Julius Kühn-Institut
Bundesforschungsinstitut für Kulturpflanzen
Federal Research Centre for Cultivated Plants

A Random Forest for Predicting Human Puumala Orthohantavirus Infections in North-Western Germany

Orestis Kazasidis, Joanna Dürger, Christian Imholt, Jens Jacob
Julius Kühn Institute (JKI)
Institute for Plant Protection in Horticulture and Forests
Vertebrate Research, Münster, Germany

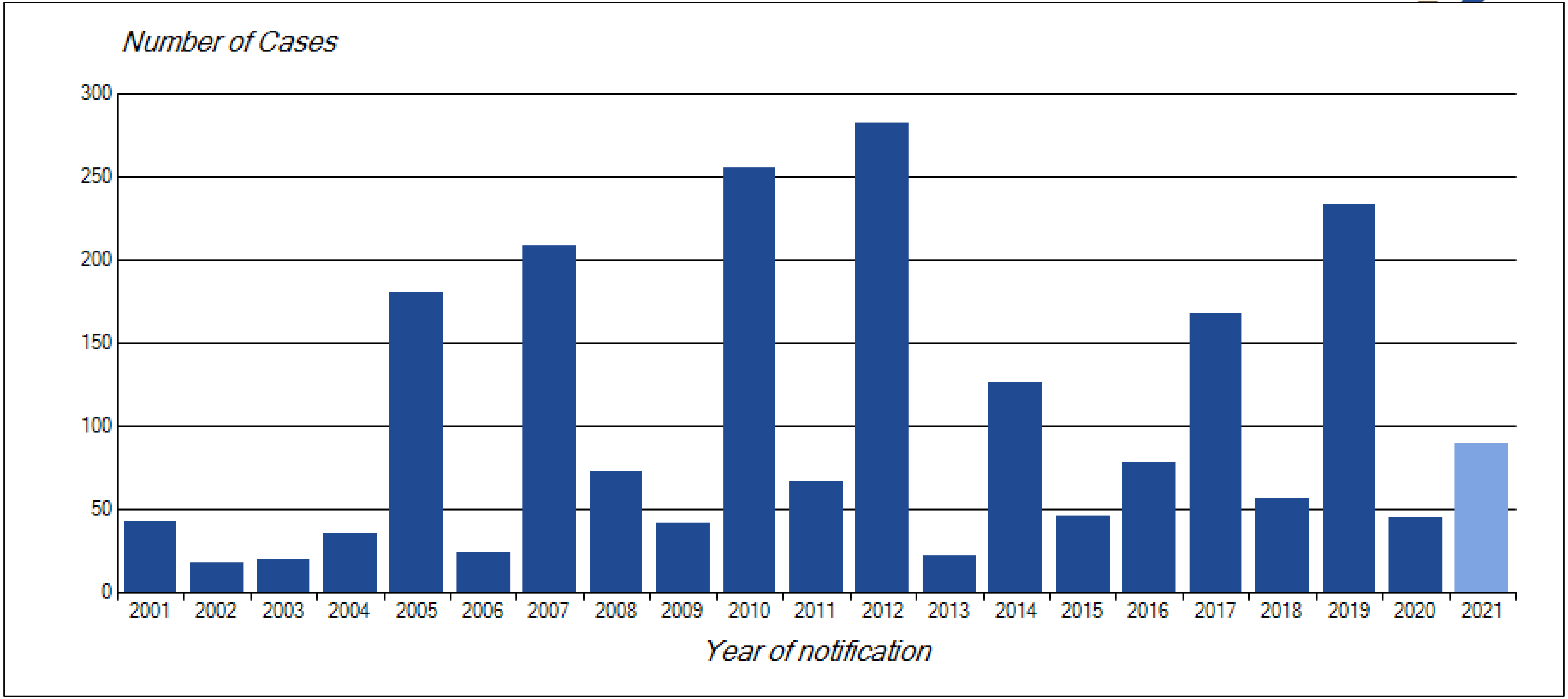
Motivation

Objective: Prediction of the human Puumala orthohantavirus infections in North-Western Germany

- Puumala orthohantavirus (PUUV) can be transmitted to humans by infected bank voles.
- Human infections fluctuate regularly. Outbreaks every 2-3 years.
- Inhomogeneous annual spatial distribution of the human infections.
- Disease outbreaks related to rodent outbreaks (driver of damage to forest trees).

Principle: Estimation of the infection risk via the reported incidence

Motivation



Annual human PUUV-infections in Lower Saxony and North Rhine-Westphalia. Figure adapted from: Robert Koch Institute, *SurvStat@RKI 2.0*, https://survstat.rki.de, Status 25-11-2021.

Part I – Fundamentals

Problem definition

- Districts
- Risk class thresholds
- Target and weights
- Predictors

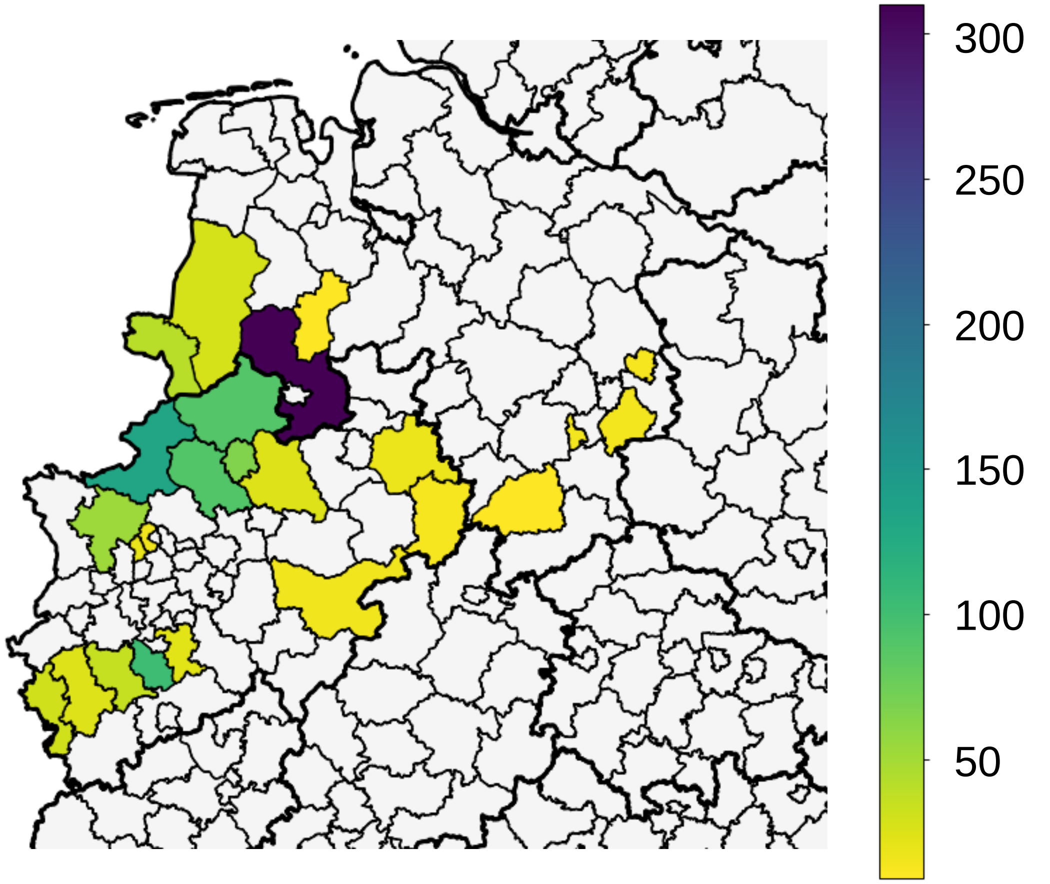
Model overview and diagram

Selection of districts (based on 2006 – 2017)

Criteria

- total infections ≥ 5
- maximum annual infections ≥ 3
- years with at least 2 infections ≥ 2
- years in the medium or high risk class ≥ 1

Incidence-based risk classes
low risk: [0, 1.5)
medium risk: [1.5, 4)
high risk: [4, ∞)



Lower Saxony
7 districts
LK Emsland
LK Grafschaft Bentheim
LK Northeim
LK Osnabrück*
LK Vechta
LK Wolfenbüttel
SK Wolfsburg

North Rhine-Westphalia
16 districts
LK Borken
LK Coesfeld
LK Düren
LK Hochsauerlandkreis
LK Höxter
LK Lippe
LK Rhein-Erft-Kreis
LK Rheinisch-Bergischer Kreis
LK Steinfurt
LK Warendorf
LK Wesel
SK Bottrop
SK Köln
SK Münster
SK Oberhausen
StädteRegion Aachen

LK = Landkreis (rural district) | SK = Stadtkreis (urban district)

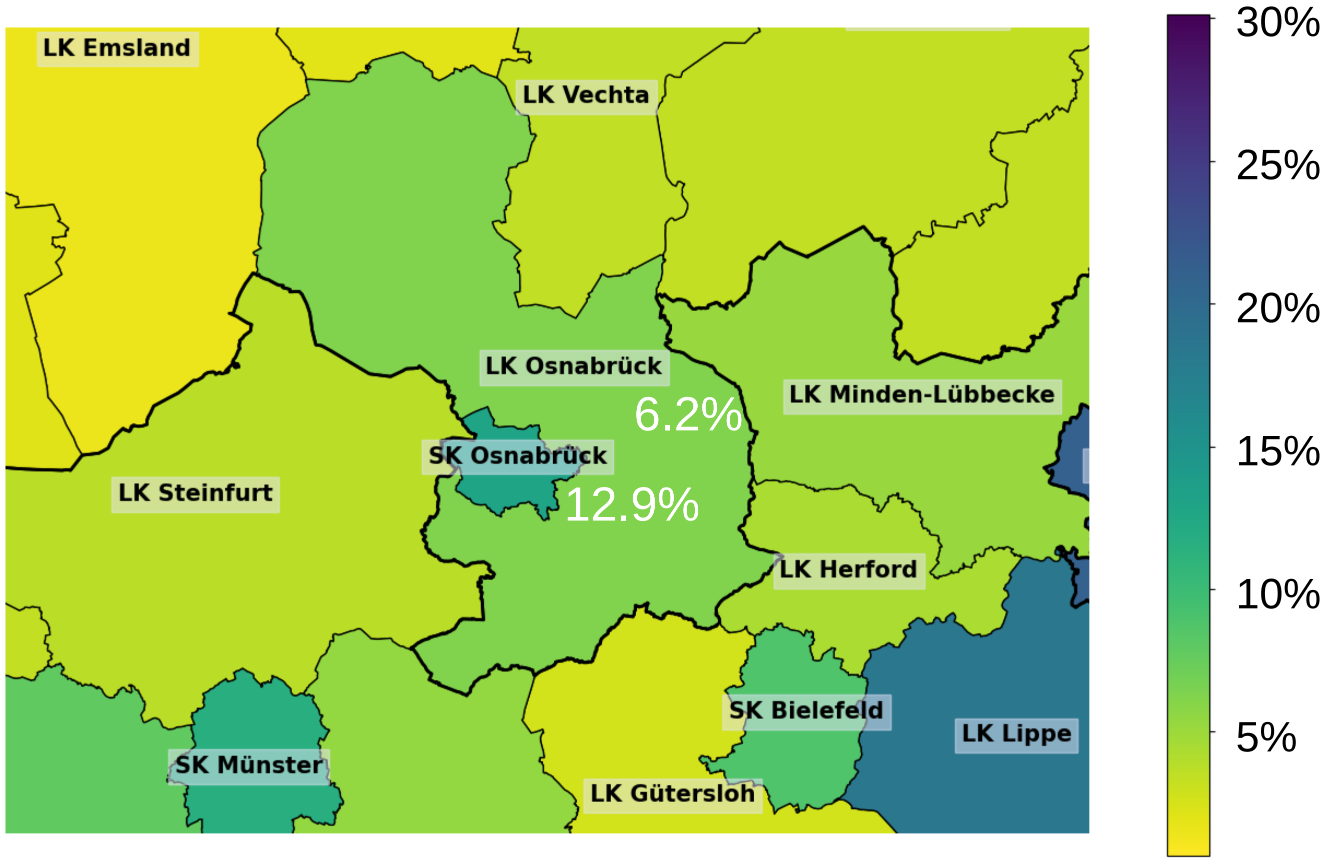
About Osnabrück (1/2)

The urban district SK Osnabrück and the rural district LK Osnabrück are combined.

Reasoning:

- ➢The infections in SK Osnabrück are expected to originate (at least partially) from the area of LK Osnabrück.
- ➢The incidences in the two districts are highly correlated (next slide).
- ➢Certain parameters that we aim to use as predictors differ significantly between the urban and rural districts of Osnabrück (land cover data, example: figure on the right).

Broad-leaved forest proportion (2018)



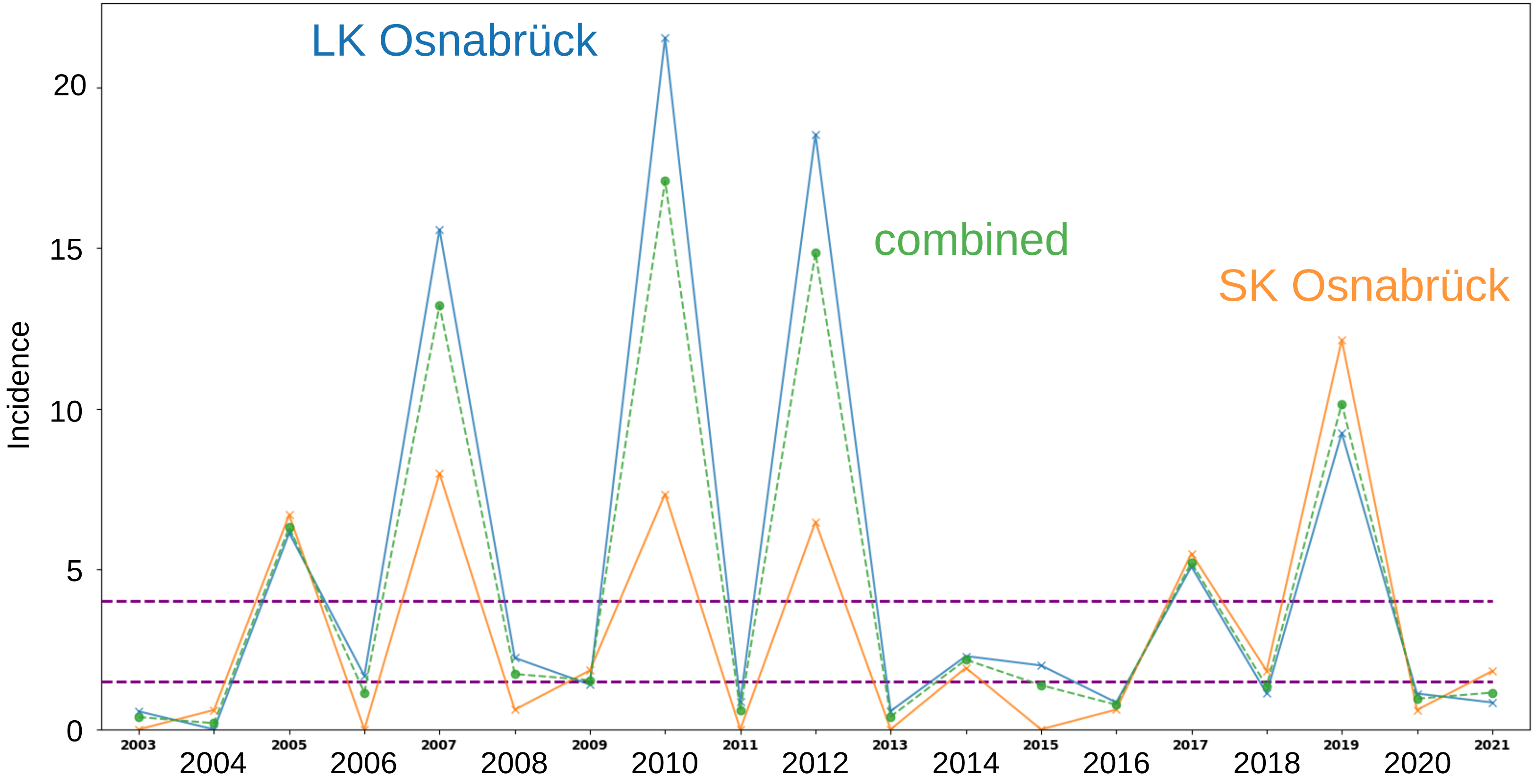
About Osnabrück (2/2)

incidence correlation = 0.767 (p<0.001)

For the most part, the following analysis holds even if the two Osnabrück districts are not combined, yet lower accuracies are expected.

For comparison, there are only three additional pairs of neighboring districts with correlation larger than 0.7.

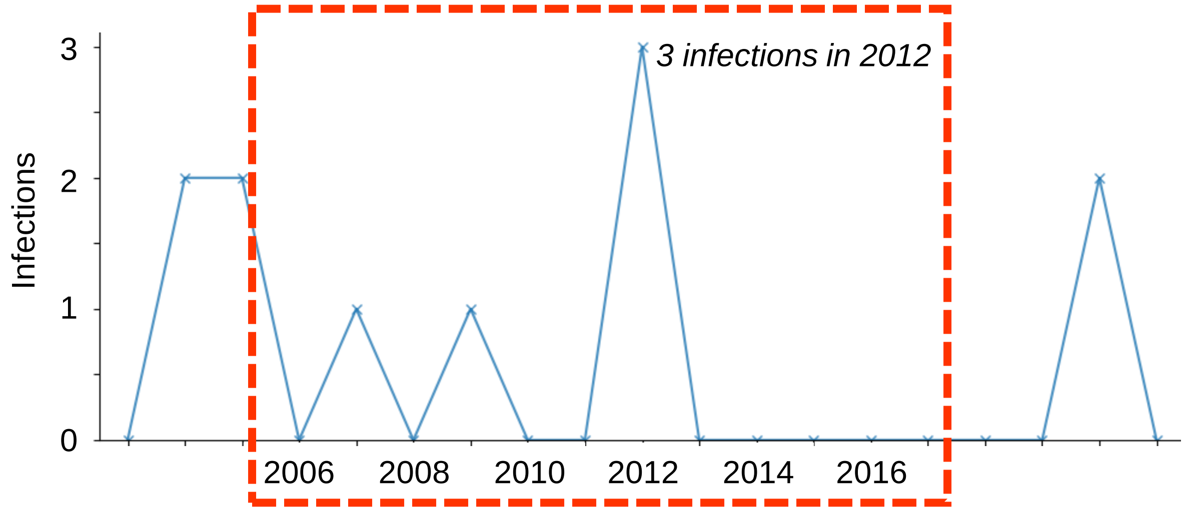
- 0.784 for the pairs:
LK Coesfeld with SK Münster,
LK Rheinisch-Bergischer Kreis with SK Köln
- 0.746 for the pair:
LK Borken with LK Wesel



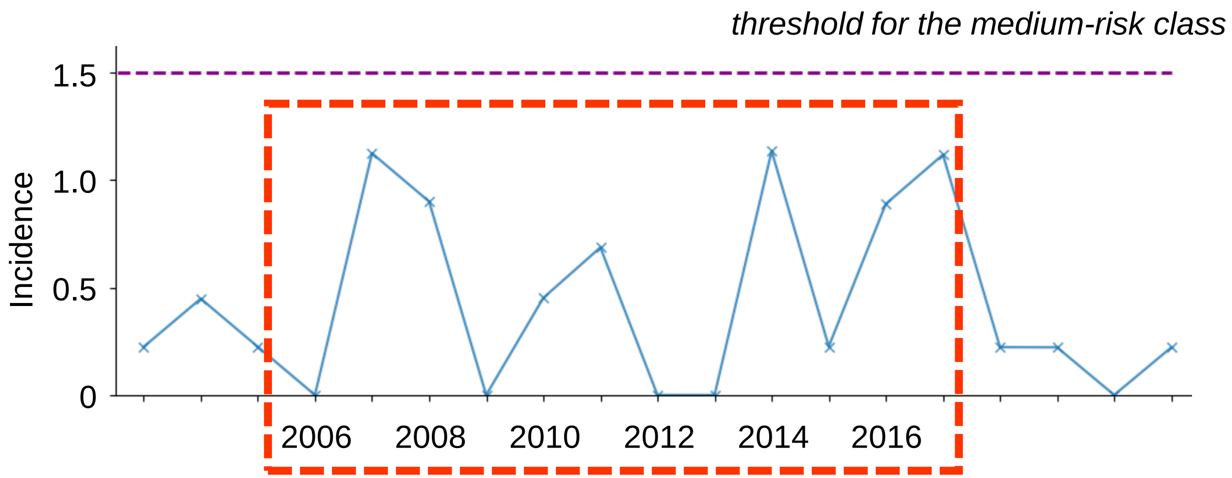
Removed districts

- 51 removed districts with at least 1 infection in 2006 – 2017
- These districts had 207 total infections (15% of the total infections in these 12 years).
- Maximum annual infections = 6 (Region Hannover in 2012)
- Maximum annual incidence = 3.3 (LK Helmstedt in 2012)

LK Helmstedt (5 infections)
removed, because only one year with at least 2 infections



LK Rhein-Kreis Neuss (29 infections)
removed, because it is always in the low-risk class



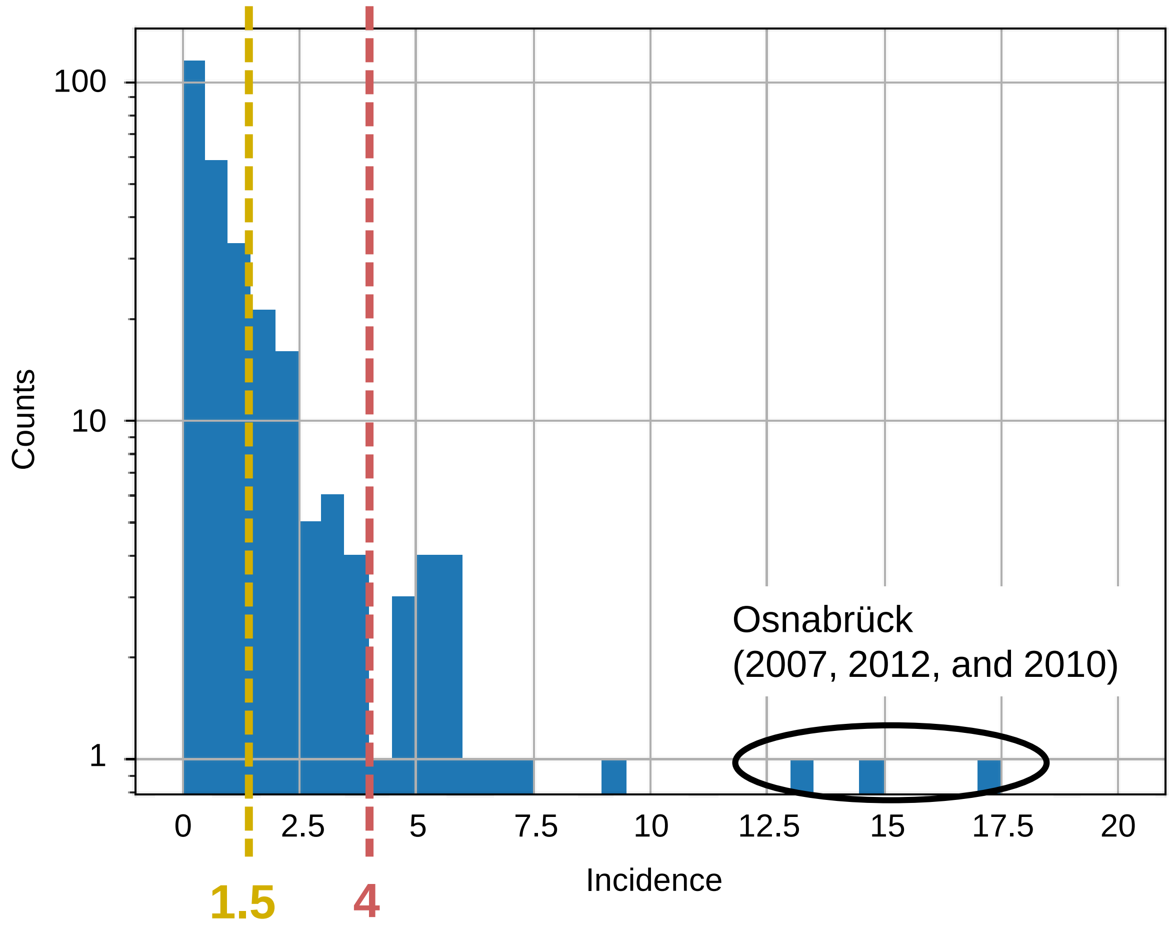
Selection of the class thresholds

Incidence-based risk classes

low risk:	**[0 , 1.5)**	205 samples ≈ 74%
medium risk:	**[1.5, 4)**	52 samples ≈ 19%
high risk:	**[4 , ∞)**	19 samples ≈ 7%

The first threshold is selected at 1.5, because for incidence < 1.5 there are several samples with only a couple (1 to 3) infections.
This is considered a balancing effect for an incidence-based metric with respect to infections.

The second threshold is selected by inspecting the histogram of the incidence values (right figure for bin width = 0.5).

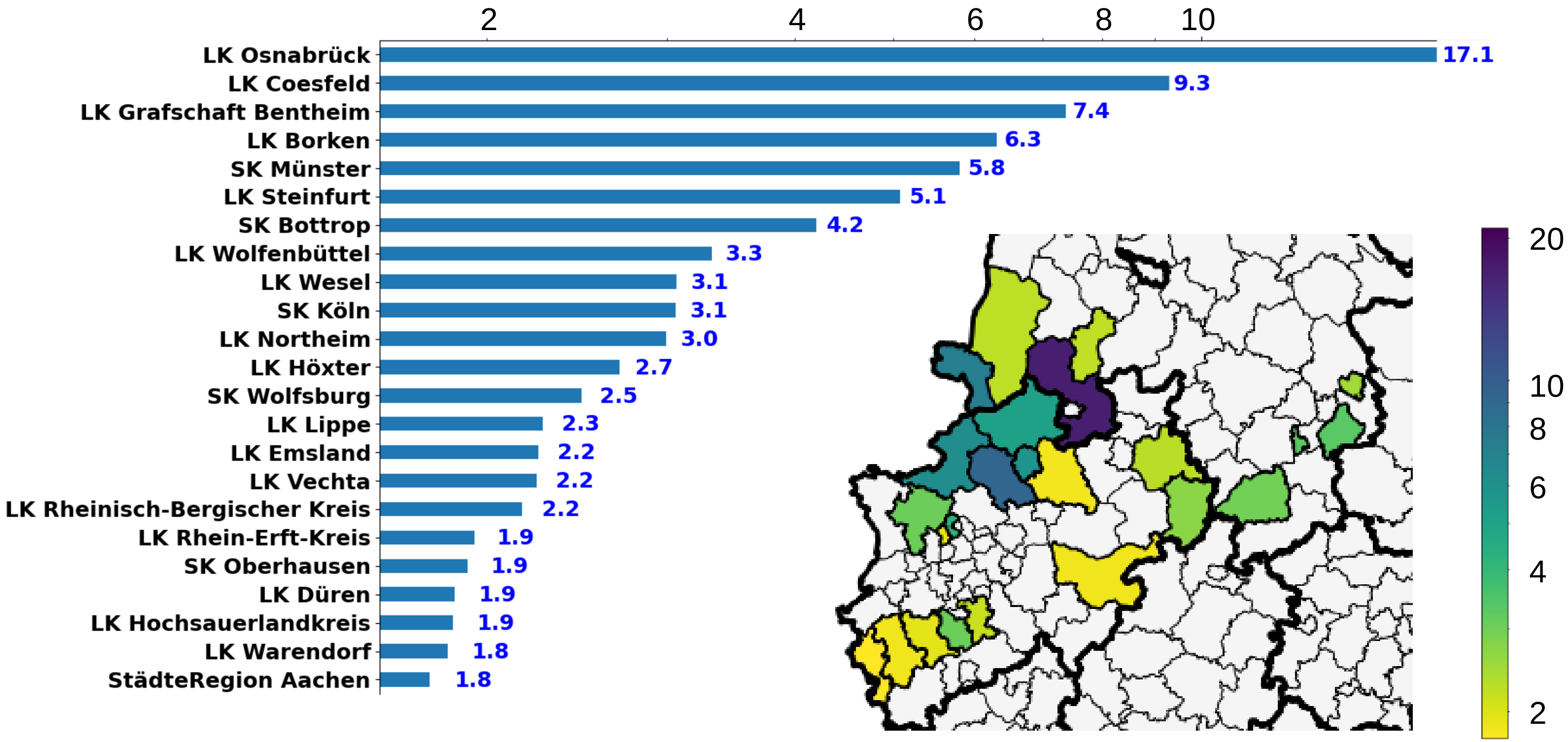


Scaling the incidence (1/3)

The annual incidence for each district is scaled to [min, max] = [0, 1] for the specific time period.

$$\text{scaled incidence} = \frac{\text{incidence} - \min(\text{district})}{\max(\text{district}) - \min(\text{district})}$$

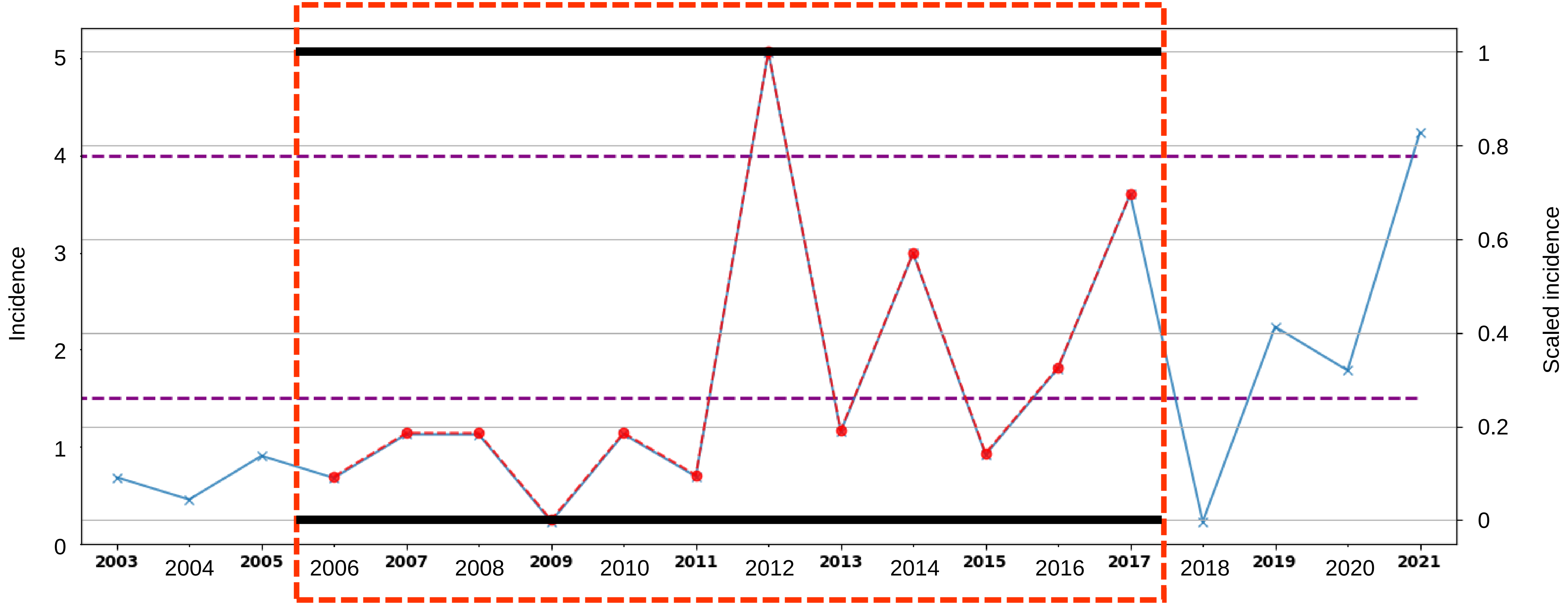
maximum annual incidence per district (log scaling on x-axis)



The min-scaling influences only LK Osnabrück (min=0.40) and LK Steinfurt (min=0.22).

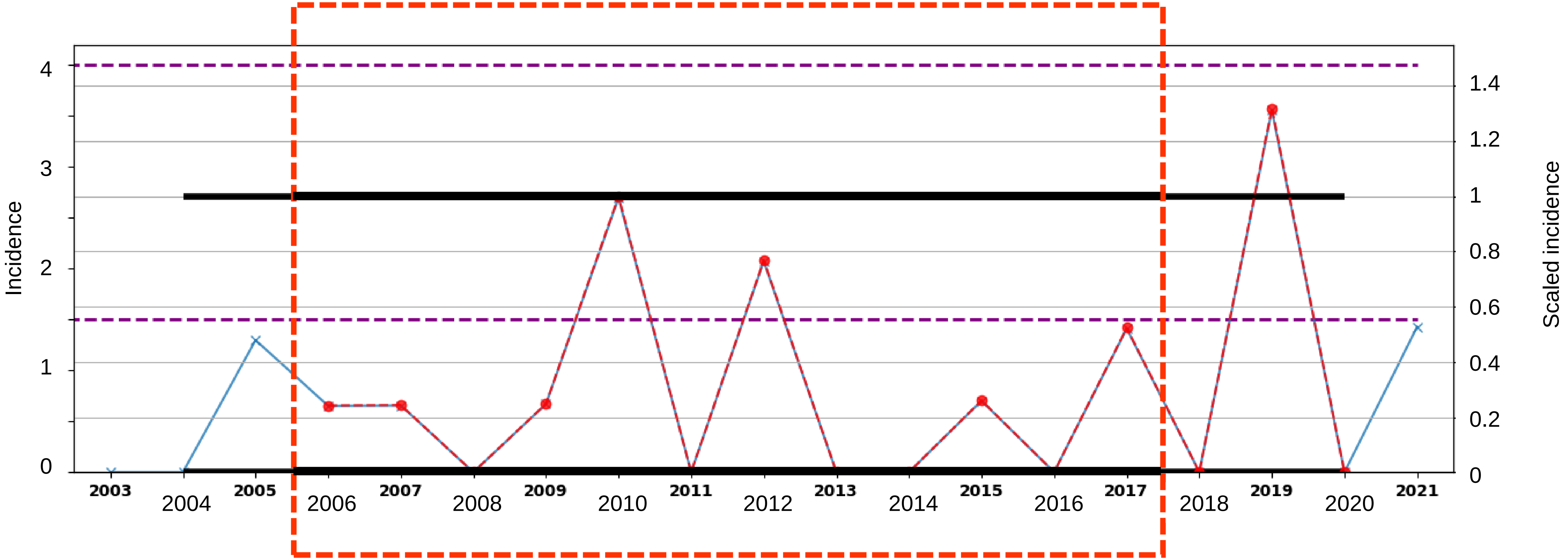
Scaling the incidence (2/3)

The effect of scaling in the district LK Steinfurt.



Scaling the incidence (3/3)

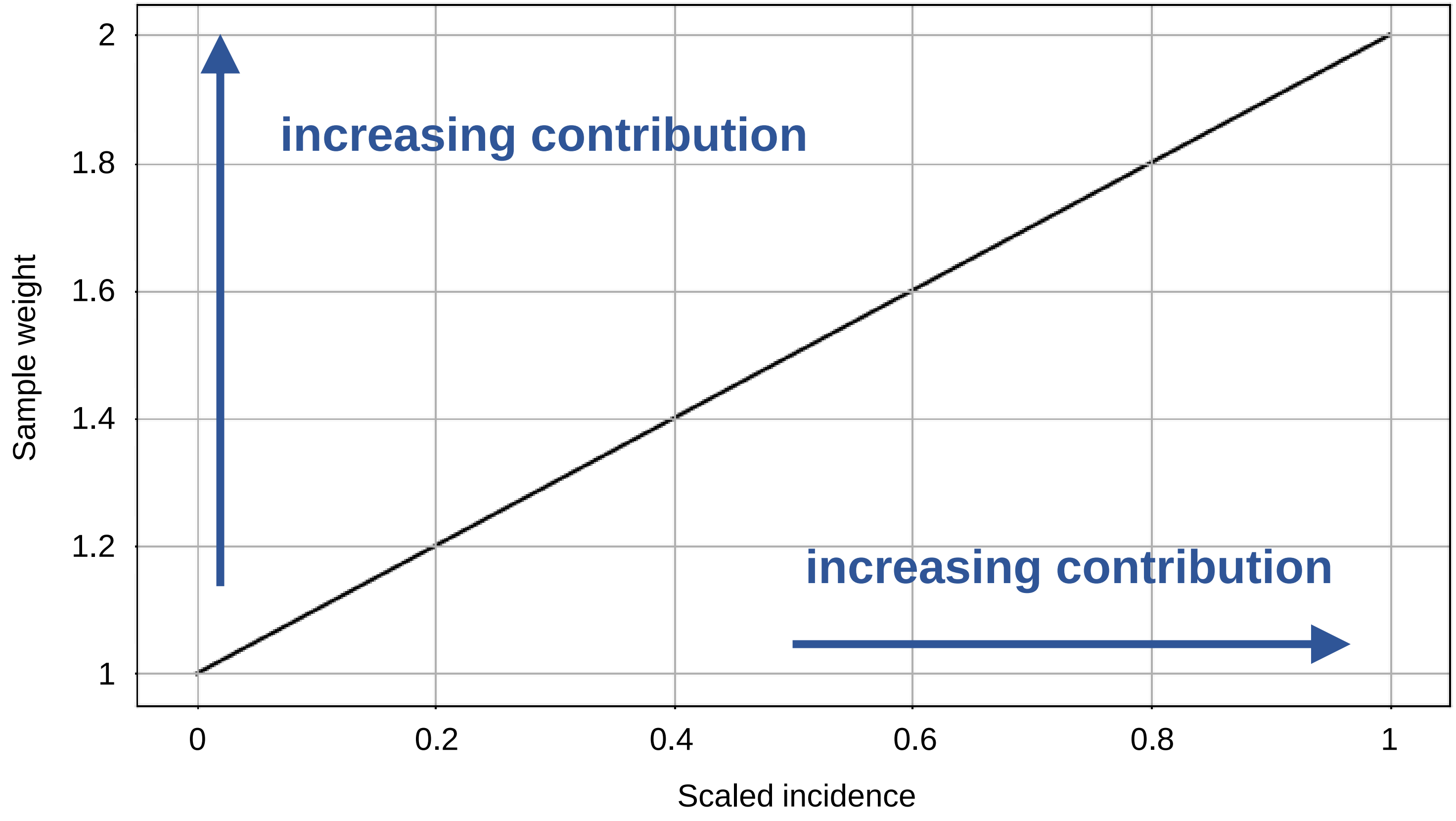
The effect of scaling in the district LK Höxter.



Weights

The samples are weighted based on their target value.

$$\text{sample weight} = \mathbf{1} \cdot \text{scaled incidence} + \mathbf{1}$$



Initial pool of predictors

Weather parameters*

Monthly

Minimum air temperature
Average air temperature
Maximum air temperature
Soil temperature
Sunshine duration
Precipitation

two preceding years (V1 and V2) until the previous September

Annual

Summer days
Ice days
Snow cover days
Beech - beginning of leaves unfolding
Beech - autumn leave coloring
Beech - autumn leave fall
Begin of the vegetation period
End of the vegetation period

two years before (V2)

* Source: DWD Climate Data Center (CDC).

Land cover*

Broad-leaved forest proportion
Mixed forest proportion
Broad-leaved + mixed forest proportion
Urban proportion

* Source: European Union, Copernicus Land Monitoring Service 2006 – 2018.

Phenology*

Beech flowering intensity

two preceding years (V1 and V2)

* Source: Dagmar Schneck, State Office for Forest Reproductive Material, Brandenburg State Forestry Office, personal communication.

In total: 140 features

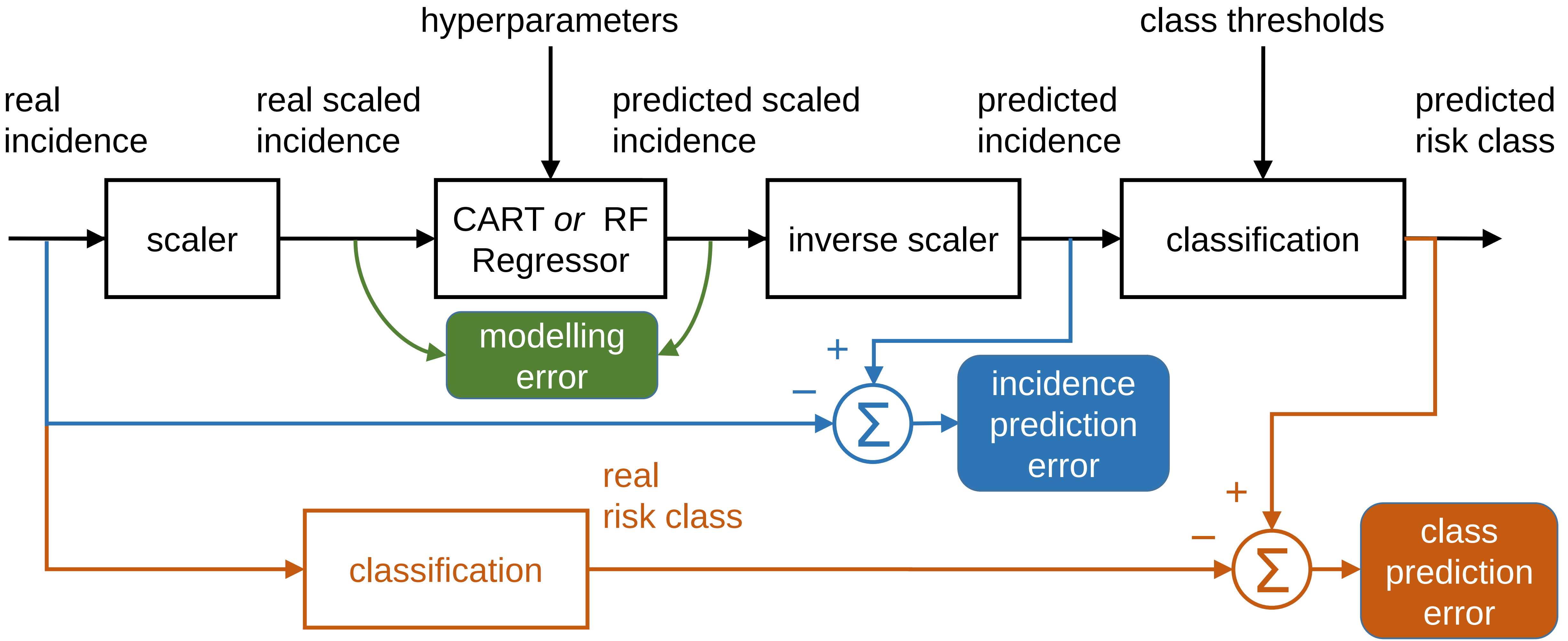
Images credits: 1. DWD, 2. Copernicus Land Monitoring Service, 3. Dagmar Schneck

Model: Overview

Parameter	***Value***
Districts	23 districts, based on specific criteria
Target	incidence scaled at the district level
Predictors	selection from 140 features
Primary method	CART *or* RF Regressor with weights
Training	in the years 2006 – 2017
Validation	external (in the years 2018 – 2020)

CART = Classification and Regression Tree
RF = Random Forest

Model: Diagram



Part II – CART and Random Forest

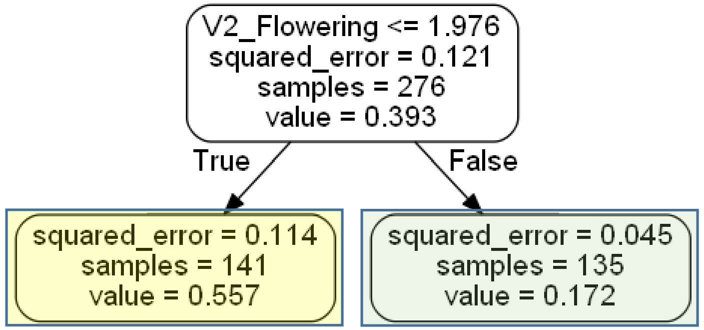
- Short description
- Comparison
- Performance metrics

CART fundamentals

Total training dataset
276 samples (23 districts x 12 years)
average incidence = 1.30
average scaled incidence = 0.31
weighted average scaled incidence = 0.39
weighted mean squared error = 0.12

Each sample corresponds to one year in one district.

Example for a single split

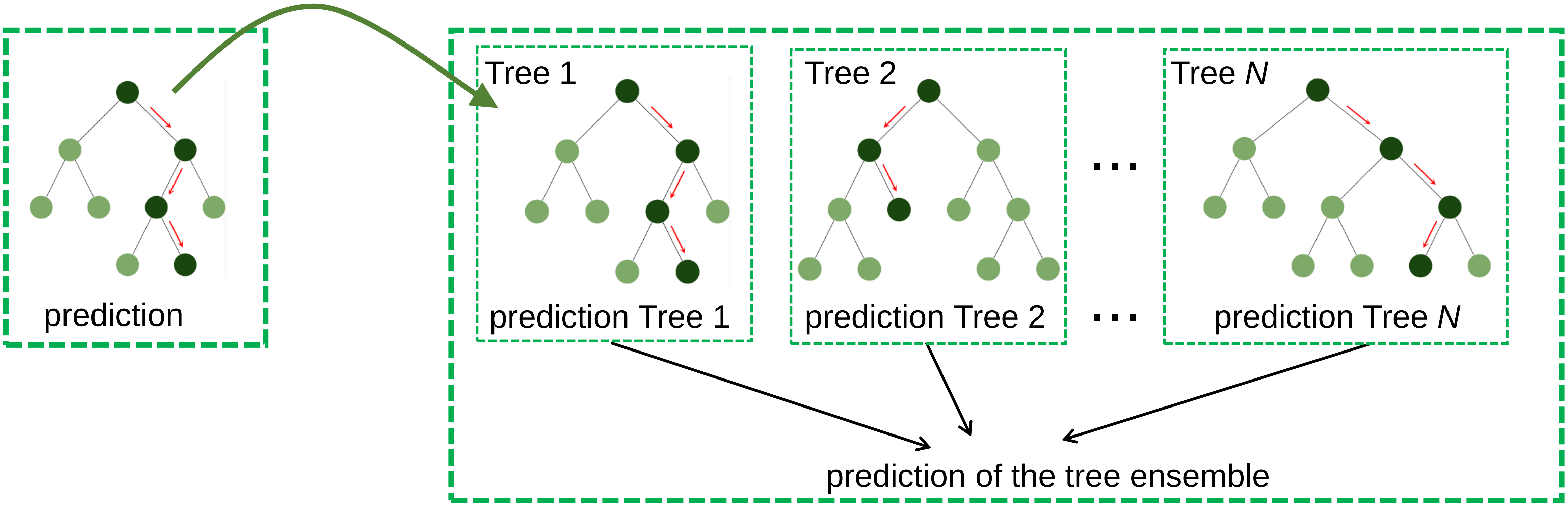


R^2 = 0.30

Random forest: a special tree ensemble

CART generates
a single decision tree.

A random forest comprises several decision trees, each
trained on **a subset of the samples**
with **a subset of the features for each split**.

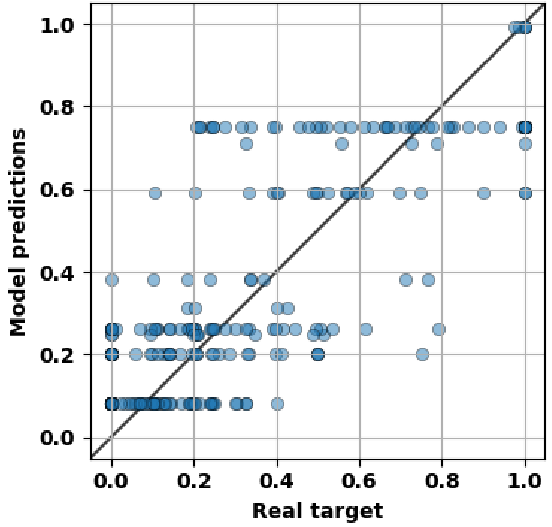


Comparison

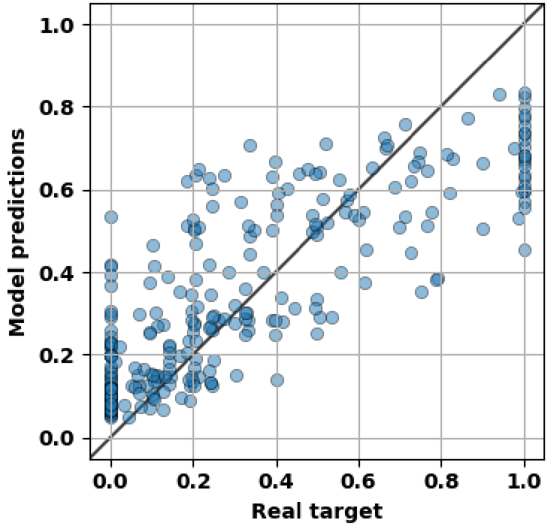
CART
deterministic

Advantages
- better performance in the training set
- requires less parametrization
- easily interpretable

CART predictions



RF predictions



Random Forest (RF)
stochastic

Advantages
- not greedy
- less overfit → better performance in a test set
- more robust
- possible estimation of prediction accuracy
- continuous output

Performance metrics

- Risk-class accuracy → accuracy paradox

- Confusion matrix
- Null class accuracy

- Regression metric: mean squared error
- R^2

- out-of-bag score (only for RF)
- Precision and recall, F-score, ...

Ideal confusion matrix
Ideal class accuracy = 100%

		Predicted class		
		low	medium	high
Real class	high	0	0	19
	medium	0	52	0
	low	205	0	0

Null confusion matrix
Null class accuracy = 74%

		Predicted class		
		low	medium	high
Real class	high	19	0	0
	medium	52	0	0
	low	205	0	0

Performance metrics – example: LK Coesfeld

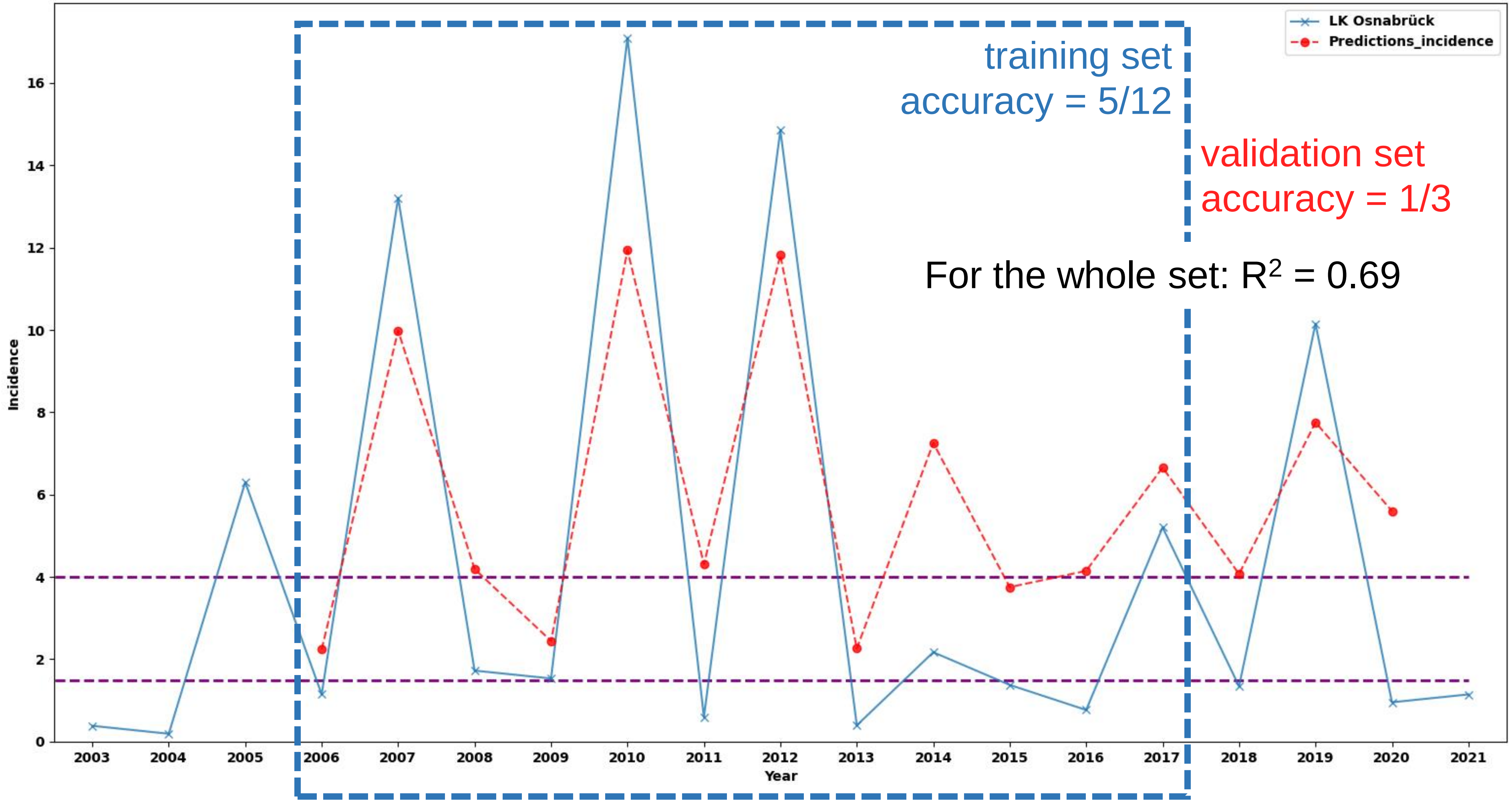
High R^2 and high accuracy

Predictions with an RF, trained in 2006-2017

Performance metrics – example: LK Osnabrück

High R^2 but low accuracy

Predictions with an RF, trained in 2006-2017



Performance metrics – example: LK Lippe

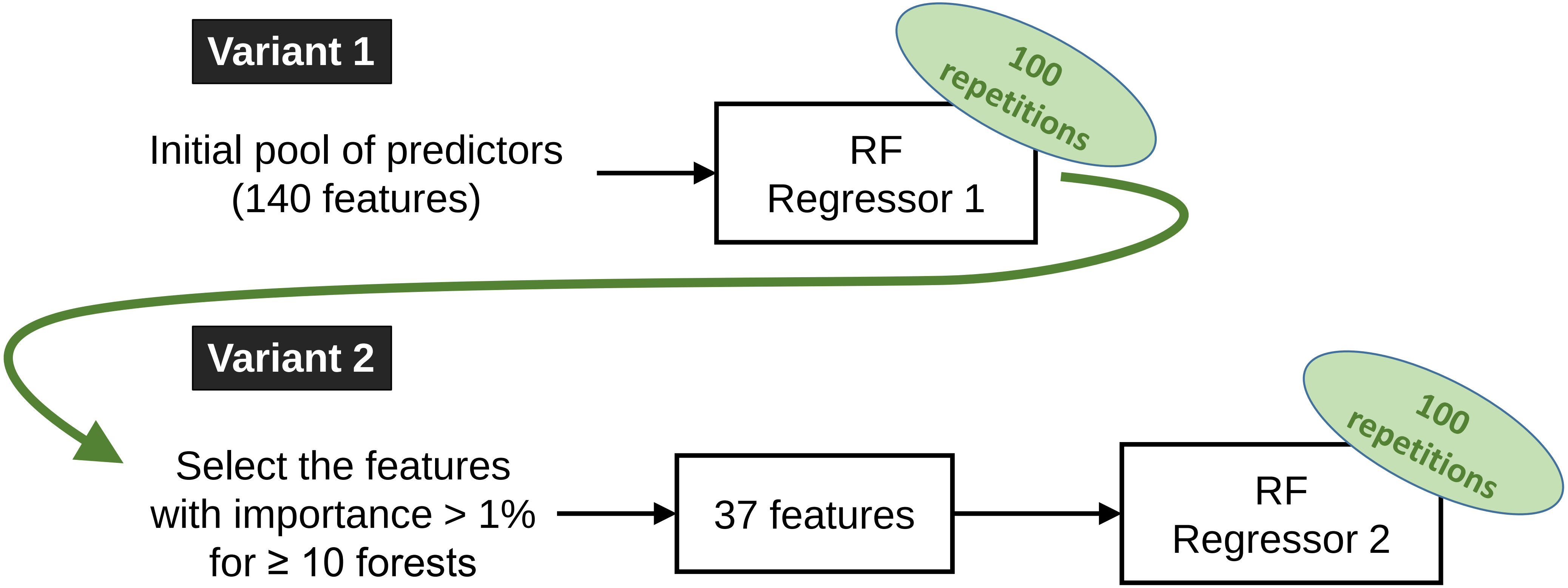
Low R^2 and high accuracy

Predictions with an RF, trained in 2006-2017

Part III – Prediction with a Random Forest

- Selection of predictors
- Performance

RF models – Selection of predictors



RF models – Performance (training dataset)

Variant 1 – all 140 features

R^2 complete model : 0.81 [0.80,0.82] (CART: 0.75)
Class accuracy: 82% [80%,83%] (81%)
Null Accuracy: 74%

an example confusion matrix

		Predicted class		
		low	medium	high
Real class	high	0	7	12
	medium	15	32	5
	low	182	22	1

Variant 2 – selected 37 features

R^2 complete model : 0.80 [0.79,0.81]
Class accuracy: 81% [80%,83%]

an example confusion matrix

		Predicted class		
		low	medium	high
Real class	high	0	6	13
	medium	17	31	4
	low	180	24	1

5 differences in the confusion matrix

RF models – Performance (validation dataset)

Variant 1 – all 140 features

R^2 complete model : 0.43 [0.36,0.51] (CART: 0.14)
Class accuracy: 77% [74%,80%] (70%)
Null Accuracy: 70%

an example confusion matrix

		Predicted class		
		low	medium	high
Real class	high	0	3	2
	medium	9	7	0
	low	44	2	2

Variant 2 – selected 37 features

R^2 complete model : 0.42 [0.33,0.49]
Class accuracy: 77% [74%,78%]

an example confusion matrix

		Predicted class		
		low	medium	high
Real class	high	0	3	2
	medium	8	8	0
	low	43	3	2

2 differences in the confusion matrix

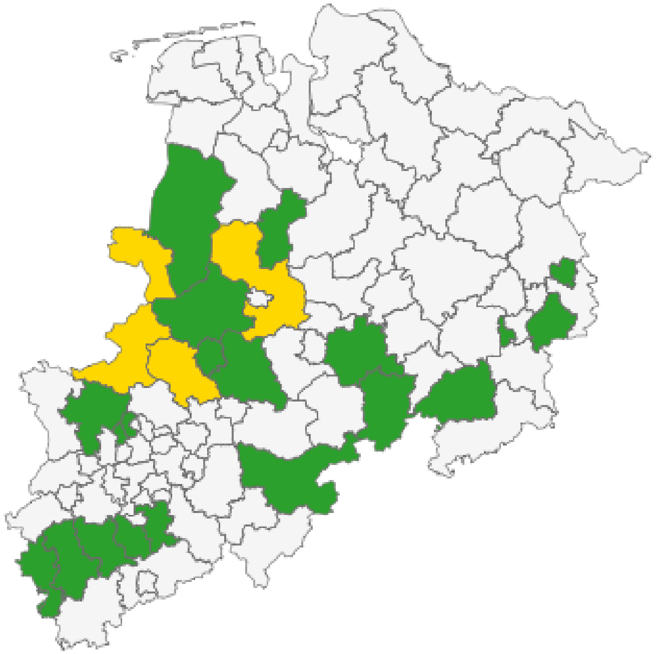
Part IV – Prediction for 2022

Model characteristics

- Training in the years 2006 – 2020
- Random Forest with 1000 estimators

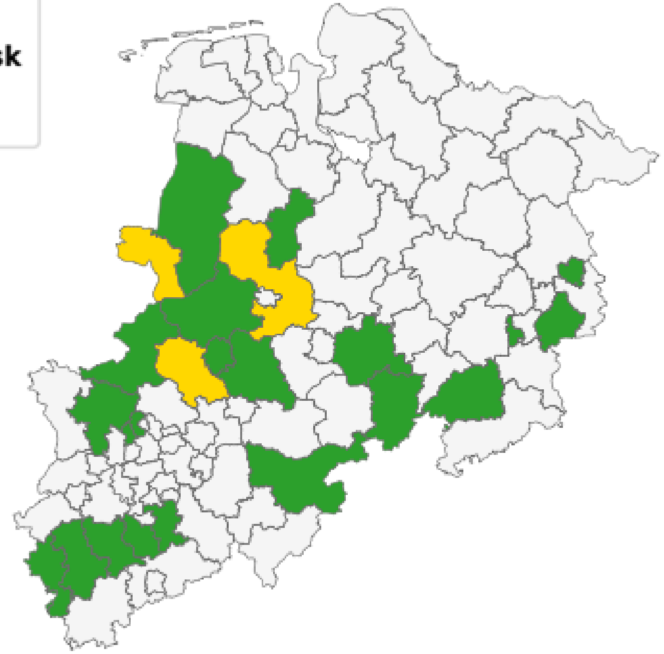
Prediction for 2022 – Risk class

Random Forest
all 140 features

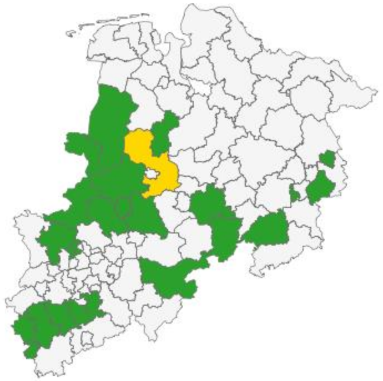


- low risk
- medium risk
- high risk

Random Forest
selected 37 features



CART



Thank you

This study was supported by the Federal Ministry of Education and Research (BMBF) for the “RoBoPub” consortium (grant number 01KI1721E), and the German Environment Agency (UBA) within the departmental research plan of the Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (research code 3720 48 401 0).