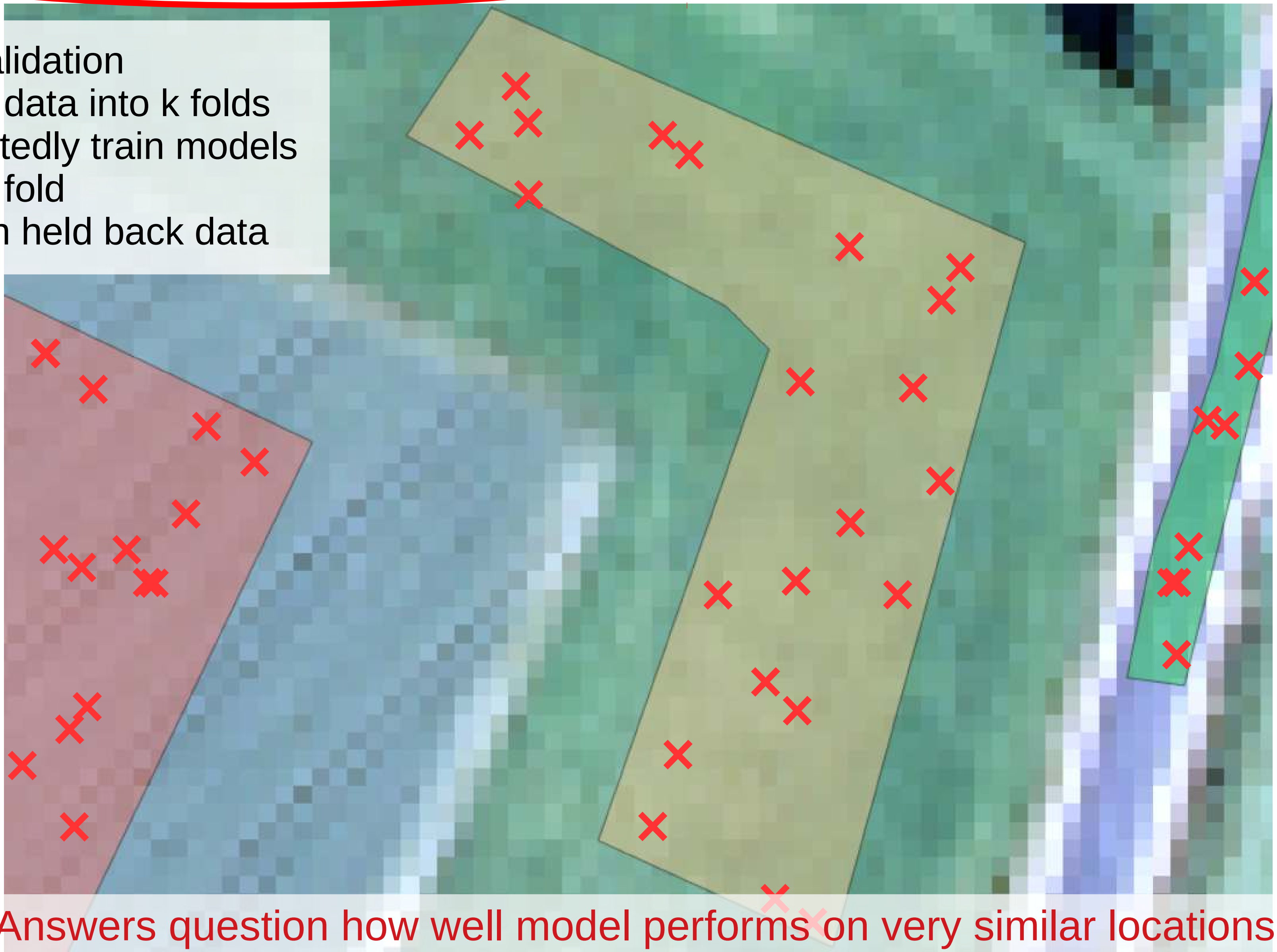
...but it doesn’t look like a perfect prediction



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

Meyer et al. (in review)

Assessment of performance by default random cross-validation



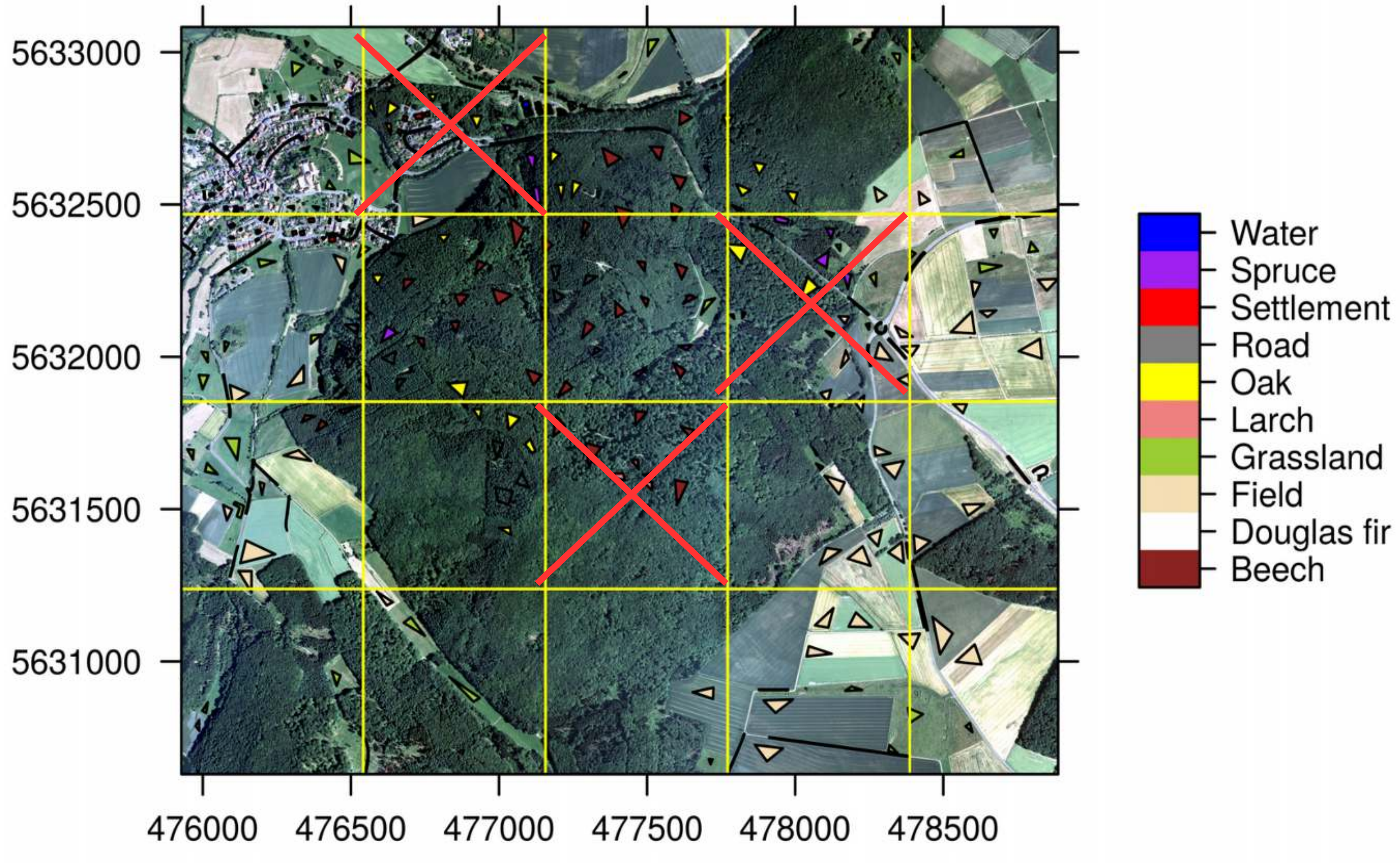
Cross-validation

- 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

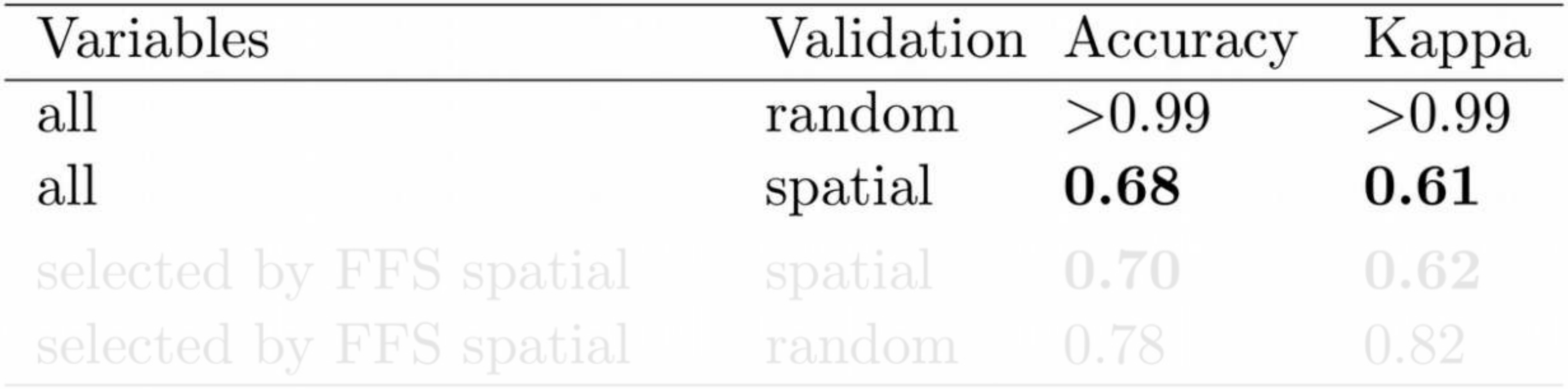
Assessment of spatial performance

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



Meyer et al. (in review)

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!

Meyer et al. (in review)

The relevance of spatial performance estimation is highly underestimated

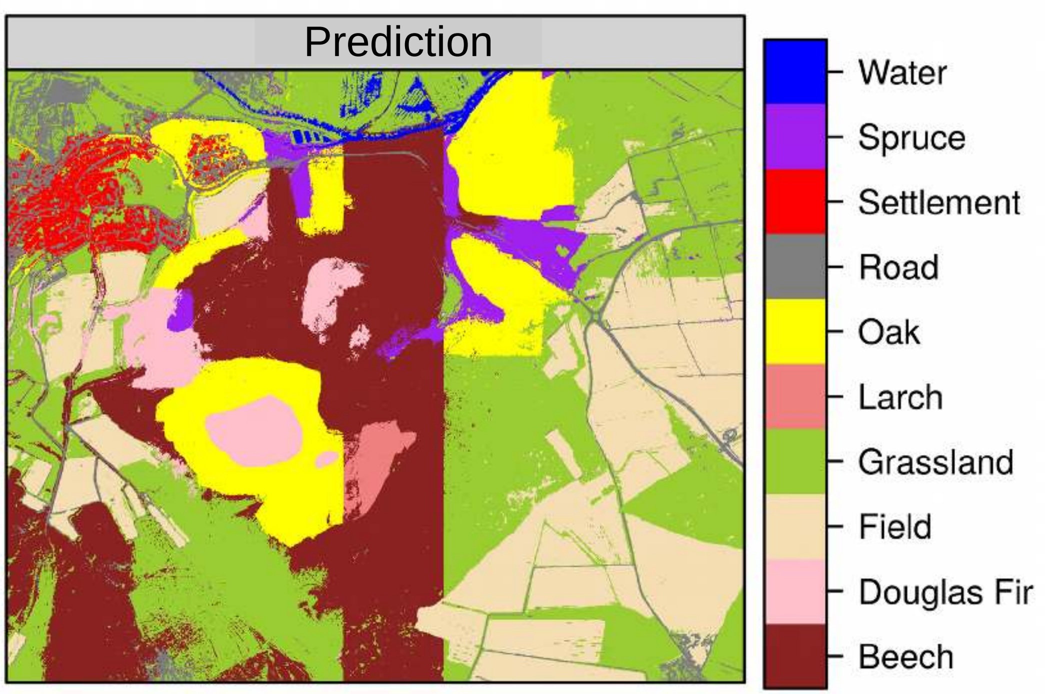
“*I am actually surprised to see the poor performance of your NN approach[...]. Typically with sufficient training data a NN approach can often reproduce the predicted variable very well even if the underlying reasons are unknown*”
(an editor from Remote Sensing of Environment)

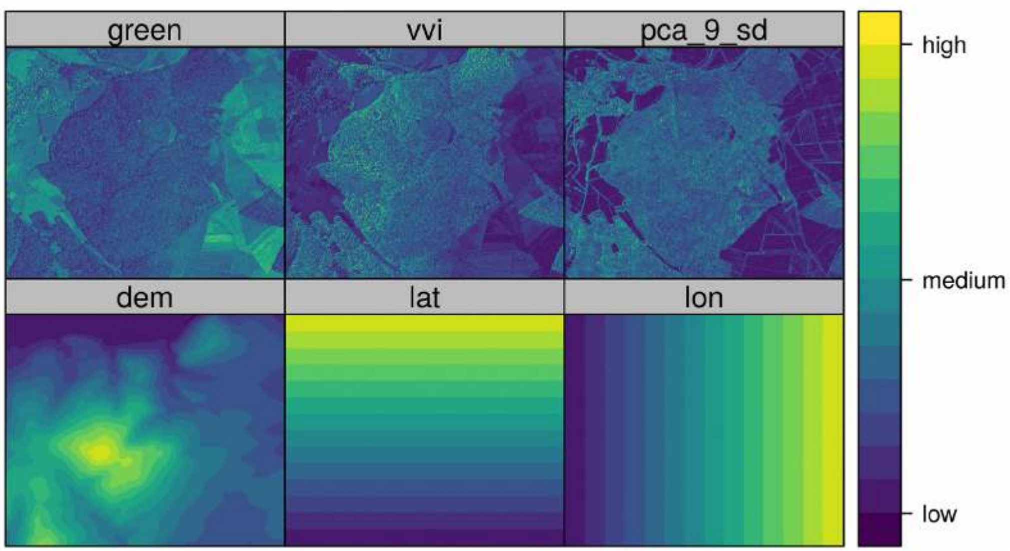
Data reproduction is not the same as data prediction!

Random
cross-validation!

Spatial
cross-validation!

...but spatial performance needs to be improved



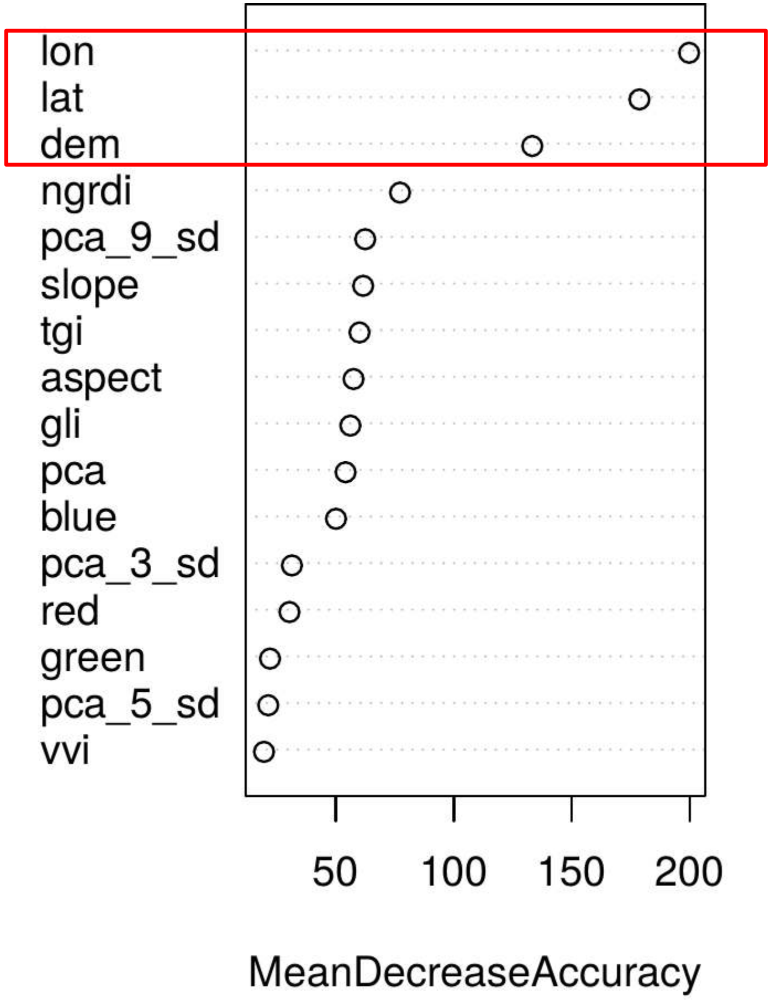


Do the spatially autocorrelated predictors lead to overfitting and prevent good spatial predictions?

Meyer et al. (in review)

Misinterpretation of autocorrelated predictors?

Variable importance

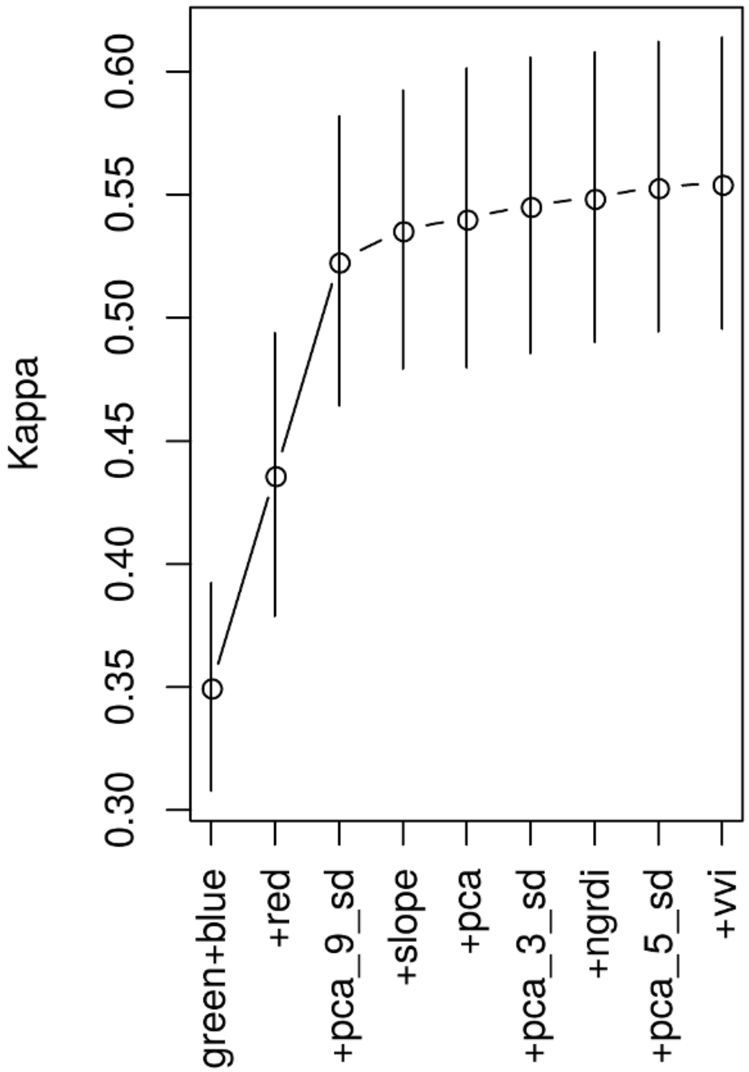


- Removing variables that lead to overfitting should improve the results
- Spatial variable selection required

Meyer et al. (in review)

Spatial Variable Selection

```
for each resampling iteration do
    Partition the data into training
    Tune and train models using a
    Predict on test data and calcu
end
Keep the best performing 2-variał
for each additional number of vaı
    for each remaining variable \
        for each resampling iterc
            Partition the data intı
            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
```

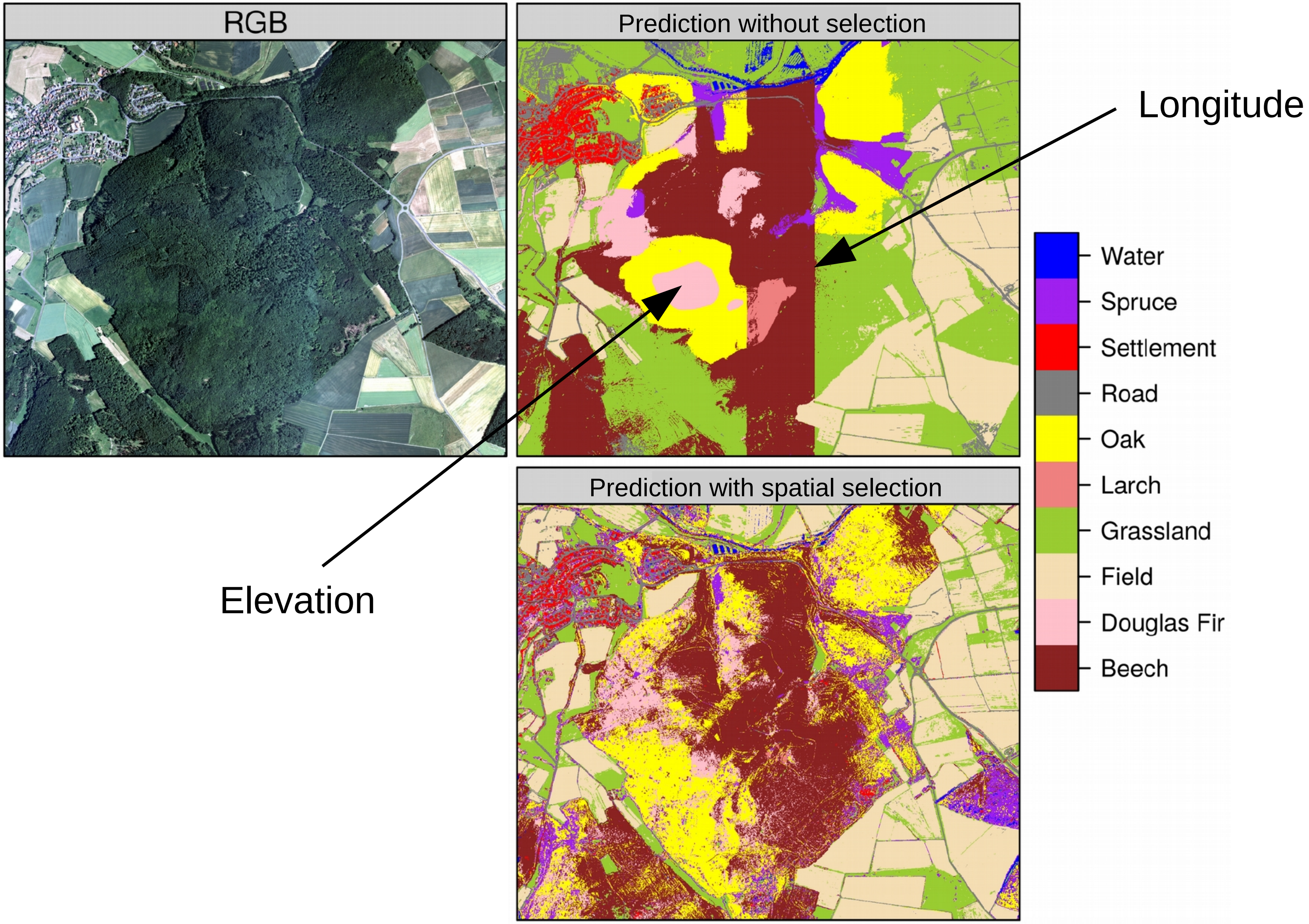


Which 2 variables lead to the best model?

Which further variables improve the model?

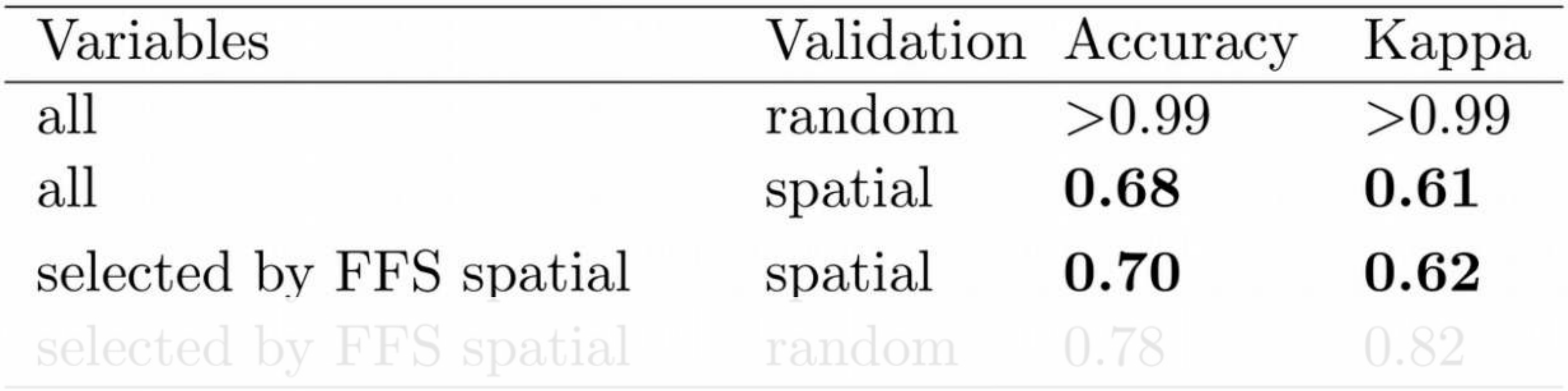
Function implemented in Meyer 2018: CAST: 'caret' Applications for Spatial-Temporal Models. R package version 0.2.0.

Improved performance by spatial variable selection



Meyer et al. (in review)

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

Meyer et al. (in review)

Conclusions

How should the performance of spatial prediction models be assessed?

- Standard validation procedures lead to an overoptimistic view on prediction performance
- Spatial validation is essential!

How can the performance be improved?

- Spatial dependencies cause misinterpretations and overfitting
- Spatial variable selection required!

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