RETHINKING VALIDATION FOR SPATIAL MACHINE LEARNING

Jakub Nowosad (Adam Mickiewicz University, Poznań and University of Münster)
Machine Learning for Earth Observation 2026, Exeter, UK
2026-06-22

Roles of spatial machine learning?

Explanatory: understanding the world

Predictive: forecasting the future or mapping the present

Both: understanding the world to map the present/future *(holy grail)*

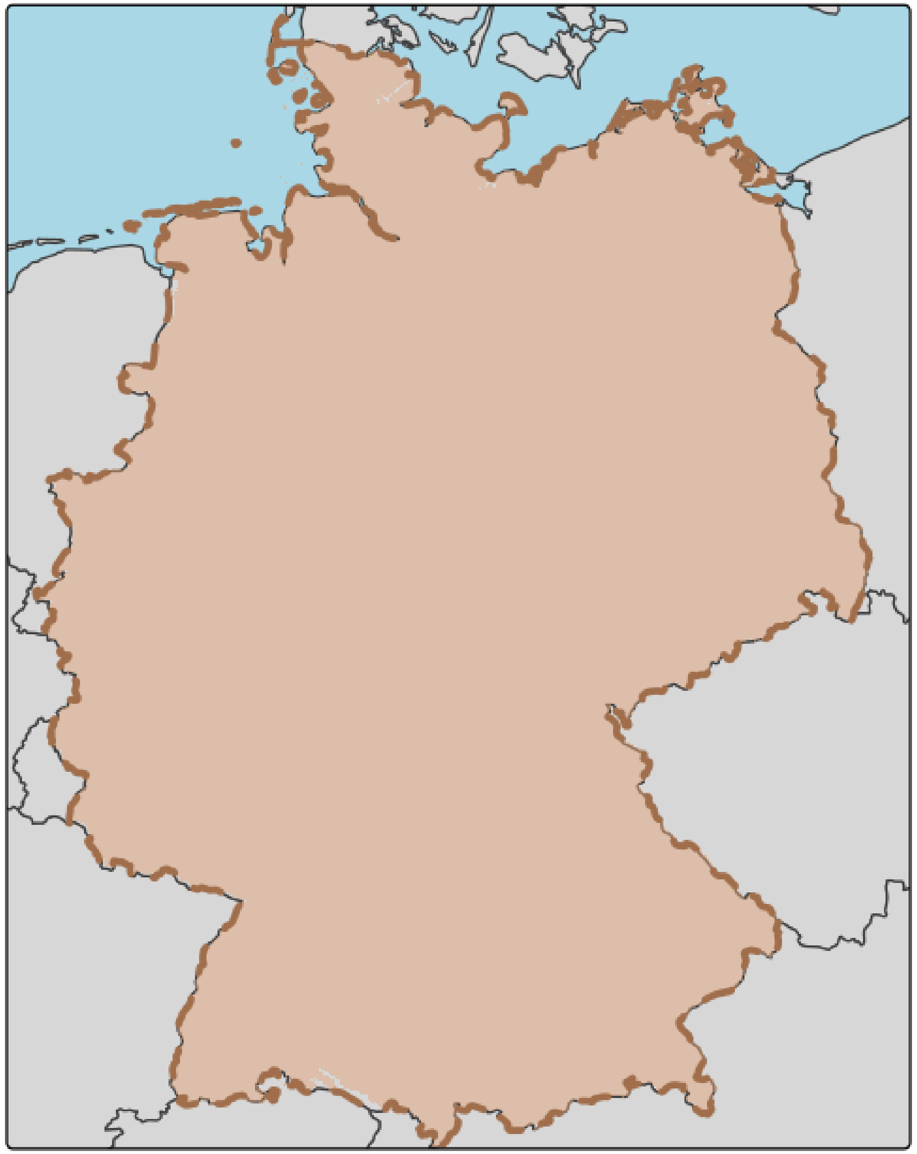
Roles of spatial machine learning?

Explanatory: understanding the world

Predictive: forecasting the future or mapping the present

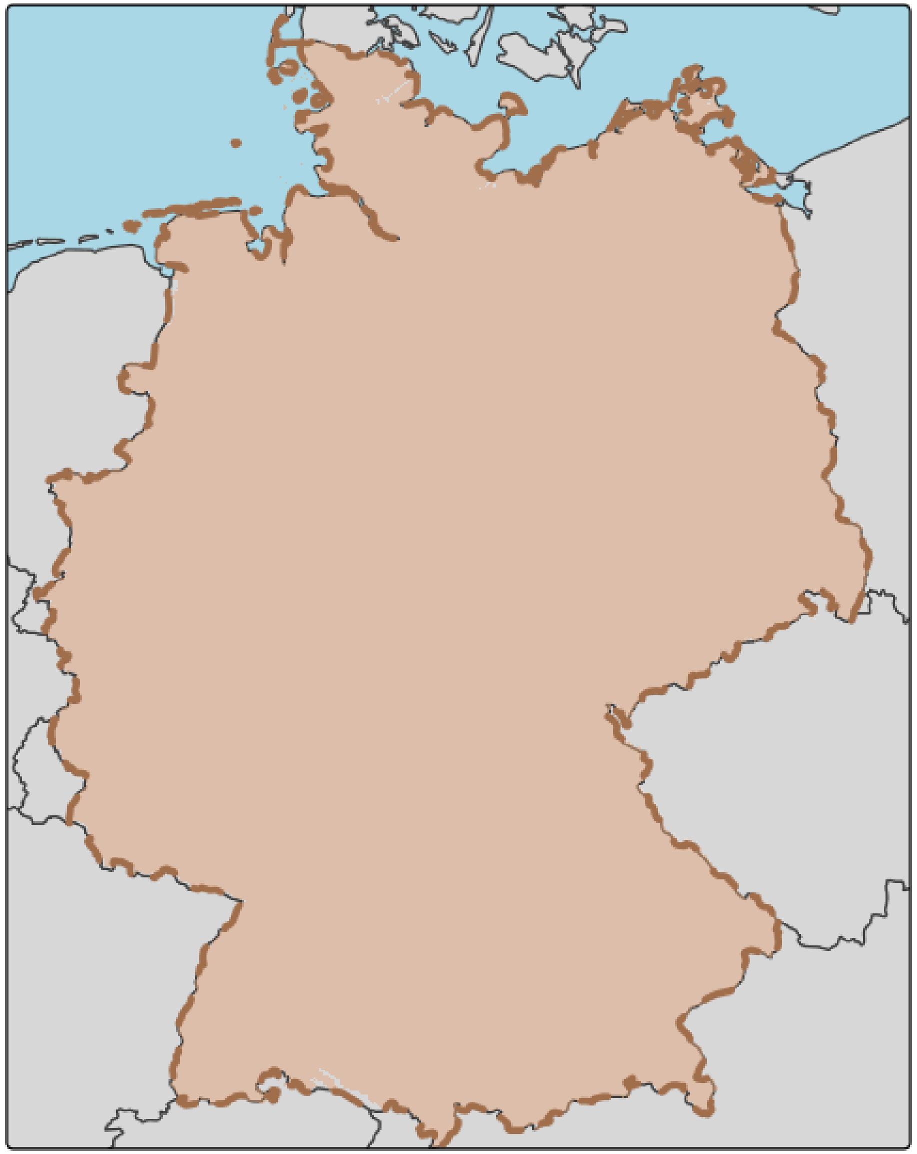
Both: understanding the world to map the present/future *(holy grail)*

Predictive spatial machine learning



Prediction domain

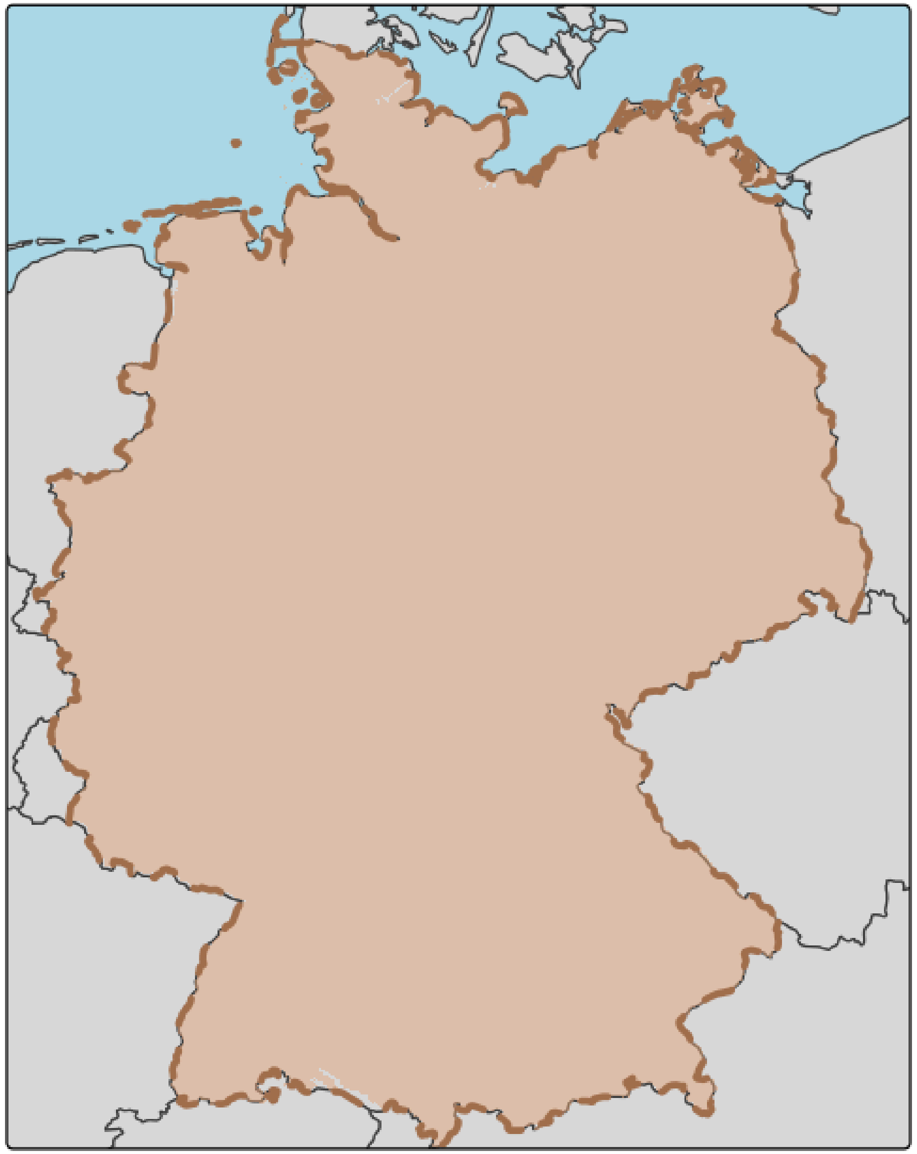
Predictive spatial machine learning



Prediction domain

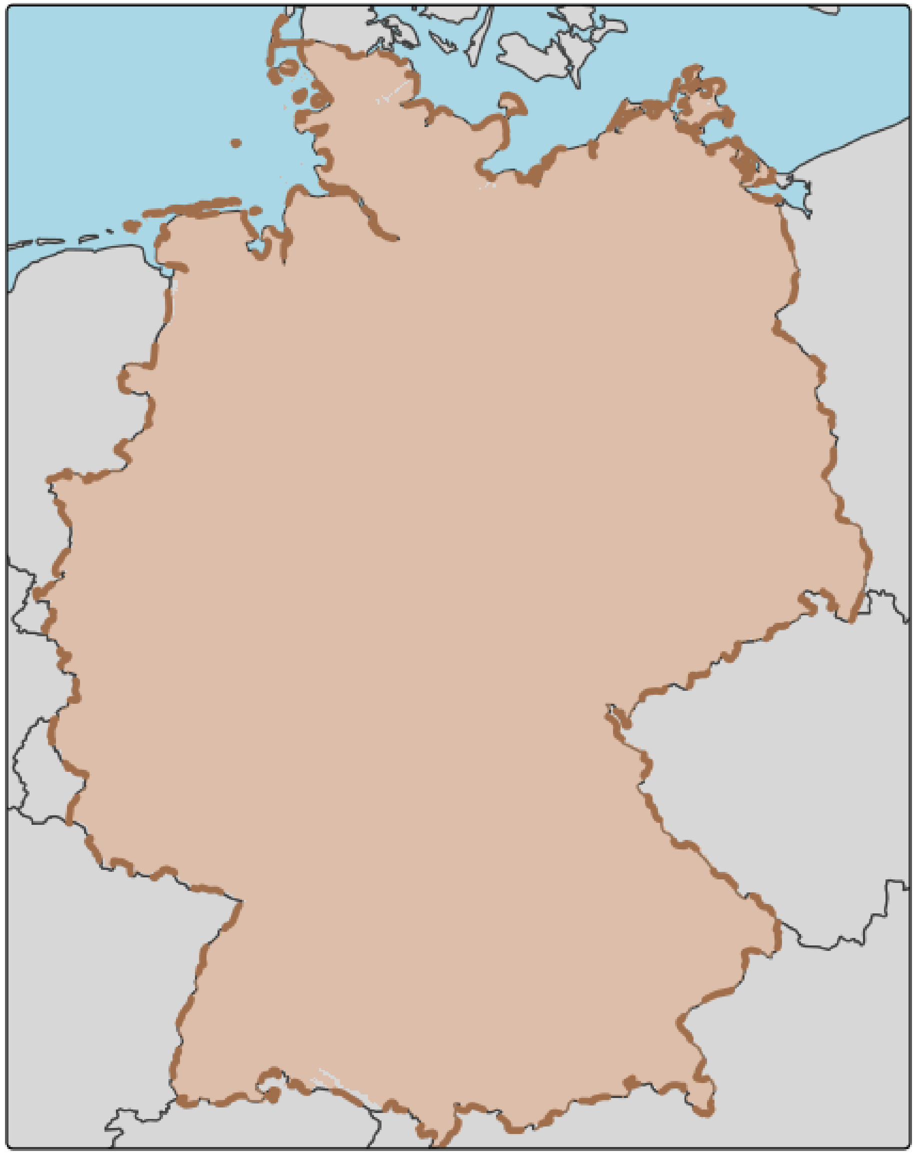
- We are interested in mapping across the spatial domain

Predictive spatial machine learning



Prediction domain

- We are interested in mapping across the spatial domain
- Thus, we should evaluate the map, not the model



Prediction domain

Predictive spatial machine learning

- We are interested in mapping across the spatial domain
- Thus, we should evaluate the map, not the model
- **But this raises a key question:** *What assumptions are we making about spatial evaluation?*

ASSUMPTION #1

Assumptions we make

We can predict everywhere

Assumptions we make

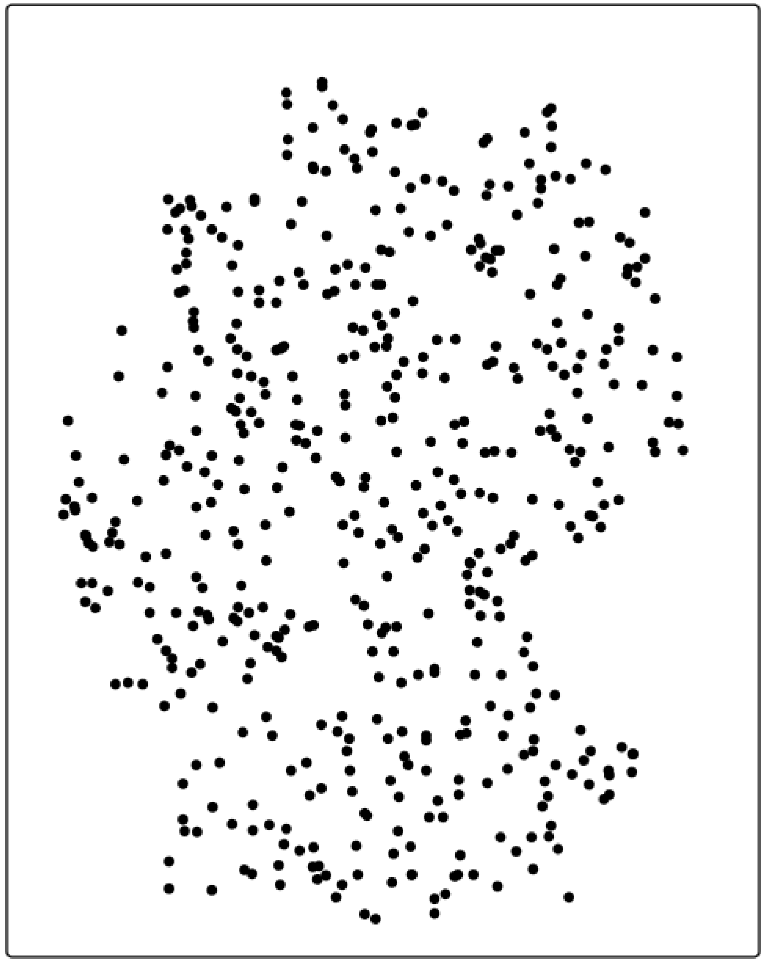
We can predict everywhere

Reality:

We validate where we have data, but predict where we do not.

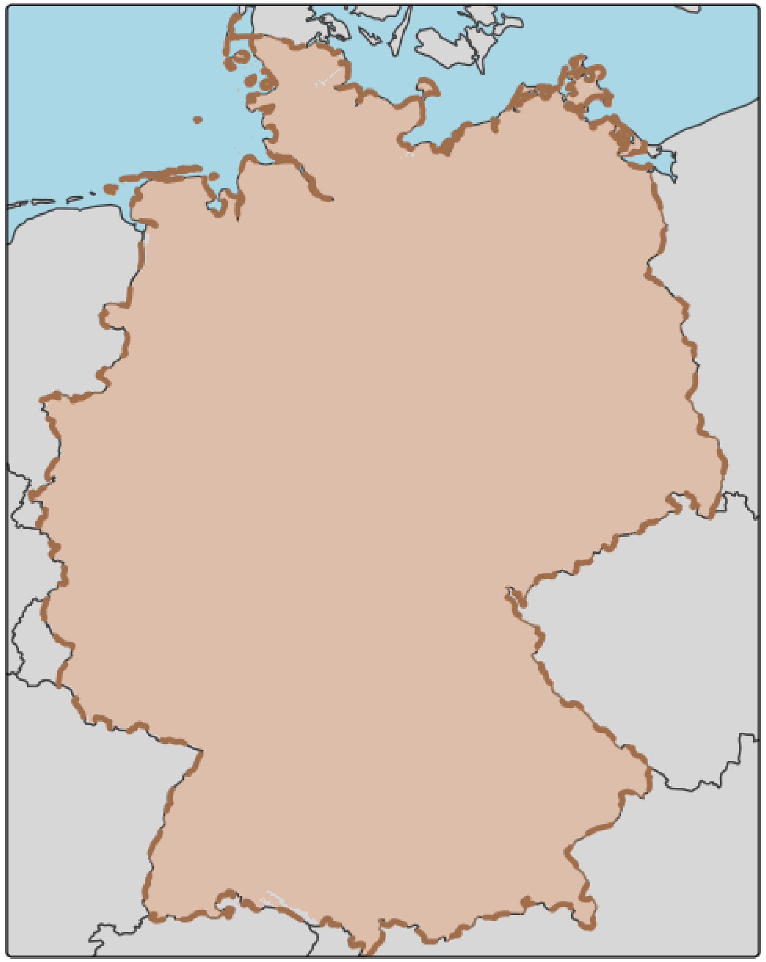
Similar application domain

We have this:



Sampling locations

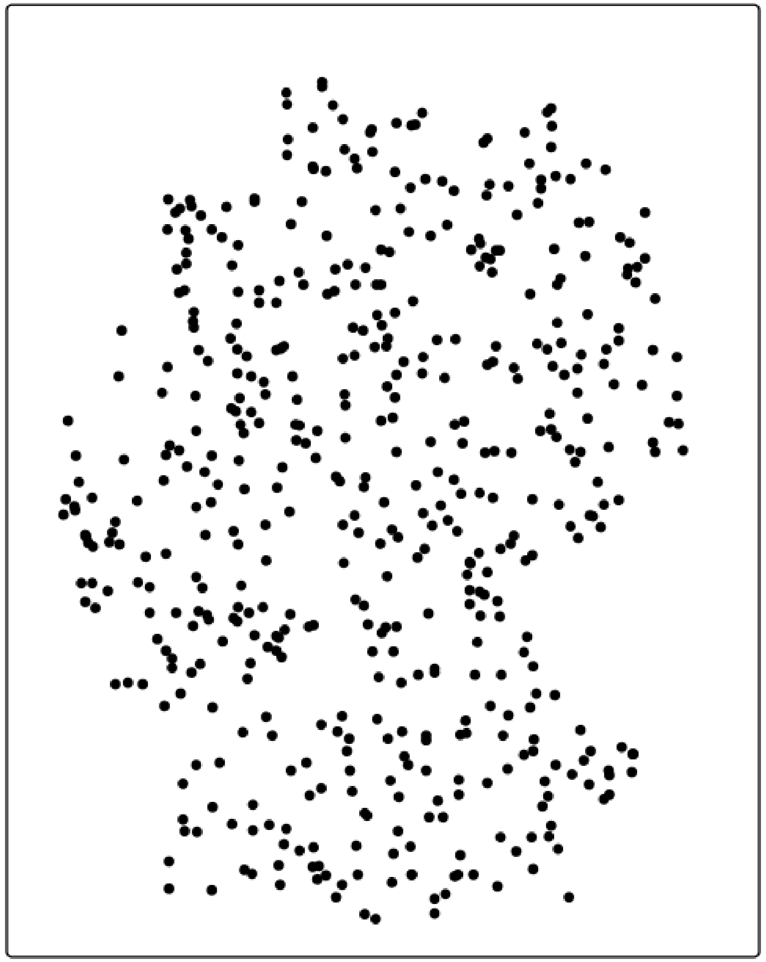
We want to predict here:



Prediction domain

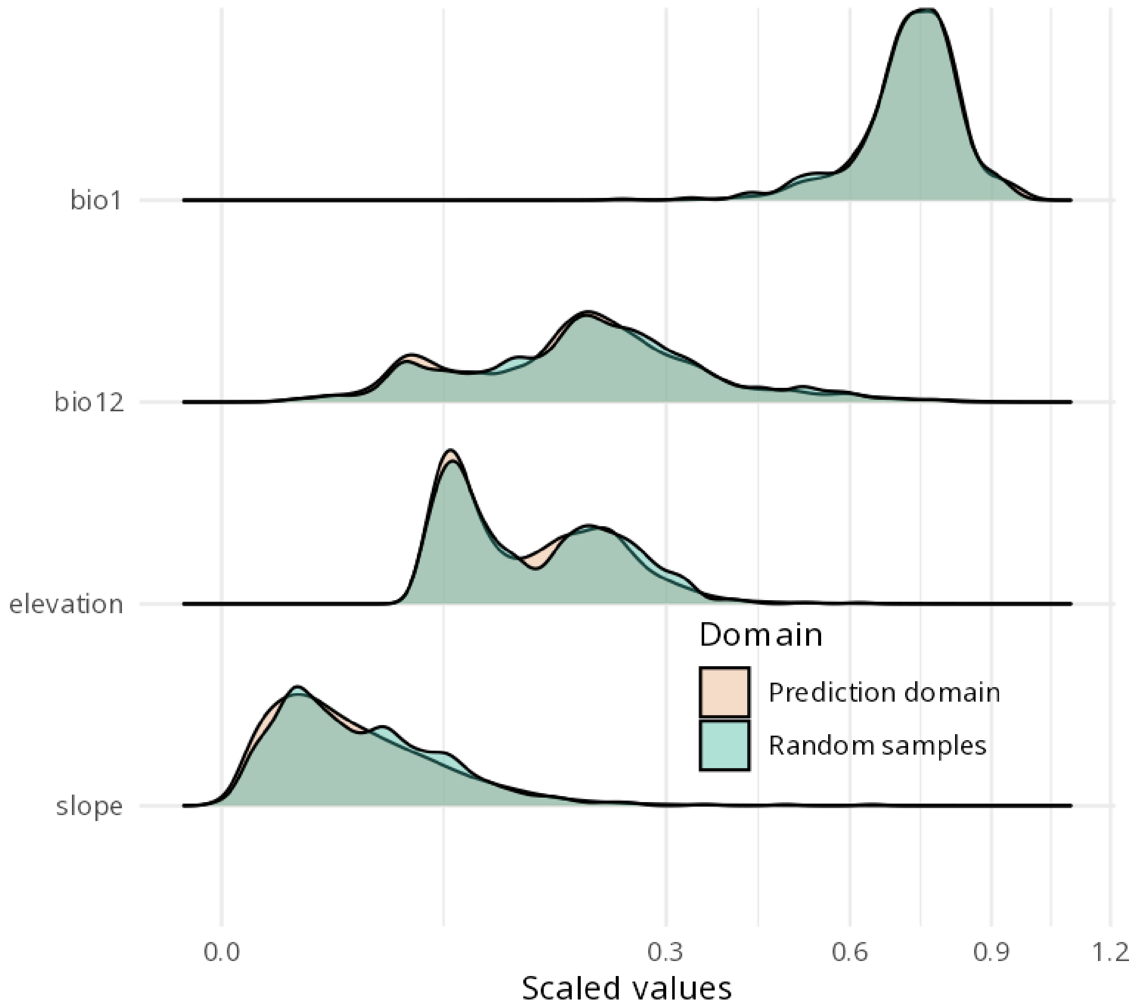
Similar application domain

We have this:



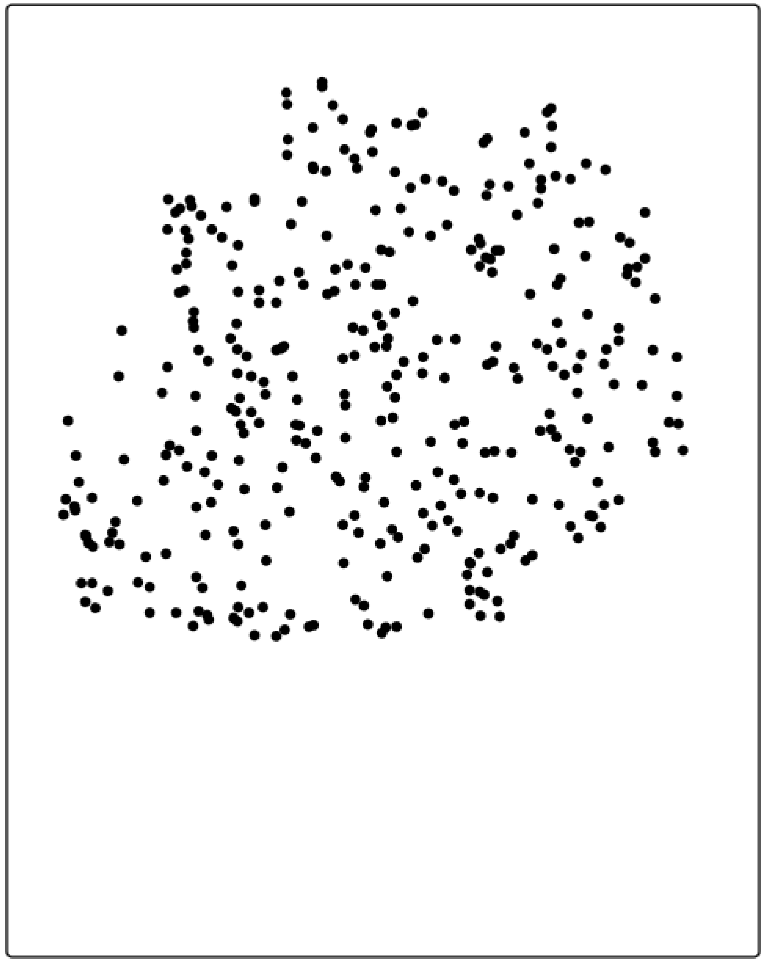
Sampling locations

Our predictor distributions are similar here:



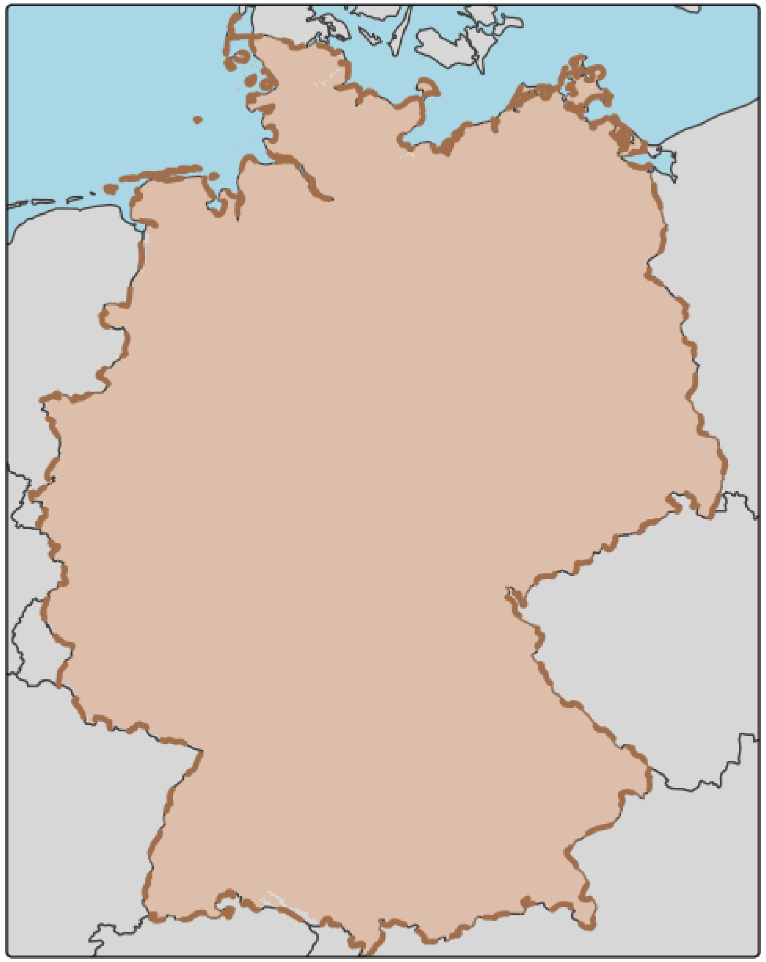
Different application domain

We have this:



Sampling locations

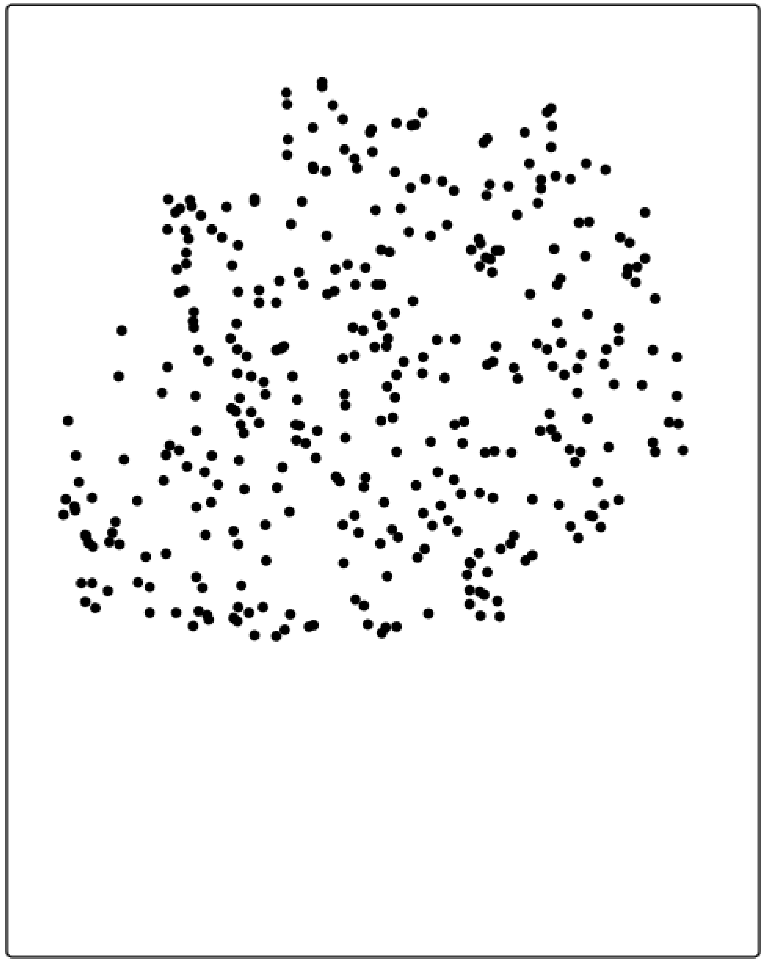
We want to predict here:



Prediction domain

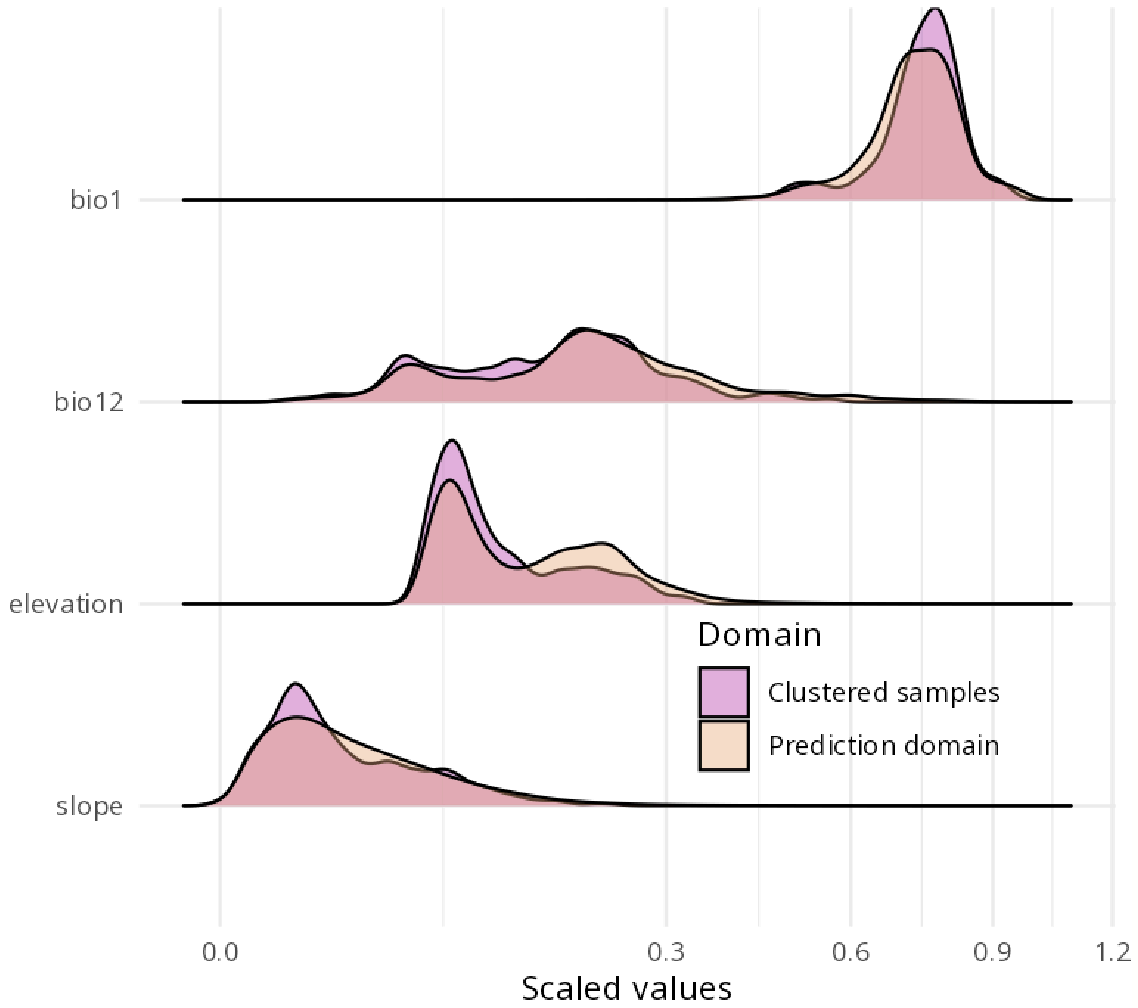
Different application domain

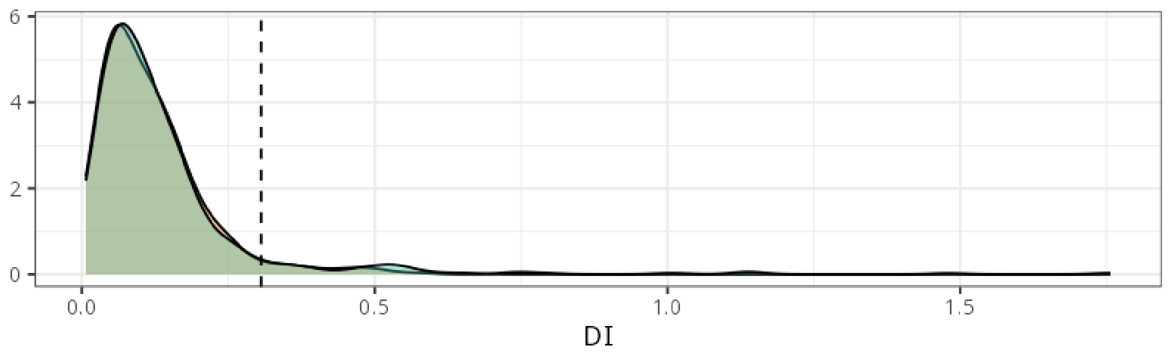
We have this:



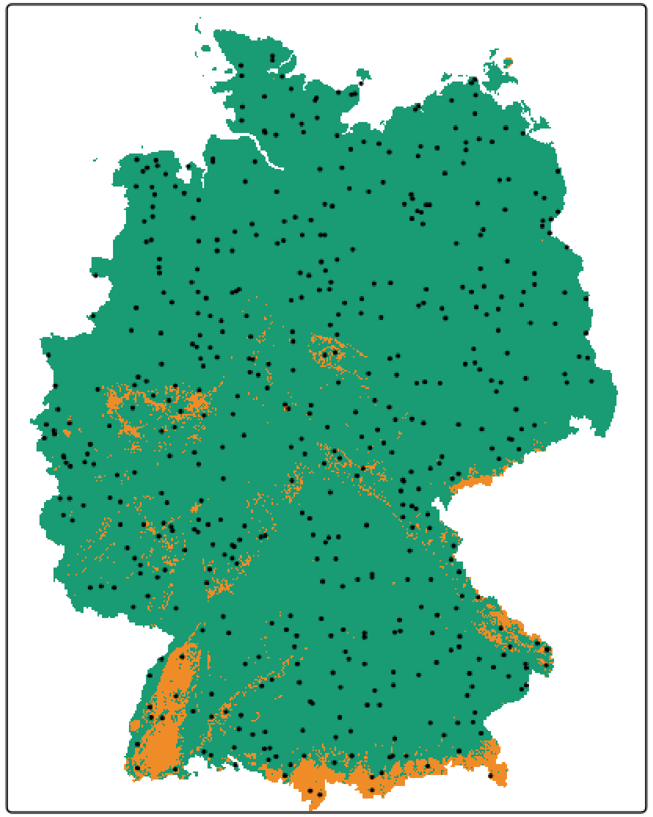
Sampling locations

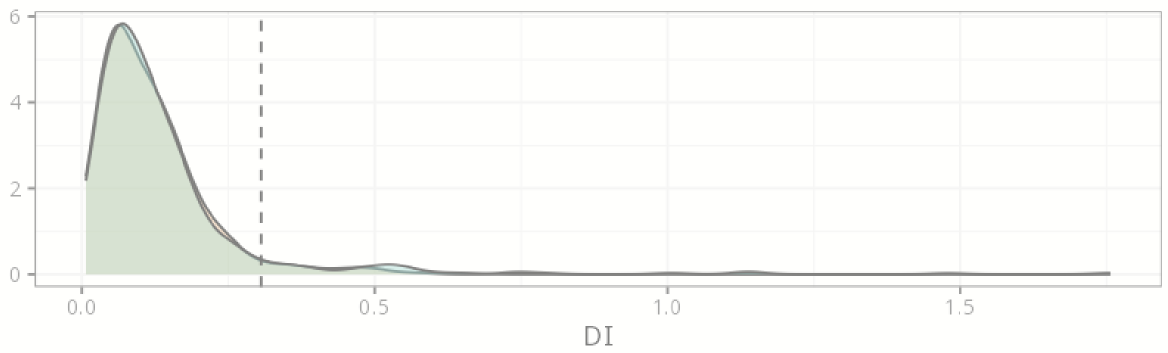
Our predictor distributions are a bit different here:



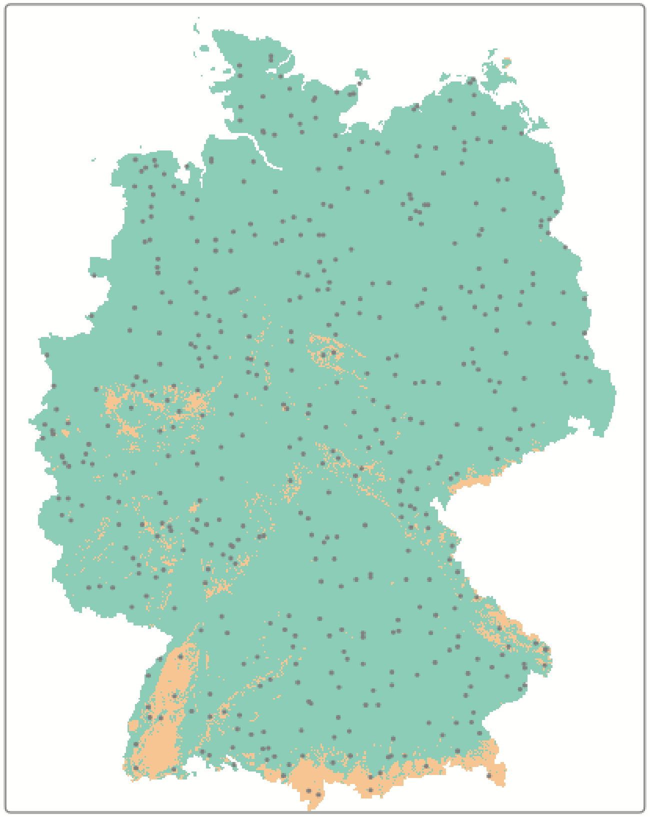


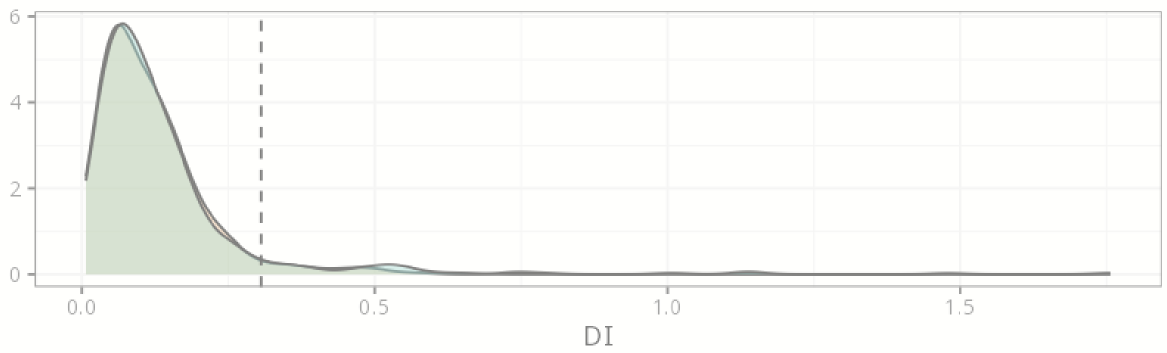
Random samples (93.9% inside AOA)



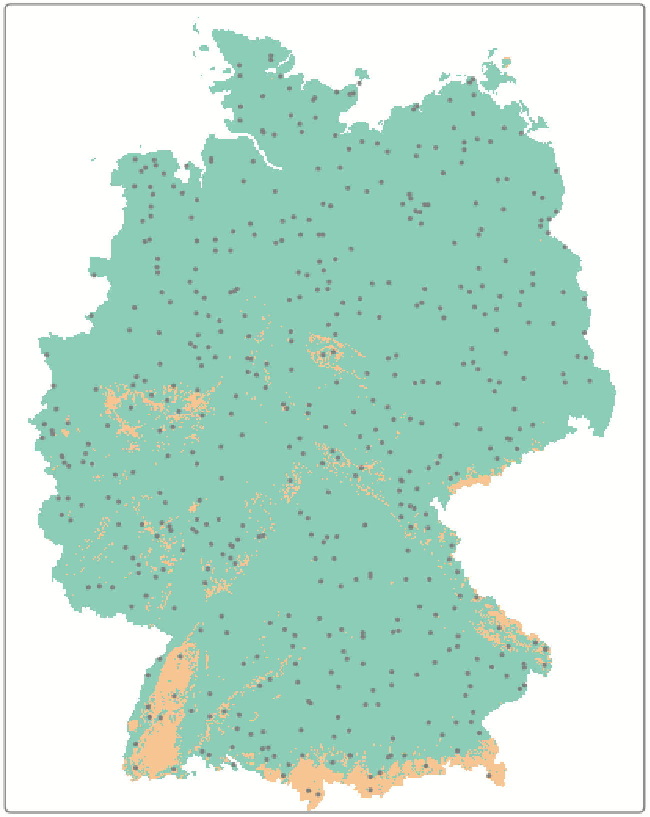


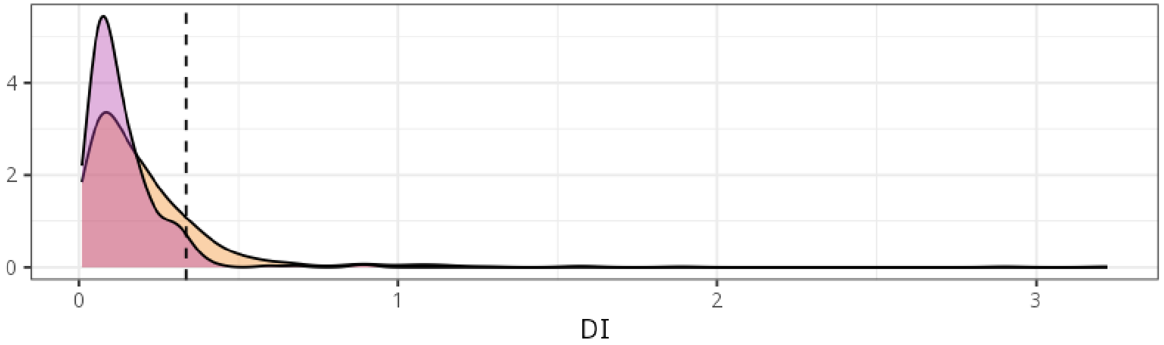
Random samples (93.9% inside AOA)



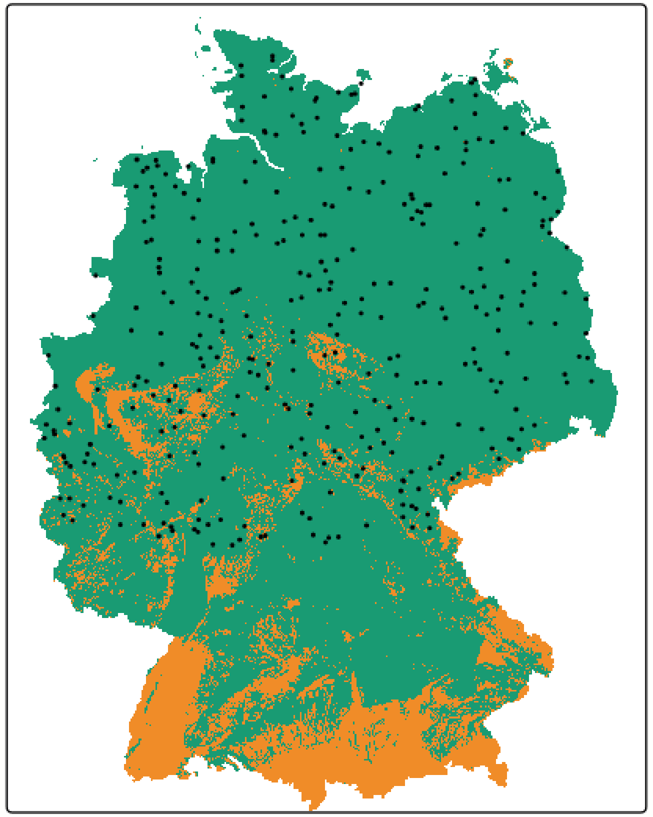


Random samples (93.9% inside AOA)





Partitioned samples (83.9% inside AOA)



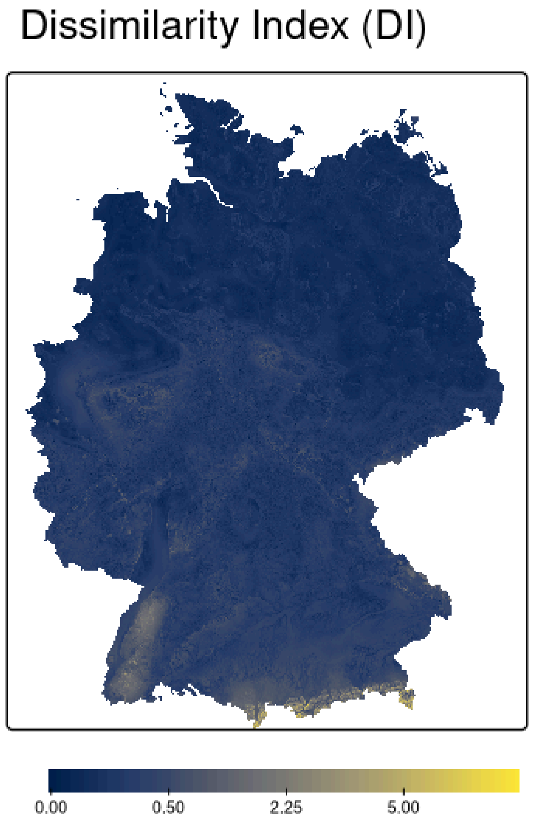
Area of applicability

Identify areas where the environment is not well represented, making predictions less trustworthy (Area of Applicability – AoA, Meyer and Pebesma, 2021); also local point density (LPD, Schumacher et al., 2025)

See also: MESS, Elith et al., 2010, Jung et al., 2020, and Velazco et al., 2023 for other related approaches

Area of applicability

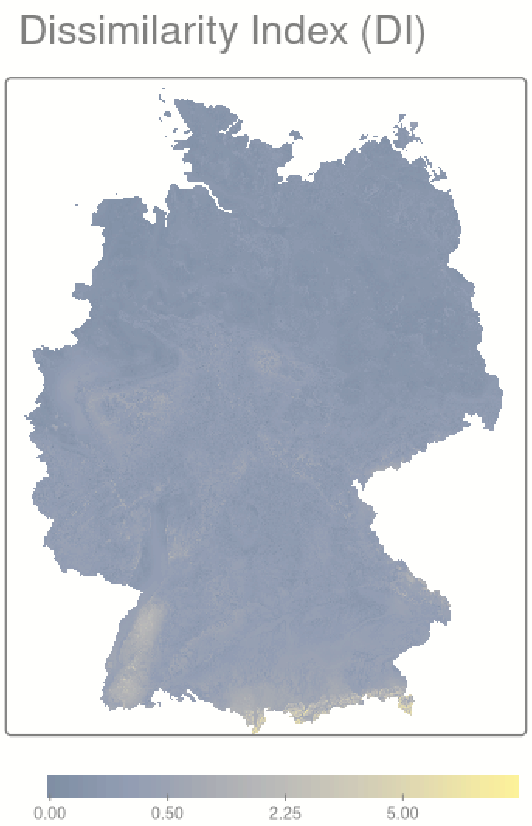
Identify areas where the environment is not well represented, making predictions less trustworthy (Area of Applicability – AoA, Meyer and Pebesma, 2021); also local point density (LPD, Schumacher et al., 2025)

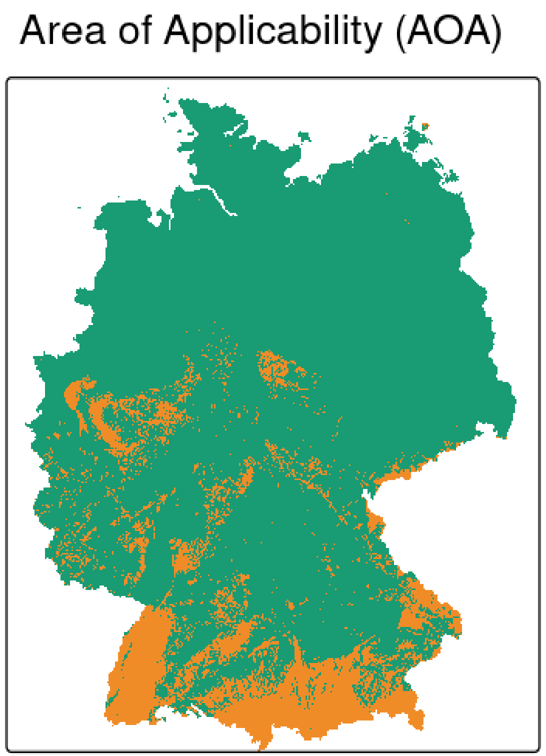


See also: MESS, Elith et al., 2010, Jung et al., 2020, and Velazco et al., 2023 for other related approaches

Area of applicability

Identify areas where the environment is not well represented, making predictions less trustworthy (Area of Applicability – AoA, Meyer and Pebesma, 2021); also local point density (LPD, Schumacher et al., 2025)





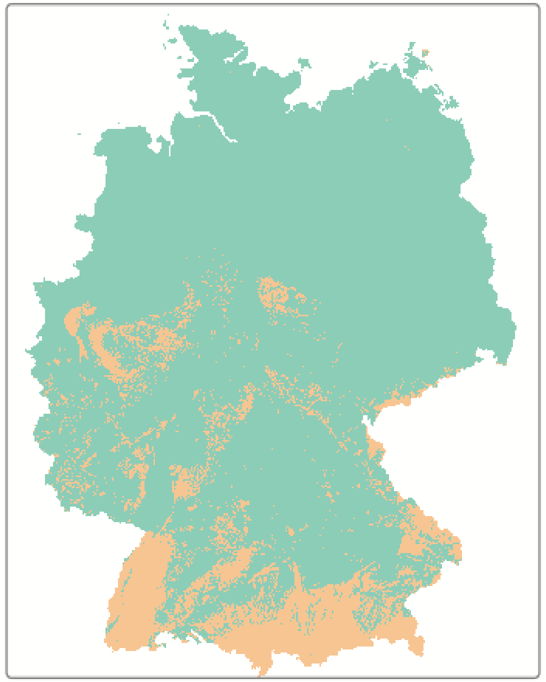
See also: MESS, Elith et al., 2010, Jung et al., 2020, and Velazco et al., 2023 for other related approaches

Area of applicability

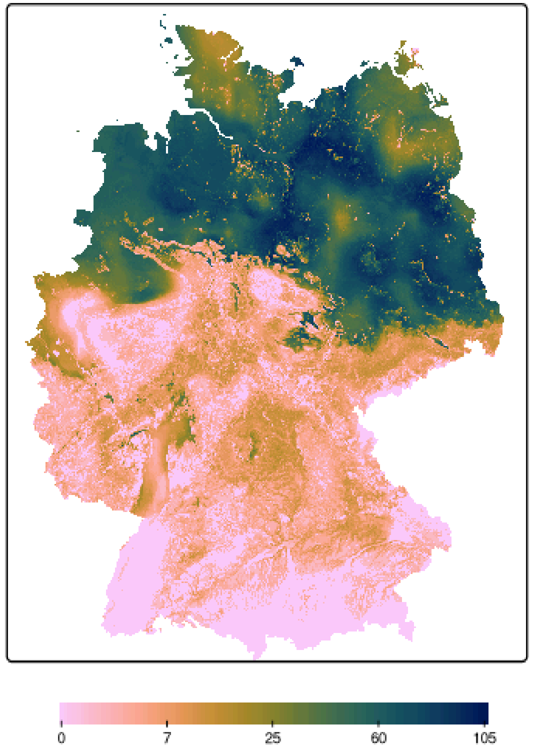
Identify areas where the environment is not well represented, making predictions less trustworthy (Area of Applicability – AoA, Meyer and Pebesma, 2021); also local point density (LPD, Schumacher et al., 2025)

Dissimilarity Index (DI)

Area of Applicability (AOA)



Local Point Density (LPD)



See also: MESS, Elith et al., 2010, Jung et al., 2020, and Velazco et al., 2023 for other related approaches

ASSUMPTION #2

Assumptions we make

We can predict everywhere

Reality:

We validate where we have data, but predict where we do not.

There is one “correct” validation approach

Assumptions we make

We can predict everywhere

Reality:

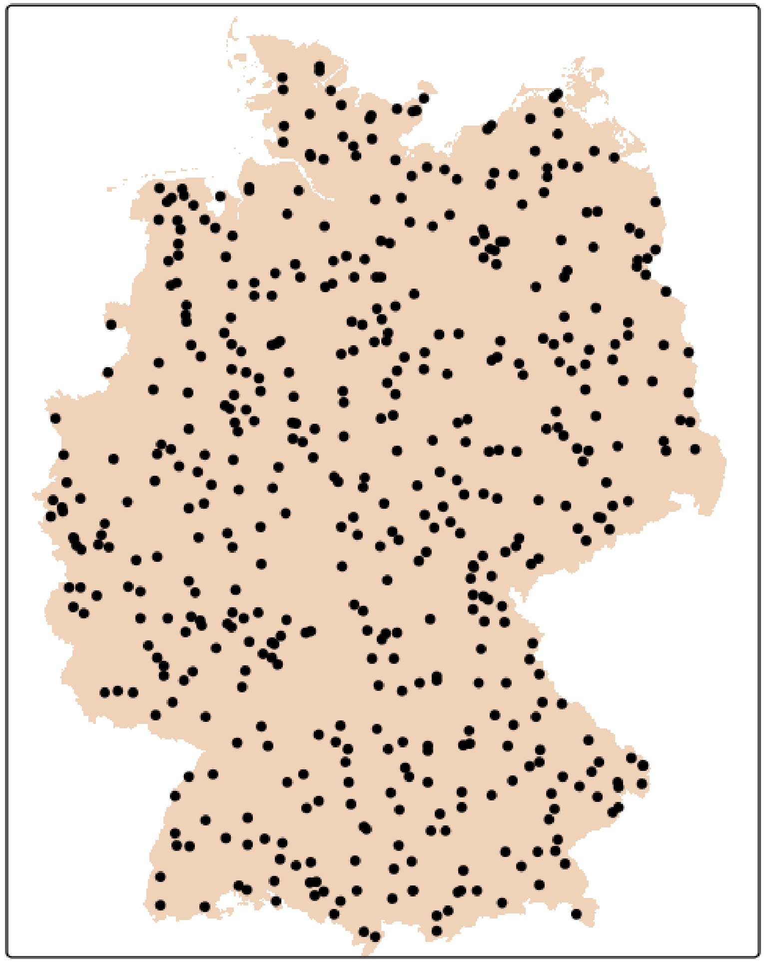
We validate where we have data, but predict where we do not.

There is one “correct” validation approach

Reality:

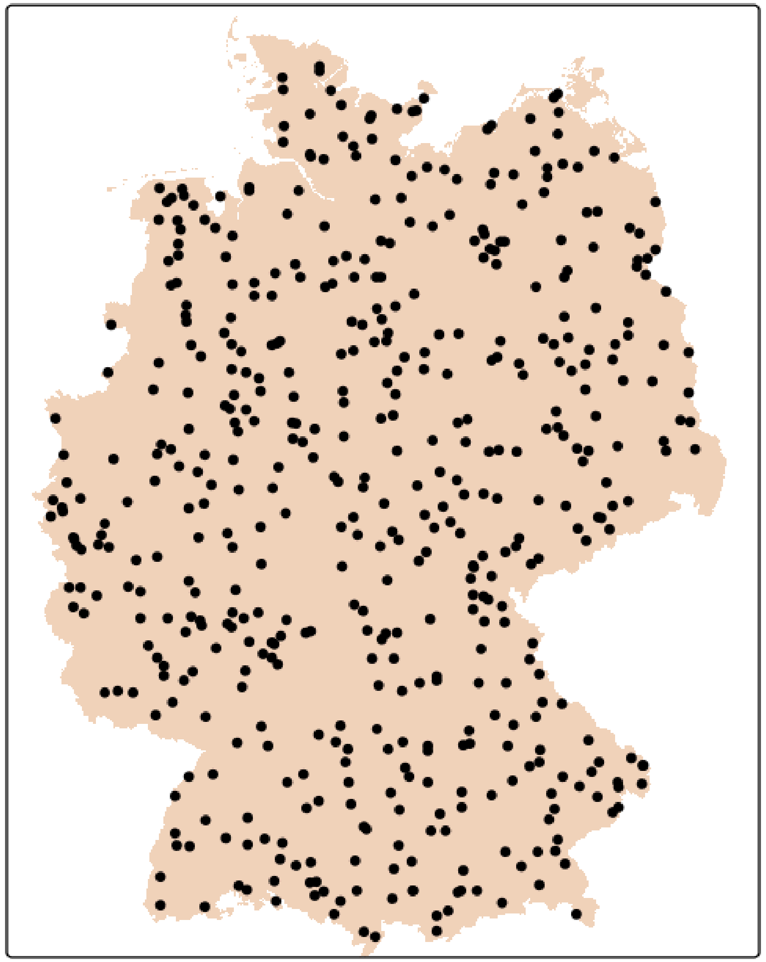
The validation strategy should follow the prediction task.

Prediction difficulty depends on prediction domain

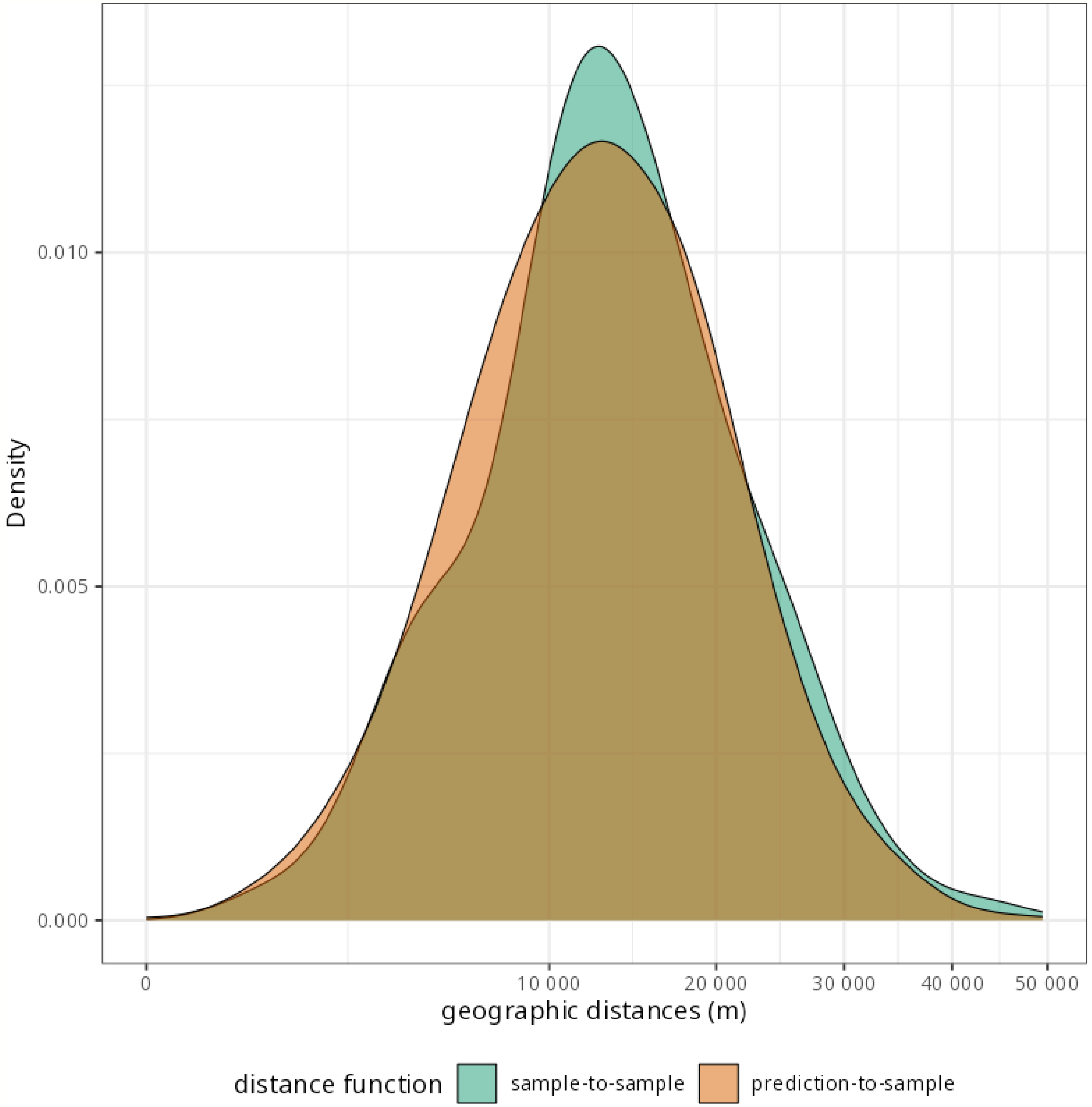


Random sampling locations

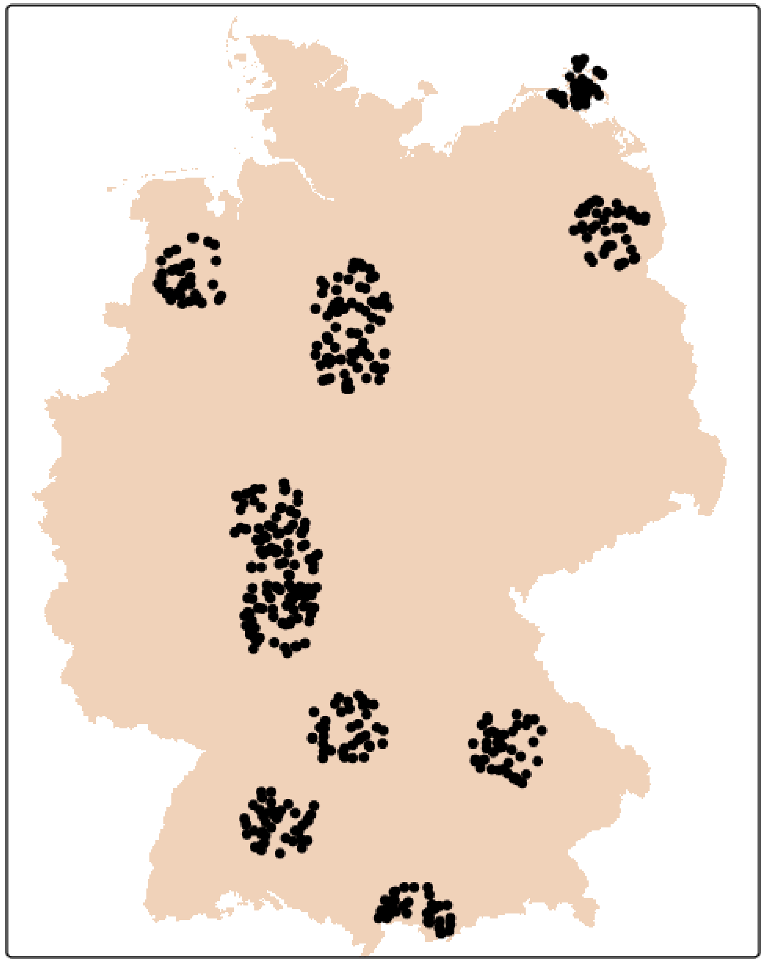
Prediction difficulty depends on prediction domain



Random sampling locations

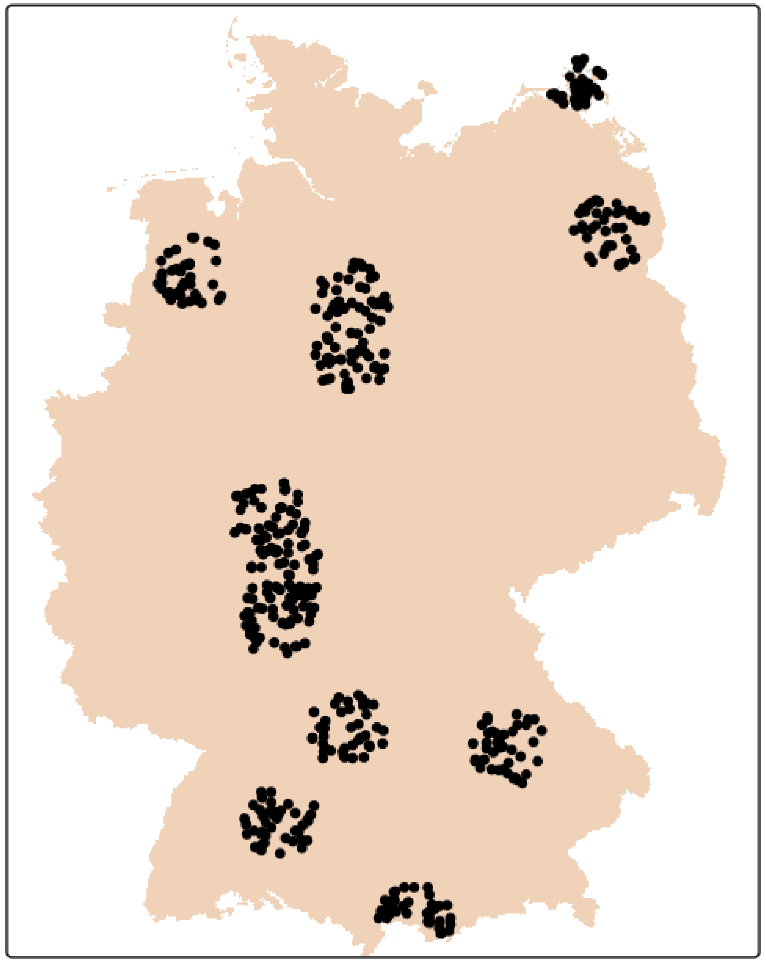


Prediction difficulty depends on prediction domain

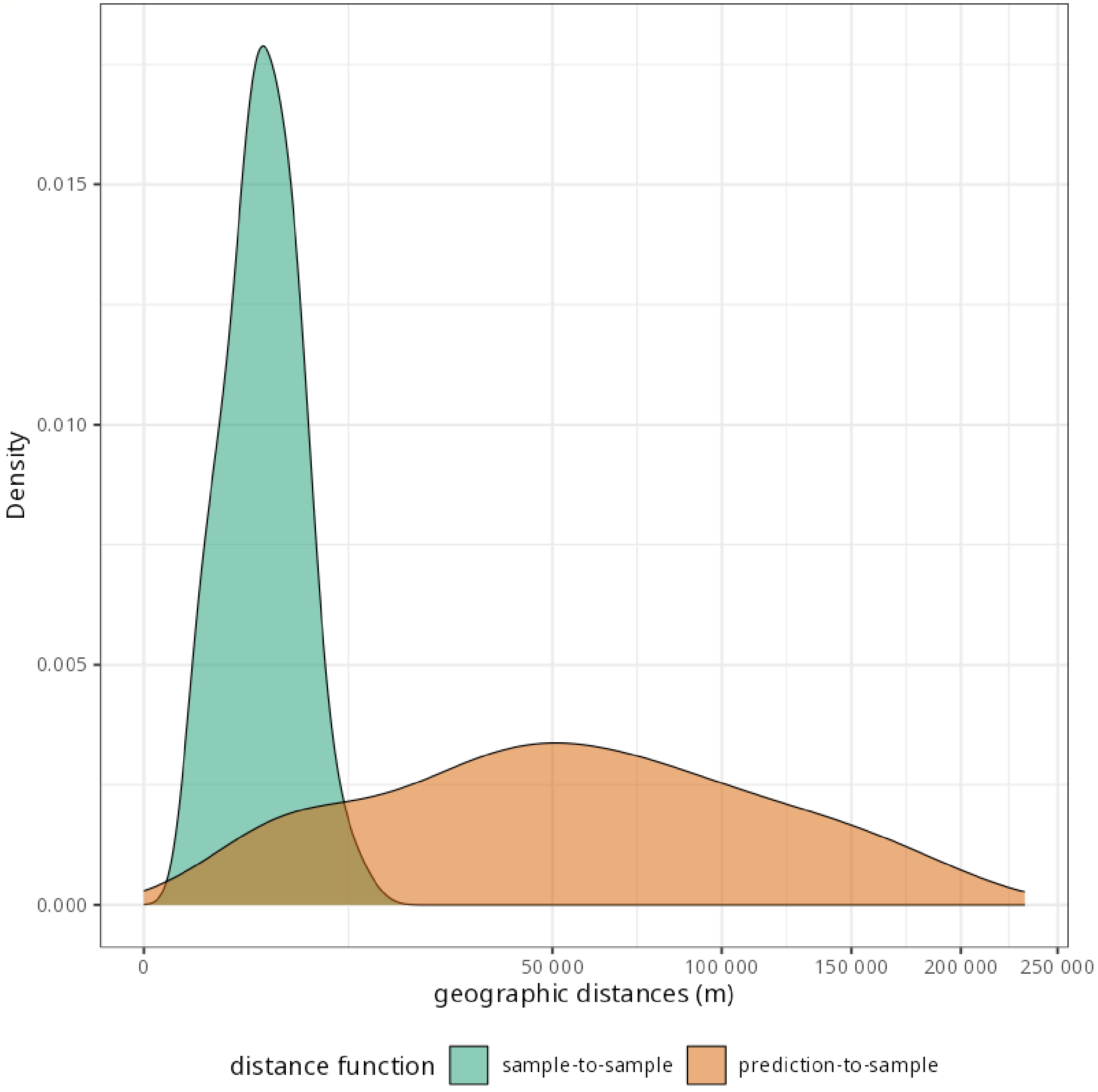


Clustered sampling locations

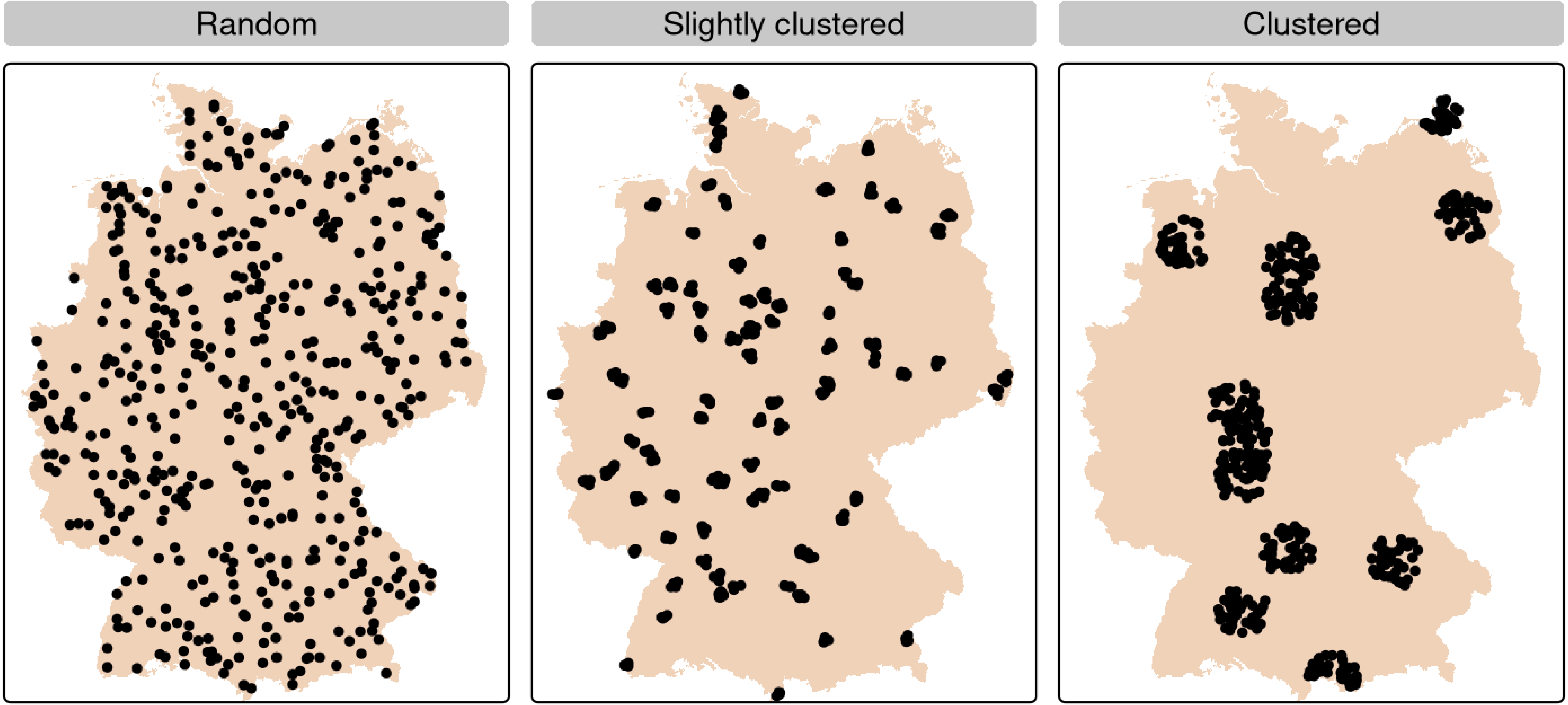
Prediction difficulty depends on prediction domain



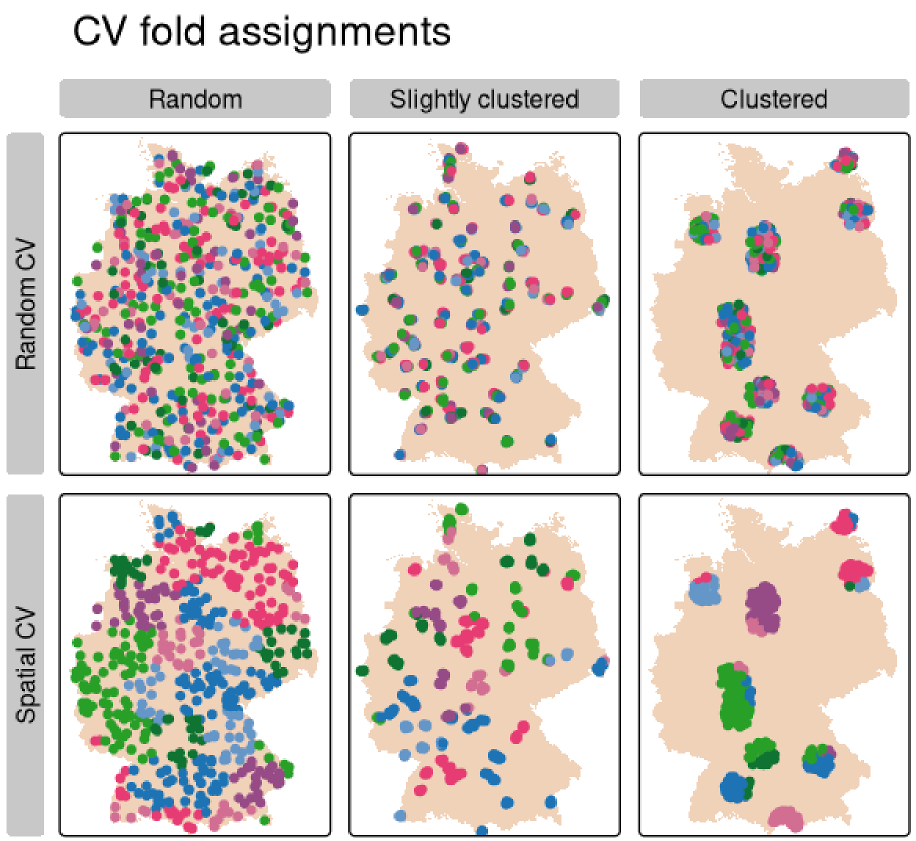
Clustered sampling locations



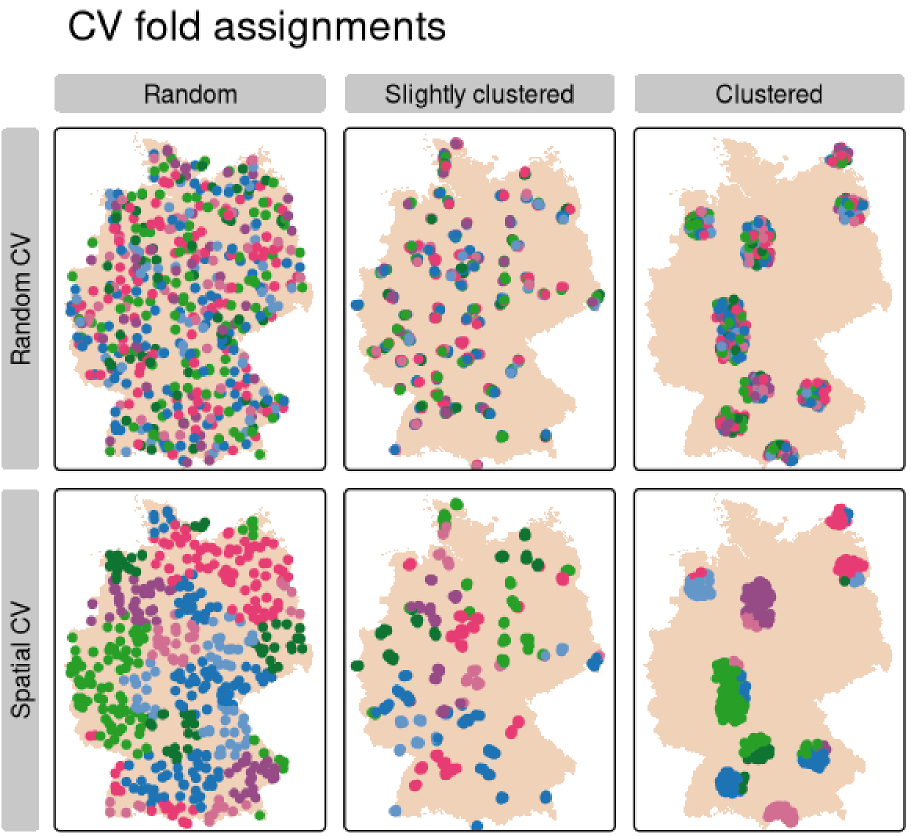
Extrapolation continuum

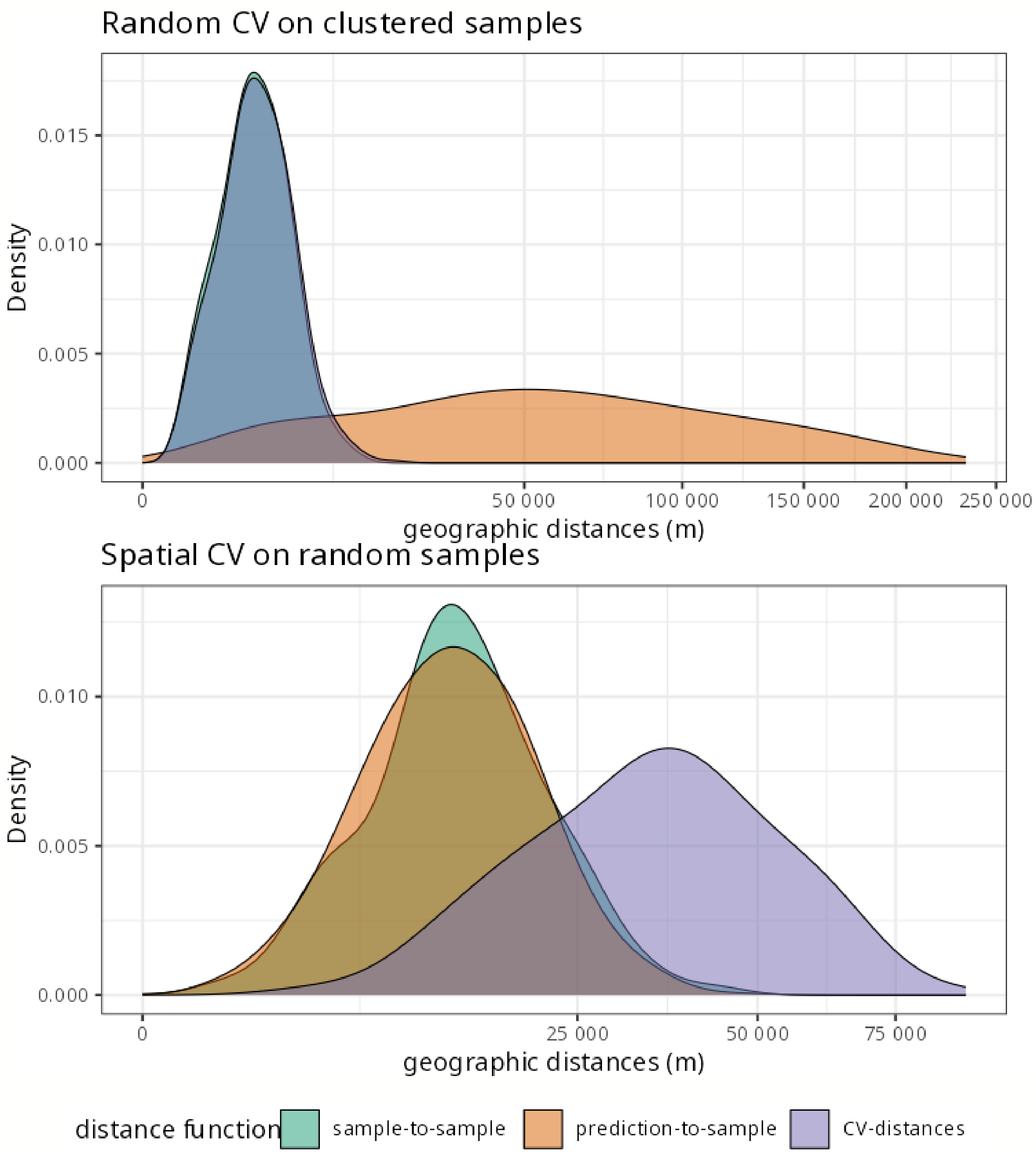


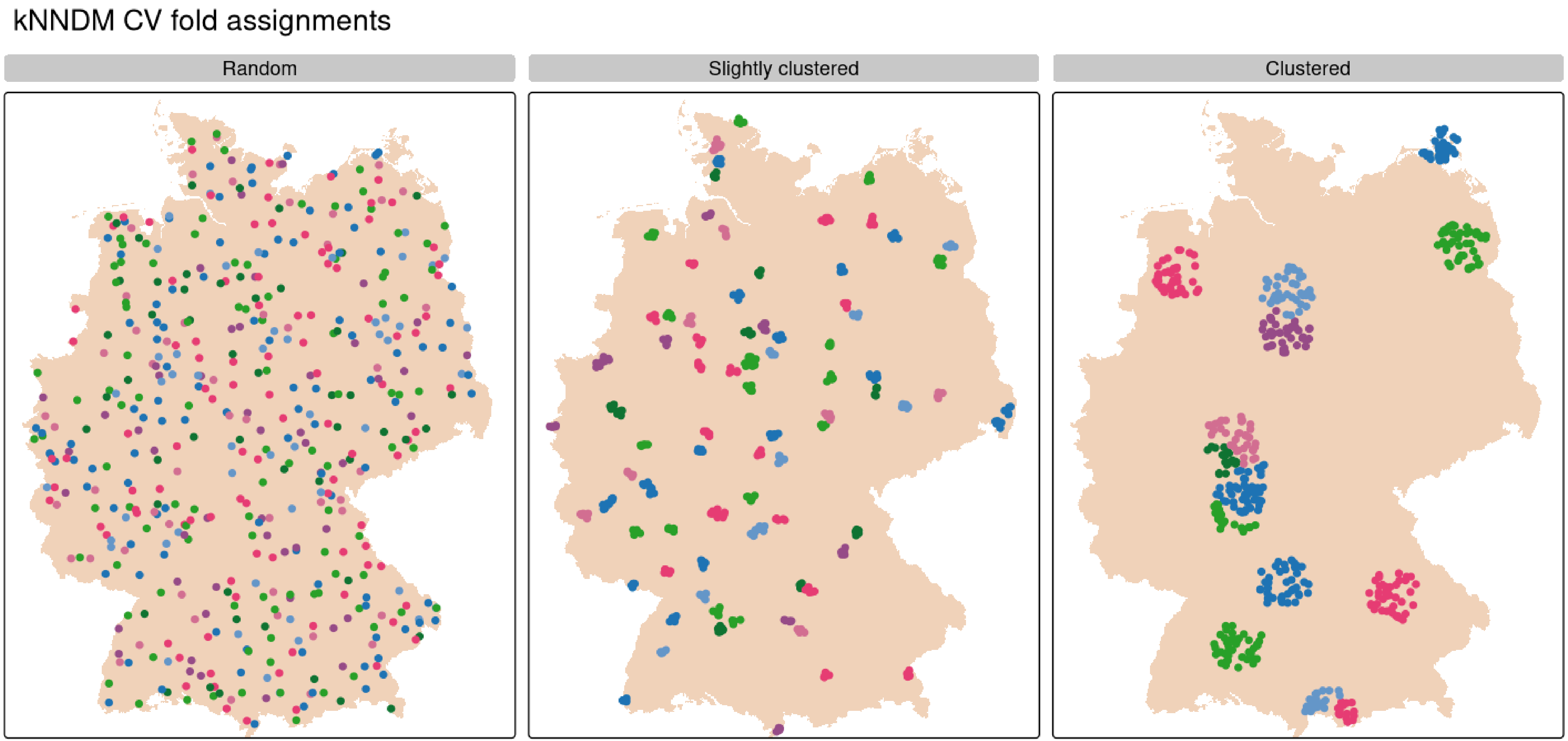
Specific evaluation strategy



Specific evaluation strategy

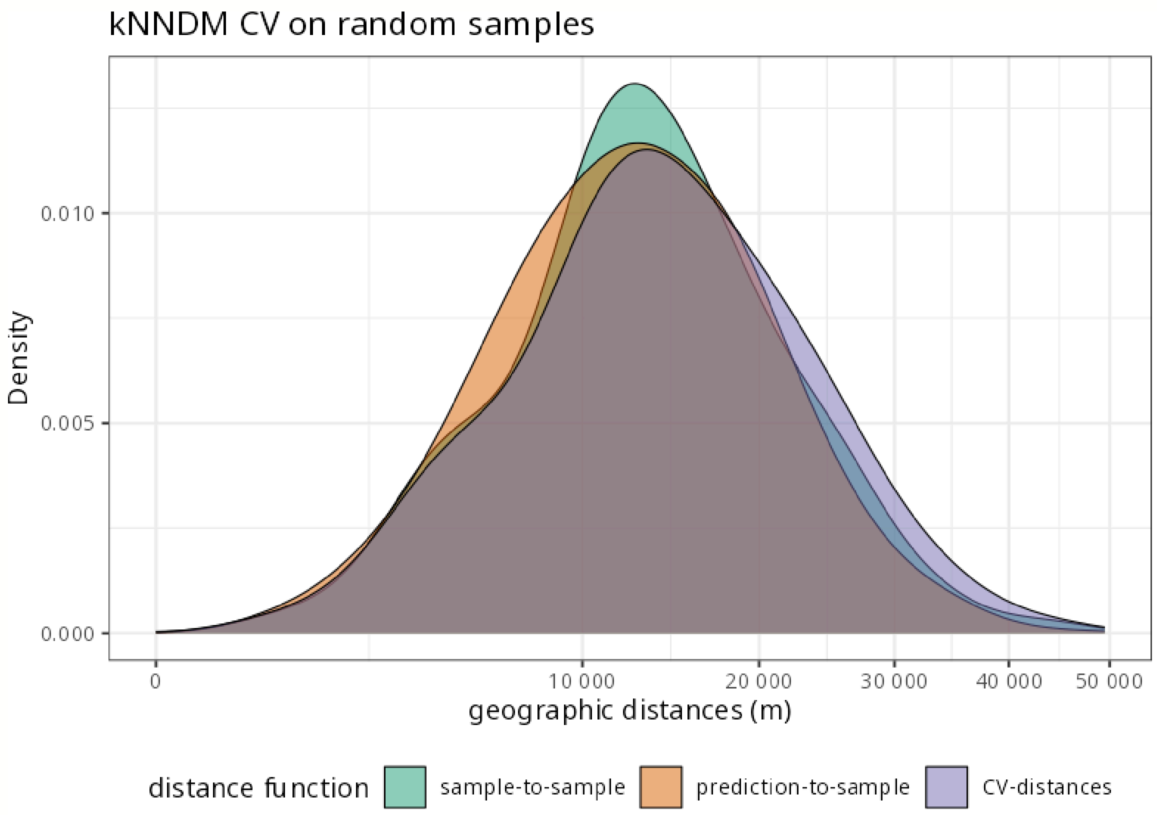





kNNDM CV fold assignments
Random
Slightly clustered
Clustered

Adaptive evaluation (kNNDM)

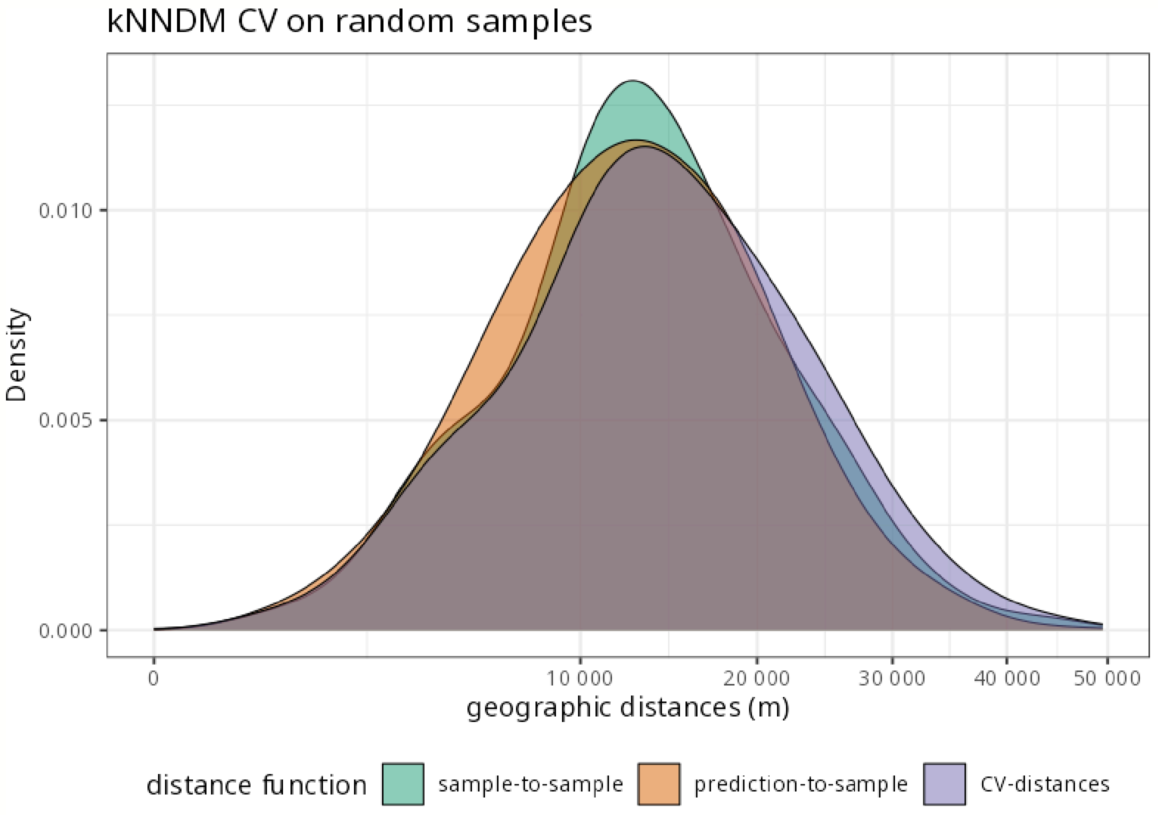
k-Nearest Neighbor Distance Matching ([kNNDM, Linnenbrink et al., 2024]) matches folds to the prediction scenario using distance structure (either in geographic or predictor space).

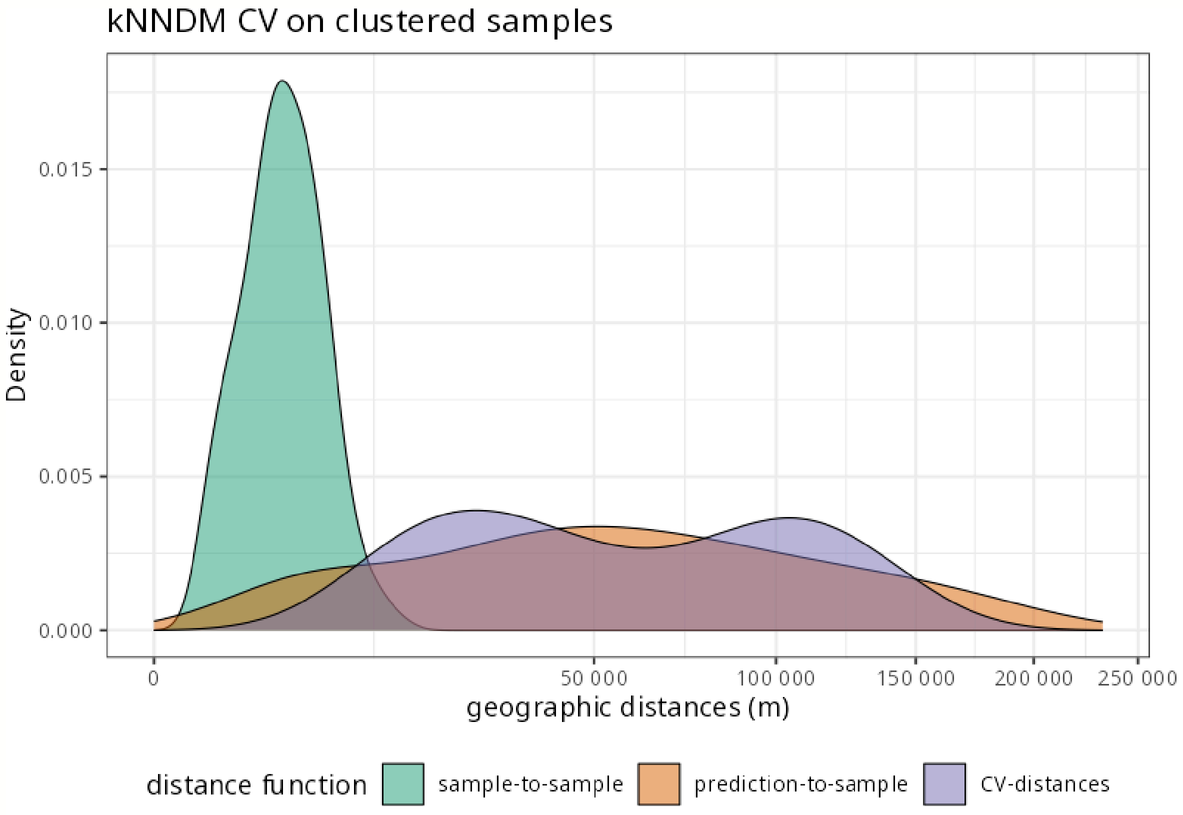


See also: NNDM, Milà et al. (2022) and DA-CV, Wang et al. (2025) for other related approaches

Adaptive evaluation (kNNDM)

k-Nearest Neighbor Distance Matching ([kNNDM, Linnenbrink et al., 2024]) matches folds to the prediction scenario using distance structure (either in geographic or predictor space).





See also: NNDM, Milà et al. (2022) and DA-CV, Wang et al. (2025) for other related approaches

ASSUMPTION #3

Assumptions we make

We can predict everywhere

Reality:

We validate where we have data, but predict where we do not.

There is one "correct" validation approach

Reality:

The validation strategy should follow the prediction task.

All validation points are equal

Assumptions we make

We can predict everywhere

Reality:

We validate where we have data, but predict where we do not.

There is one “correct” validation approach

Reality:

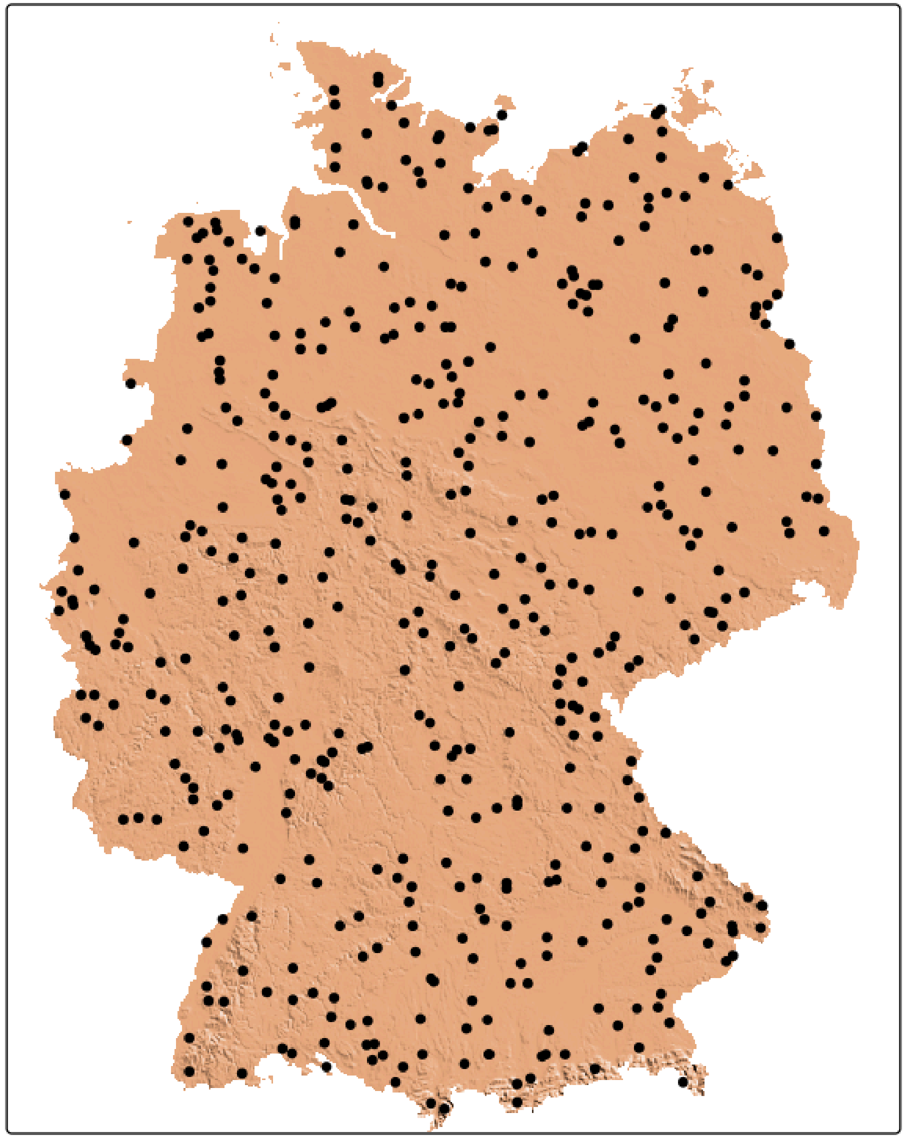
The validation strategy should follow the prediction task.

All validation points are equal

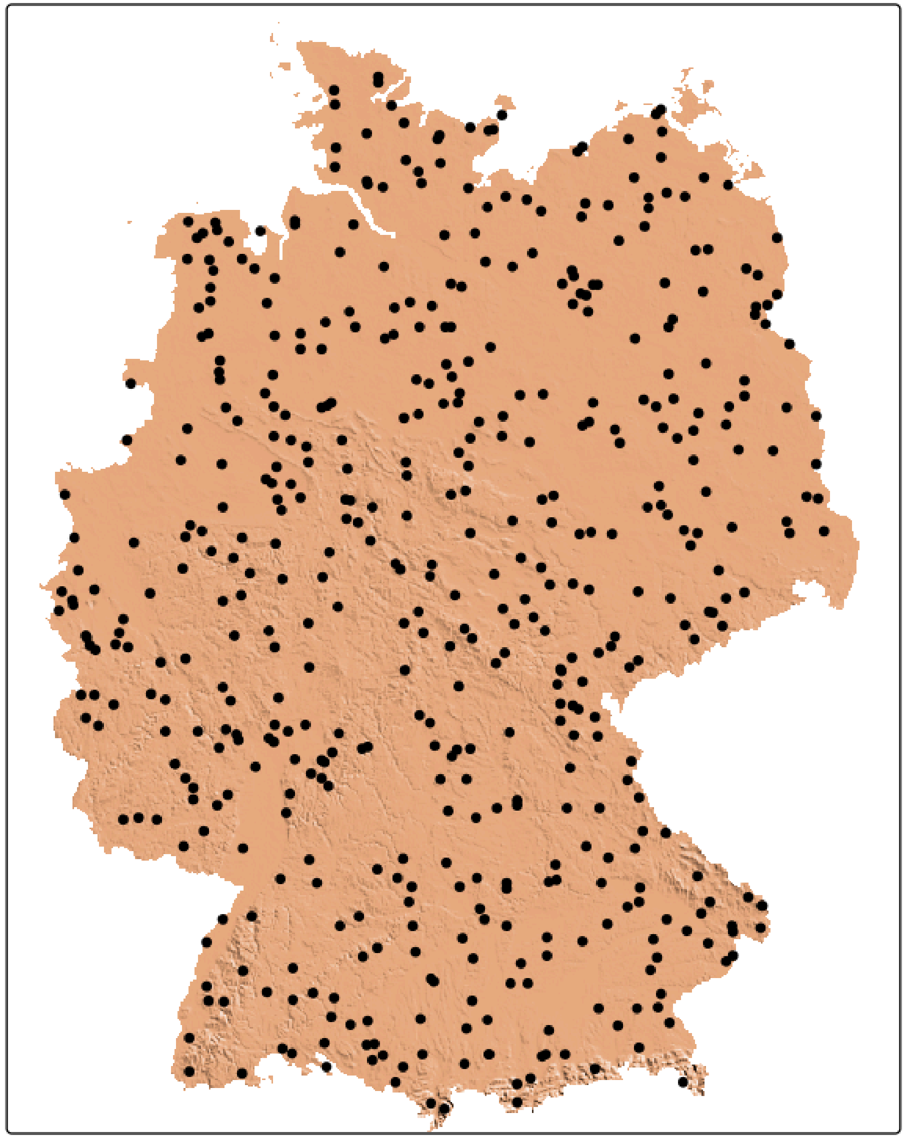
Reality:

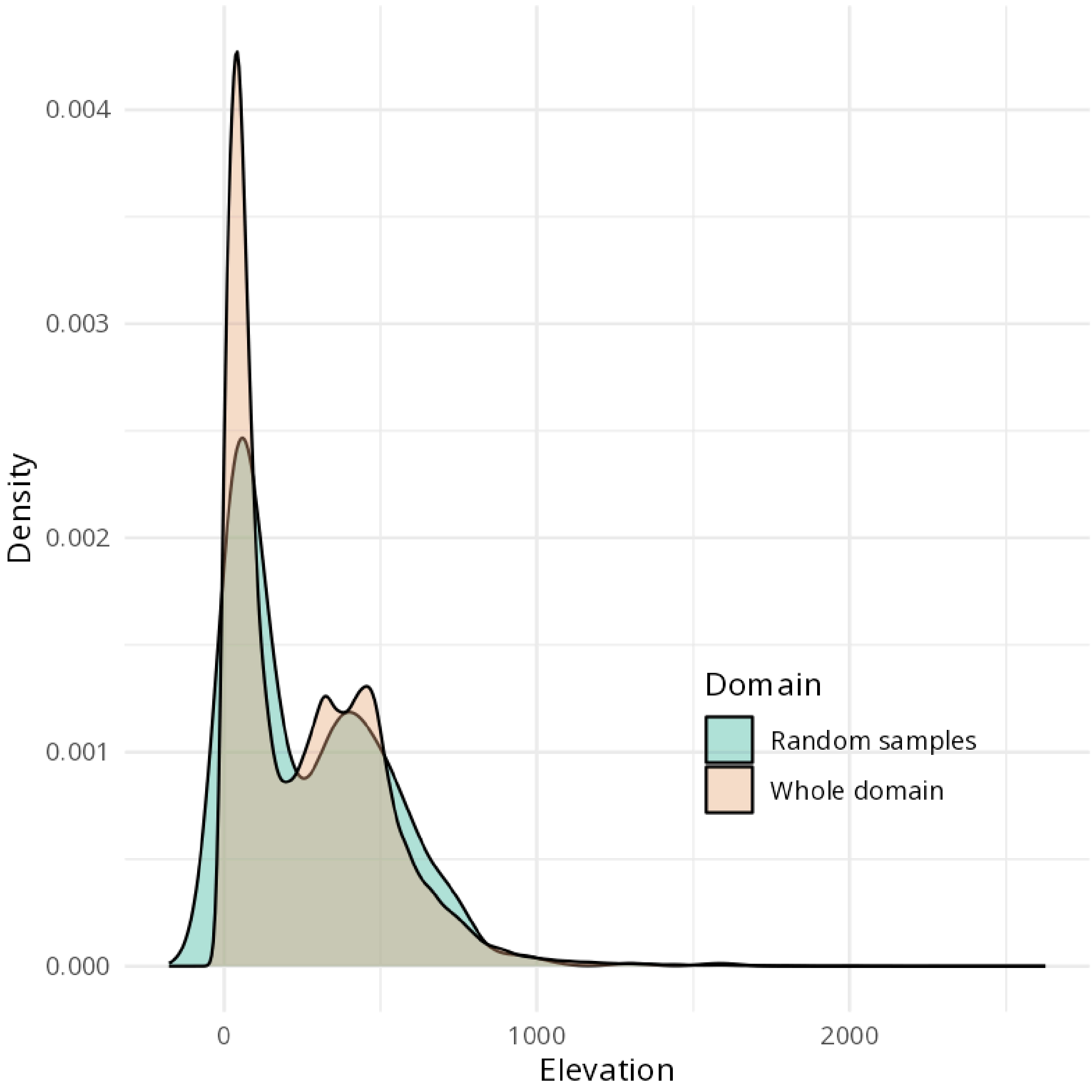
Prediction conditions are not equally common.

Overlapping predictor distribution(s)

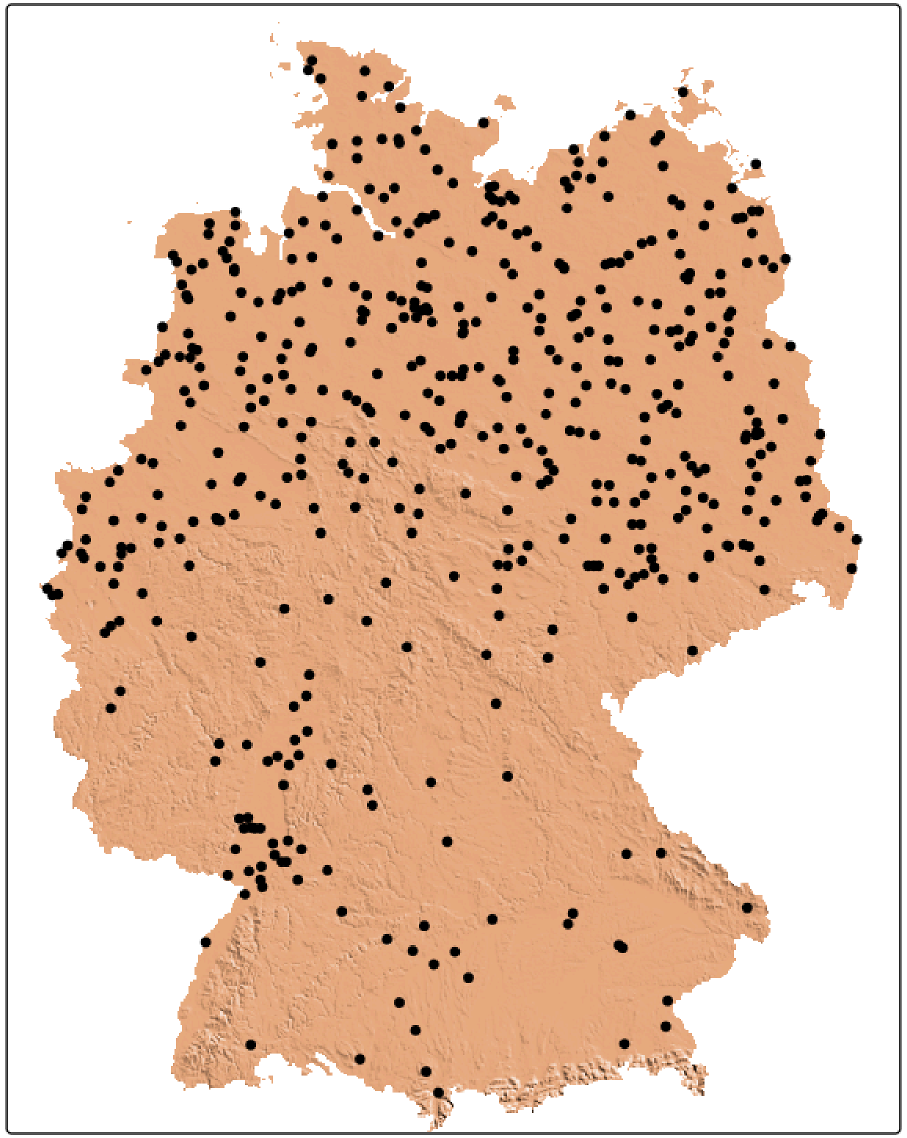


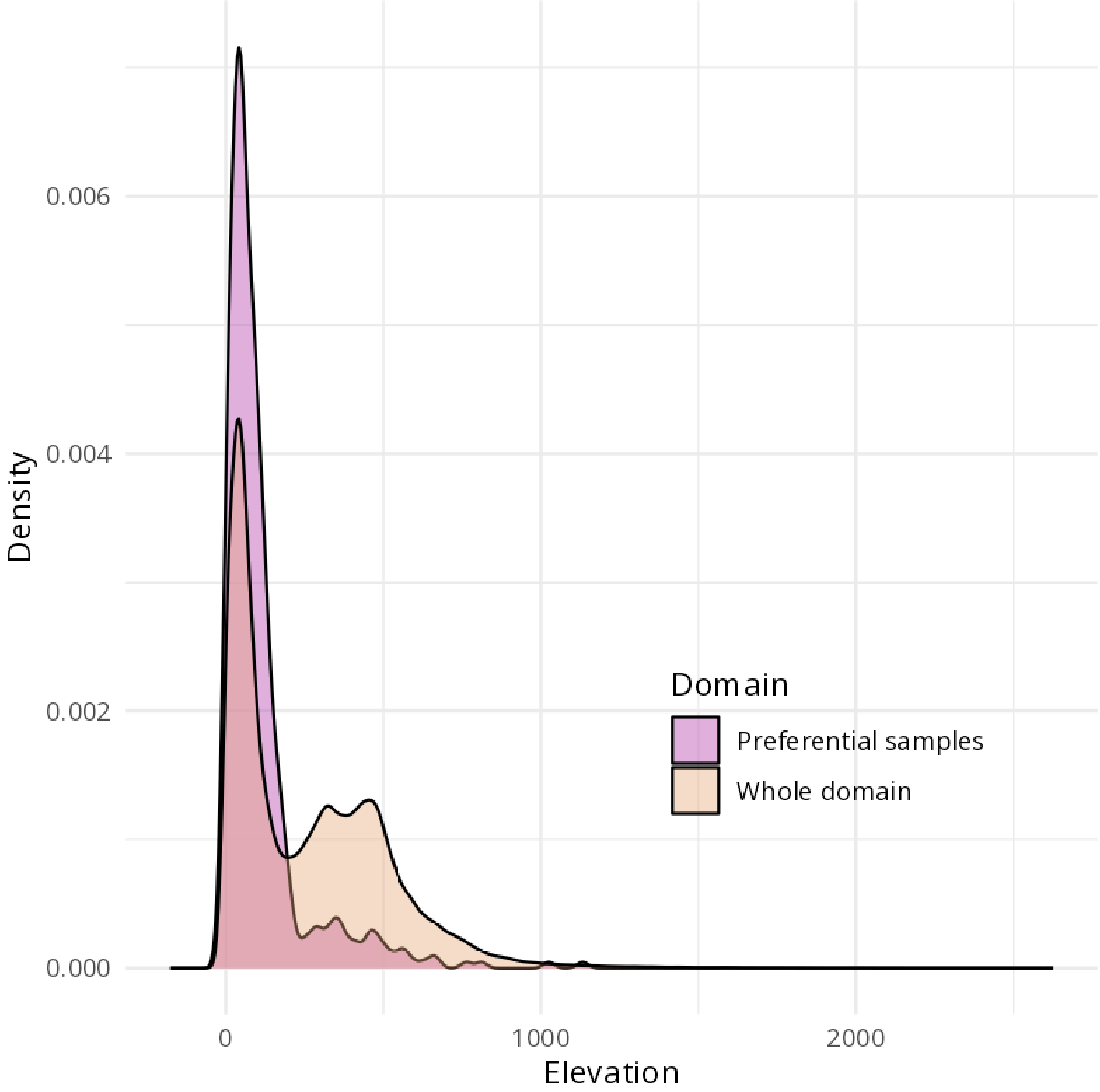
Overlapping predictor distribution(s)





Partially overlapping predictor distribution(s)





Weighting validation points

Evaluation approach	Lowland-area weight (%)	Highland-area weight (%)	Overall RMSE
Germany domain (target distribution)	50	50	0.667

Weighting validation points

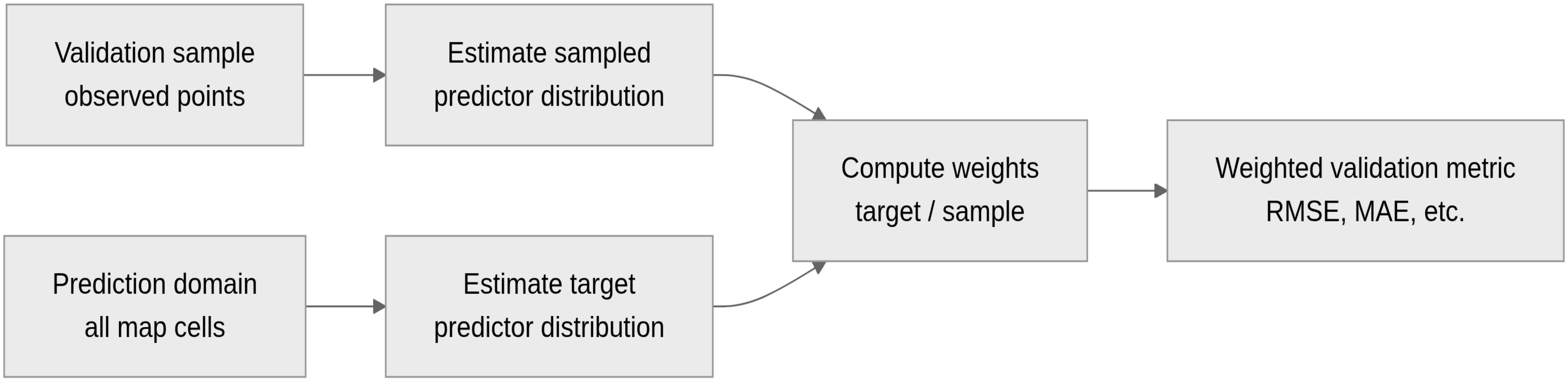
Evaluation approach	Lowland-area weight (%)	Highland-area weight (%)	Overall RMSE
Germany domain (target distribution)	50	50	0.667
Preferential sample (unweighted)	89	11	0.541

Weighting validation points

Evaluation approach	Lowland-area weight (%)	Highland-area weight (%)	Overall RMSE
Germany domain (target distribution)	50	50	0.667
Preferential sample (unweighted)	89	11	0.541
Preferential sample (reweighted)	50	50	0.667

Approaches to reweighting

Target-Weighted Cross-Validation ([TWCV, Brenning and Suesse, 2026]) adjusts cross-validation weights to align evaluation with the prediction domain rather than the sampled data distribution.



From assumptions to results

Assumptions we make

~~We can predict everywhere~~

Reality:

We validate where we have data, but predict where we do not.

~~There is one "correct" validation approach~~

Reality:

The validation strategy should follow the prediction task.

~~All validation points are equal~~

Reality:

Prediction conditions are not equally common.

Assumptions we make

~~We can predict everywhere~~

Reality:

We validate where we have data, but predict where we do not.

~~There is one “correct” validation approach~~

Reality:

The validation strategy should follow the prediction task.

~~All validation points are equal~~

Reality:

Prediction conditions are not equally common.

Possible solution: identify regions of reliable prediction.

Assumptions we make

~~We can predict everywhere~~

Reality:

We validate where we have data, but predict where we do not.

Possible solution: identify regions of reliable prediction.

~~There is one "correct" validation approach~~

Reality:

The validation strategy should follow the prediction task.

Possible solution: use adaptive evaluation strategies to create folds that resemble the prediction scenario.

~~All validation points are equal~~

Reality:

Prediction conditions are not equally common.

Assumptions we make

~~We can predict everywhere~~

Reality:

We validate where we have data, but predict where we do not.

Possible solution: identify regions of reliable prediction.

~~There is one “correct” validation approach~~

Reality:

The validation strategy should follow the prediction task.

Possible solution: use adaptive evaluation strategies to create folds that resemble the prediction scenario.

~~All validation points are equal~~

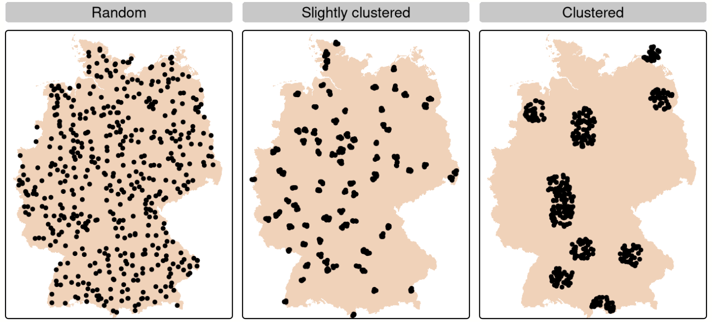
Reality:

Prediction conditions are not equally common.

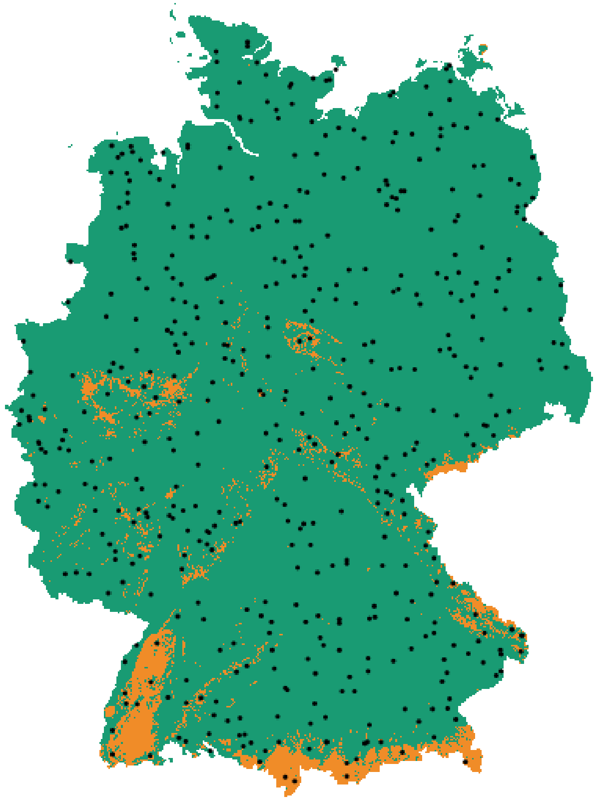
Possible solution: weight validation points according to their prevalence in the prediction area.

A piece of evidence

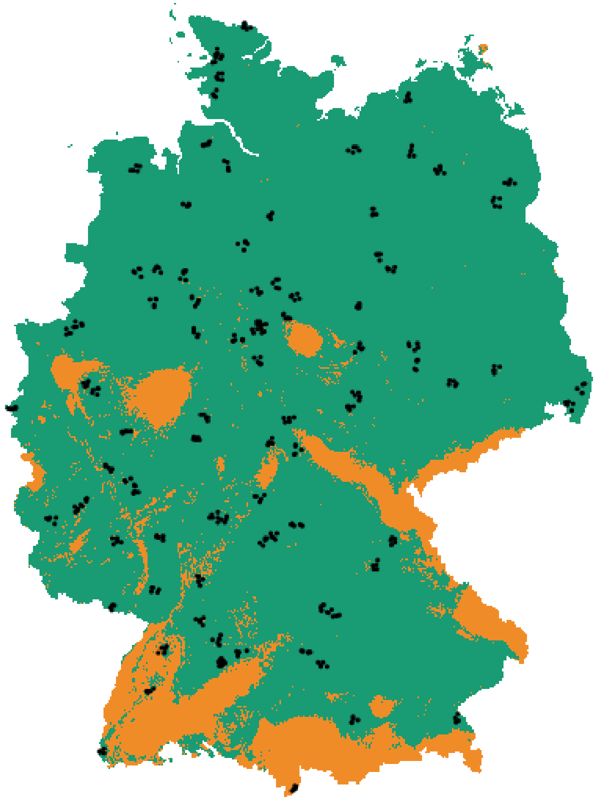
Area of applicability for different sampling designs



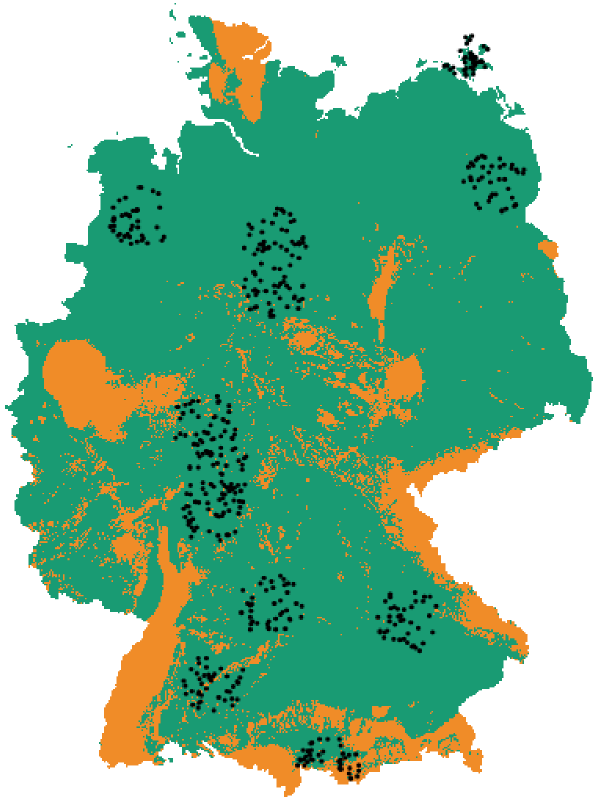
Random samples (93.9% inside AOA)



Slightly clustered samples (85.7% inside AOA)

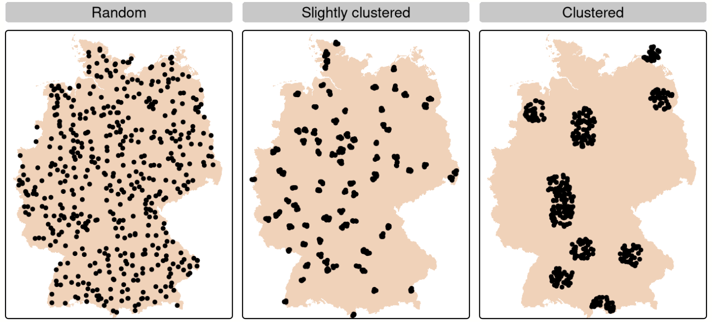


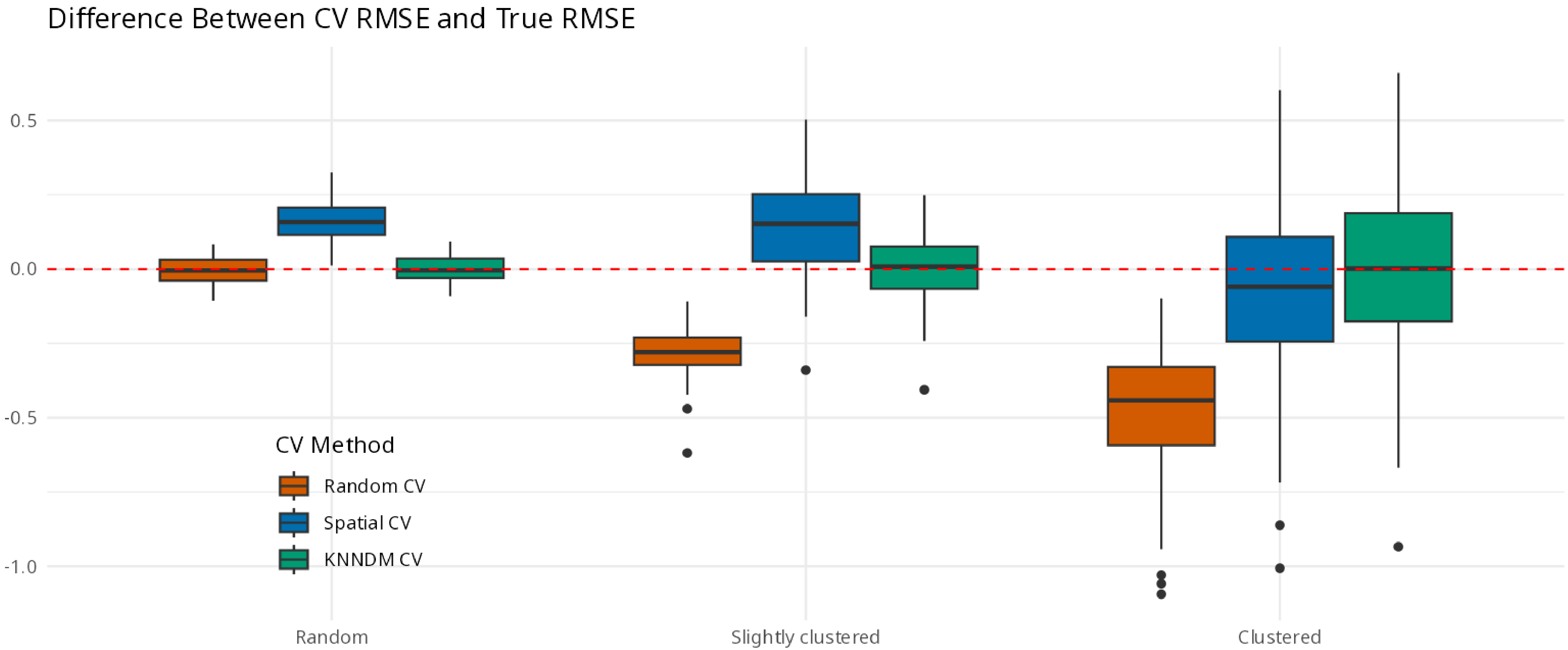
Clustered samples (81.8% inside AOA)



A piece of evidence

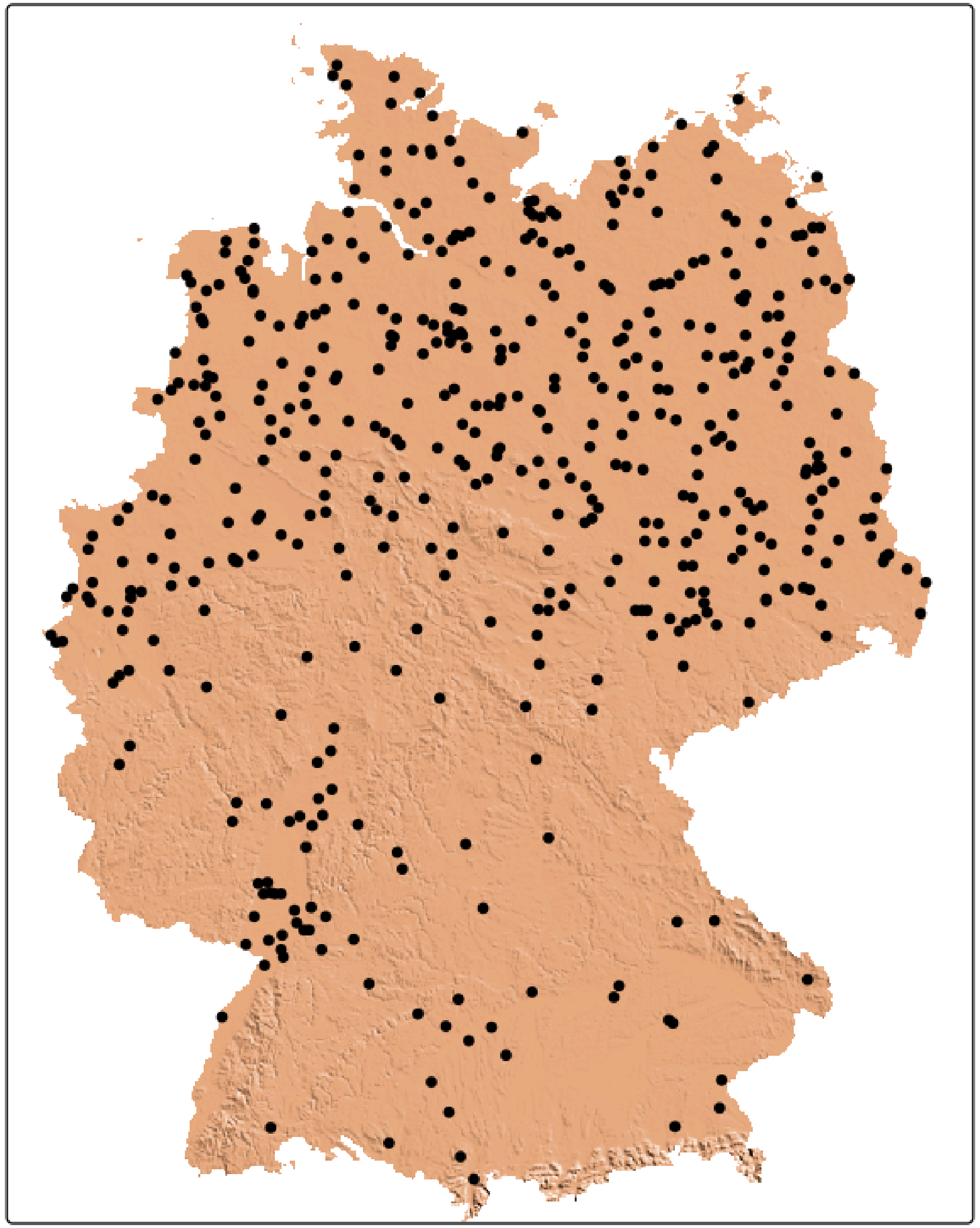
Evaluation results for different validation strategies

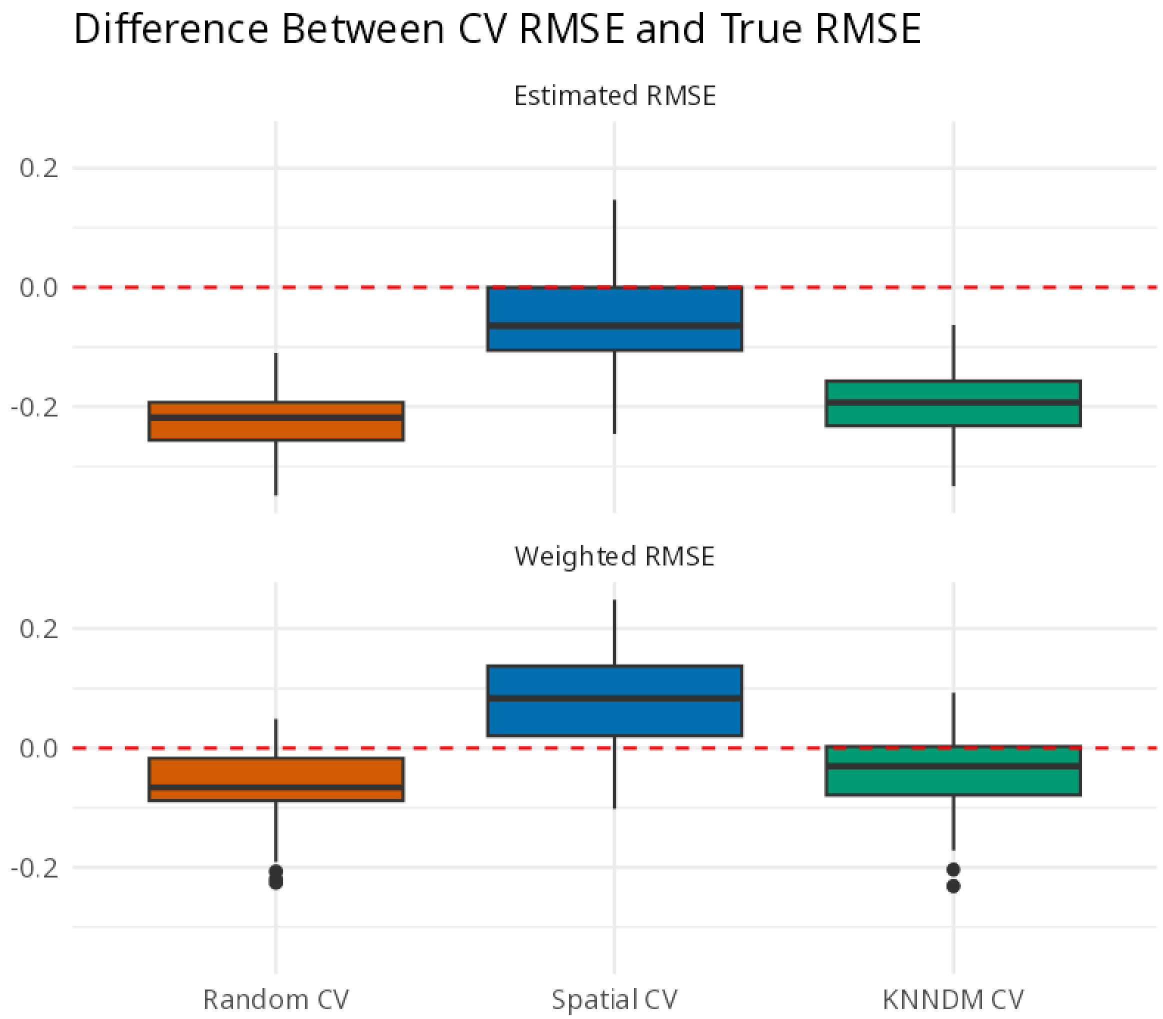




A piece of evidence

Effect of weighting validation points





Key components of *prediction-domain adaptive evaluation*

Contact

https://jakubnowosad.com

Resources

Papers: Nowosad et al., 2026,
Linnenbrink et al., 2026

Slides:

Acknowledgements

Remote Sensing & Spatial Modelling

Machine Learning for Earth Observation
22 - 24th June 2026

Funded by the European Union

Key components of *prediction-domain adaptive evaluation*

1. **Define the prediction domain**

Contact

https://jakubnowosad.com

Resources

Papers: Nowosad et al., 2026, Linnenbrink et al., 2026

Slides:

Acknowledgements

Remote Sensing & Spatial Modelling

Machine Learning for Earth Observation 22 - 24th June 2026

Funded by the European Union

Key components of *prediction-domain adaptive evaluation*

1. **Define the prediction domain**
2. **Construct validation folds that reflect the prediction domain**

Contact

https://jakubnowosad.com

Resources

Papers: Nowosad et al., 2026, Linnenbrink et al., 2026

Slides:

Acknowledgements

Remote Sensing & Spatial Modelling

Machine Learning for Earth Observation
22 - 24th June 2026

Funded by the European Union

Key components of *prediction-domain adaptive evaluation*

1. **Define the prediction domain**
2. **Construct validation folds that reflect the prediction domain**
3. **Weight validation samples by their prevalence in the prediction domain**

Contact

https://jakubnowosad.com

Resources

Papers: Nowosad et al., 2026, Linnenbrink et al., 2026

Slides:

Acknowledgements

Remote Sensing & Spatial Modelling

Machine Learning for Earth Observation
22 - 24th June 2026

Funded by the European Union

Key components of *prediction-domain adaptive evaluation*

1. **Define the prediction domain**
2. **Construct validation folds that reflect the prediction domain**
3. **Weight validation samples by their prevalence in the prediction domain**

Open questions remain, including how to mix these three components together.

Also: these are three important components, not a complete theory of spatial ML evaluation.

Contact

https://jakubnowosad.com

Resources

Papers: Nowosad et al., 2026, Linnenbrink et al., 2026

Slides:

Acknowledgements

Remote Sensing & Spatial Modelling

Machine Learning for Earth Observation
22 - 24th June 2026

Funded by the European Union