Universität Zürich[UZH]

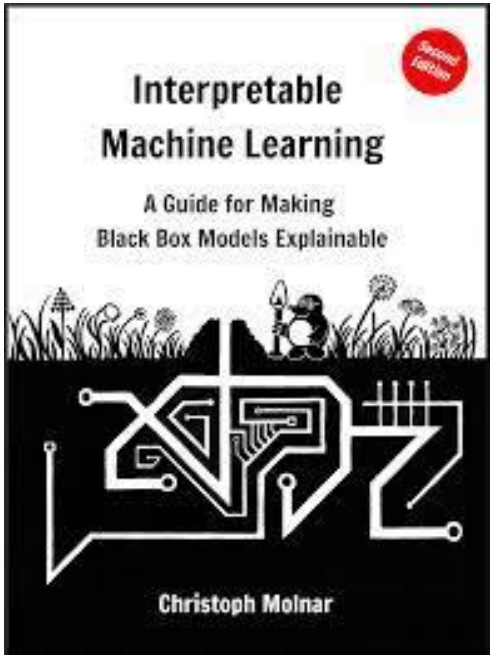
Epidemiology, Biostatistics and Prevention Institute

Interpretable machine learning: applications in public health

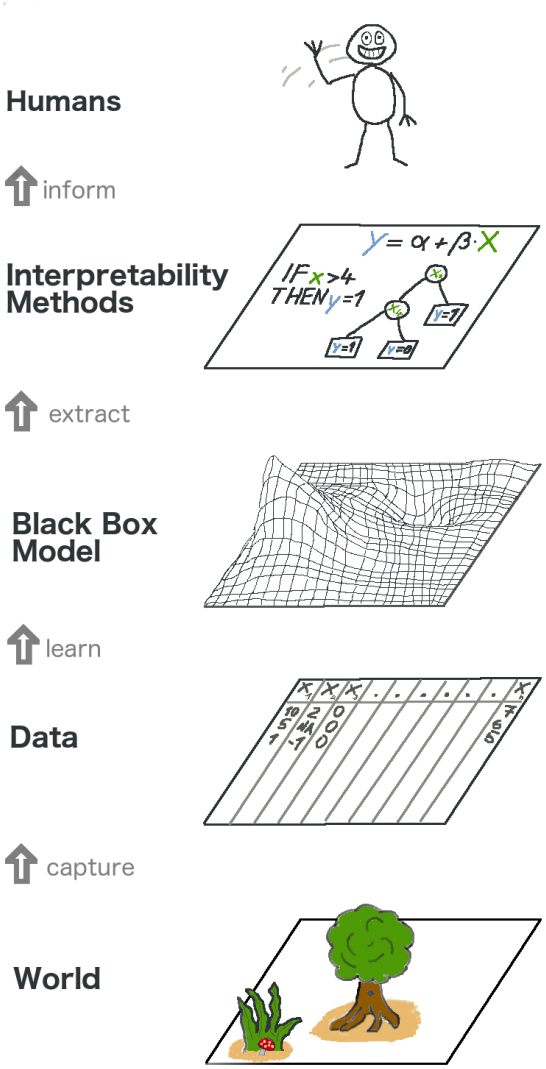
Andrea Farnham, PhD

Zurich R User Meetup

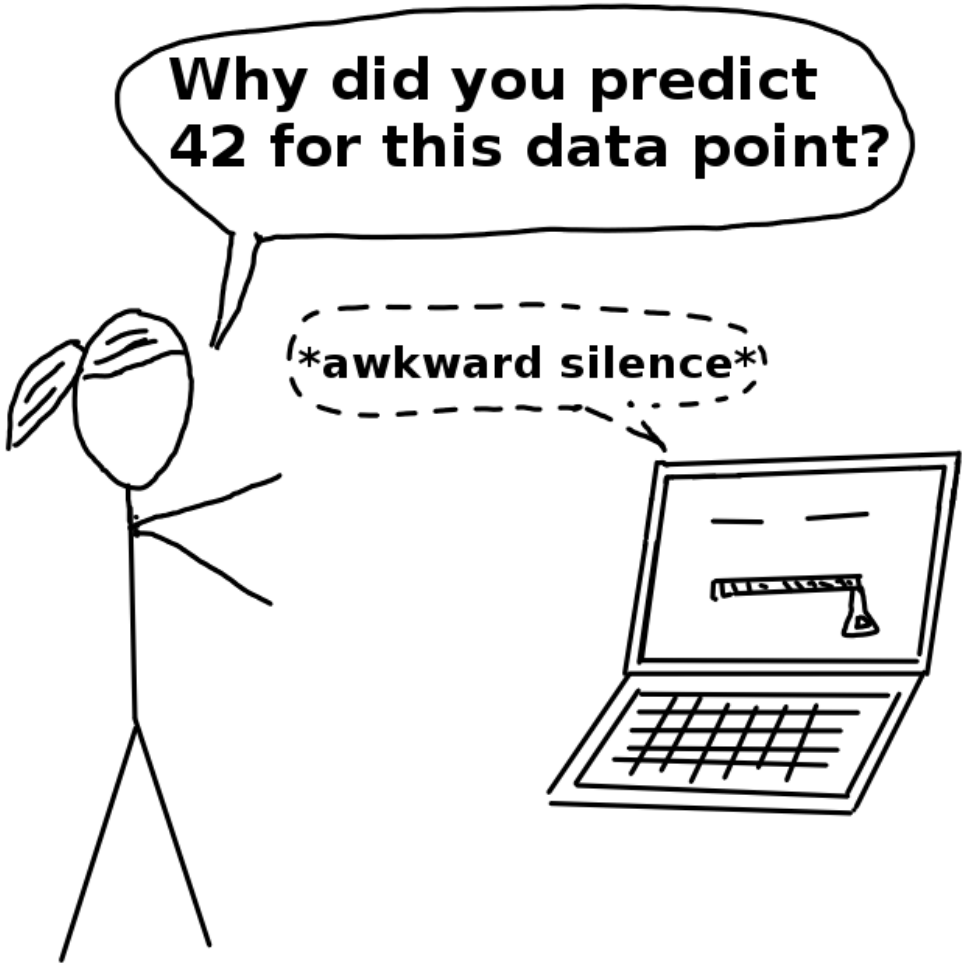
2 November 2023



Why do we need interpretable machine learning methods?



Source, Christoph Molnar, Interpretable Machine Learning



Source, Christoph Molnar, Interpretable Machine Learning

Why do we need interpretability?

- Imagine you came up with an algorithm that “learns” how to administer the exact right dose of pain medication automatically and continuously for every patient
- You don’t know **why** the machine administers the dose it does, but you know it isn’t random
- One day, the machine kills a terminally ill patient by administering her 17x the normal dose

Universität Zürich[UZH]

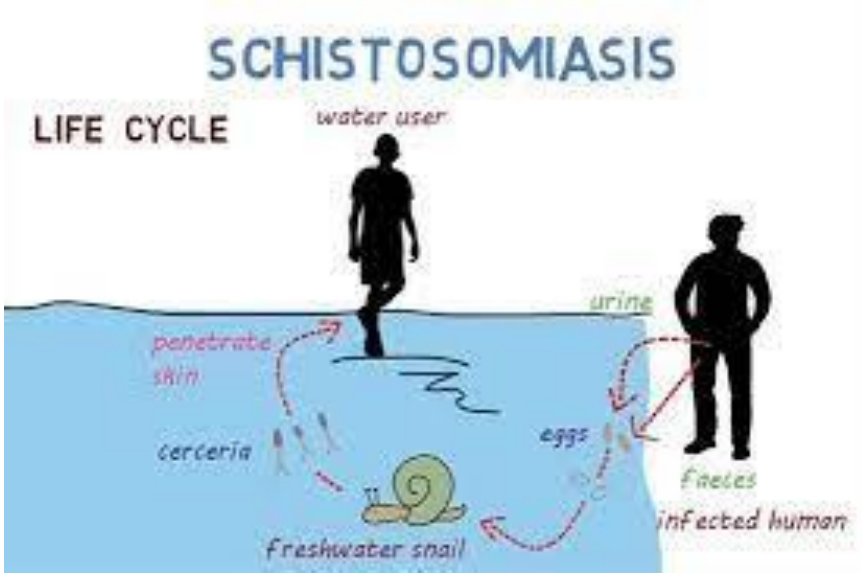
Epidemiology, Biostatistics and Prevention Institute

A case study from my work: can we use interpretable machine learning to better understand schistosomiasis and hookworm?

Swiss TPH

Swiss Tropical and Public Health Institute
Schweizerisches Tropen- und Public Health-Institut

What are schistosomiasis and hookworm?



Source: WHO

Neglected tropical diseases (NTDs) cause a huge burden of disease

Huge burden: Parasitic worm infections such as schistosomiasis and intestinal worm infections affect more than 1 billion people globally

Risk factors: They affect mostly people in the poorest communities and those without access to clean water and sanitation. The highest levels of infection are in school-age children

Chronic health problems and reduced productivity: Infection can lead to chronic illness. The worms can damage organs, such as the liver, bladder, and intestines, which can cause pain, fatigue, and long-term health problems.

Preventable and treatable

Understanding of transmission is driven largely through modelling

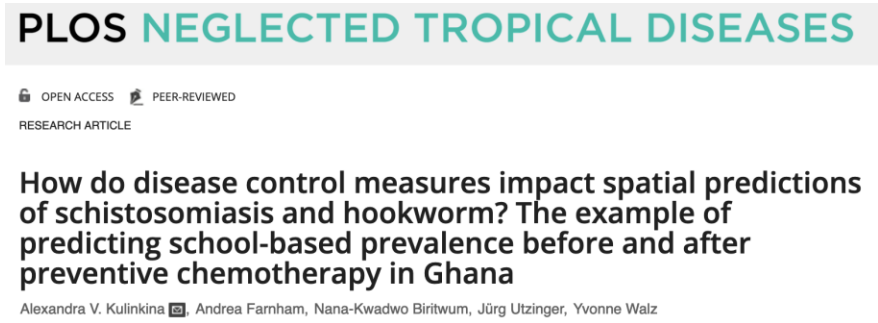
To model prevalence and transmission, scientists traditionally use periodic school-based or community-based prevalence surveys coupled with remotely sensed (RS) environmental predictors

Recent innovations:

- utilizing fine resolution RS data (e.g., Landsat 8)
- employing a larger number of relevant environmental indicators derived from the spectral bands (e.g., modified normalized difference water index [MNDWI])
- using a variable distance radius to extract and aggregate environmental indicator variables around point-prevalence locations

Our research question: are these models still valid in an era of widespread preventative chemotherapy?

- Setting: Ghana
- Two nationally representative school-based prevalence surveys conducted before (2008, n=118 schools) and after (2015, n=158 schools) the launch of large-scale preventive chemotherapy
- Primary outcome: prevalence of infection by *S. haematobium* and hookworm among school-age children.
- Compared model performance before and after the national level intervention



Random forest input predictors

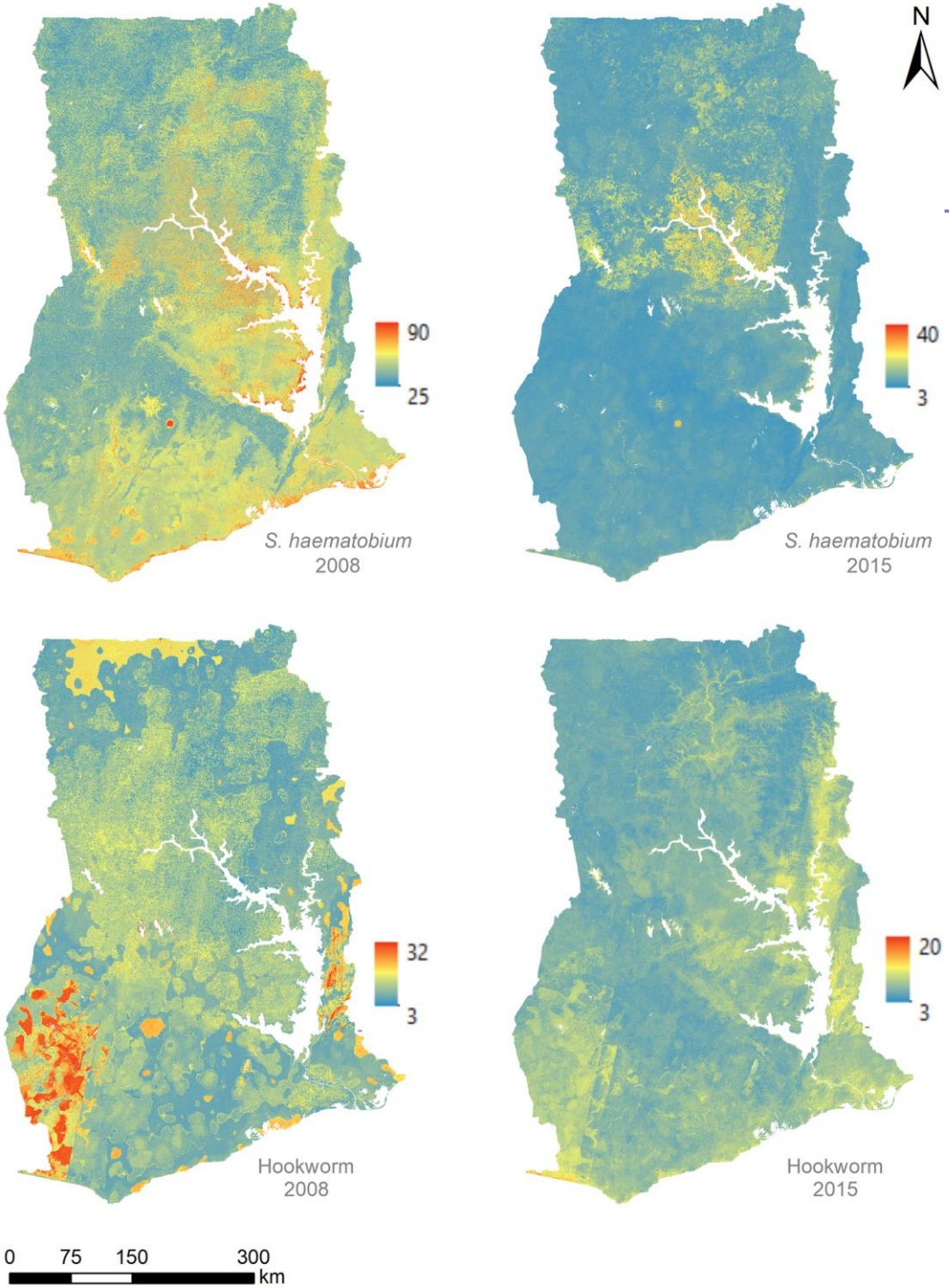
Source	*Variable name*	*Variable type*	*Resolution*	*Aggregation*	*Value range**
OLI	NDVI	Continuous	30 m	Median	0.16 to 0.80
OLI	MNDWI	Continuous	30 m	Median	-0.56 to 0.05
TIRS	LST (˚C)	Continuous	100 m	Median	21.5 to 38.2
DEM	Elevation (m)	Continuous	30 m	Median	12.0 to 537
DEM	Slope (˚)	Continuous	30 m	Median	2.47 to 13.2
DEM	Streams	Binary	30 m	Sum	115 to 3,656
DEM	Stream order	Ordinal	30 m	Maximum	1 to 8
DHS	Access to improved water (%)	Continuous	5 km	Median	29 to 99
DHS	Lack of sanitation facility (%)	Continuous	5 km	Median	0.7 to 98

* Range represents minimum and maximum values present in the dataset (within the largest buffer radius of 5 km in the unmasked dataset).

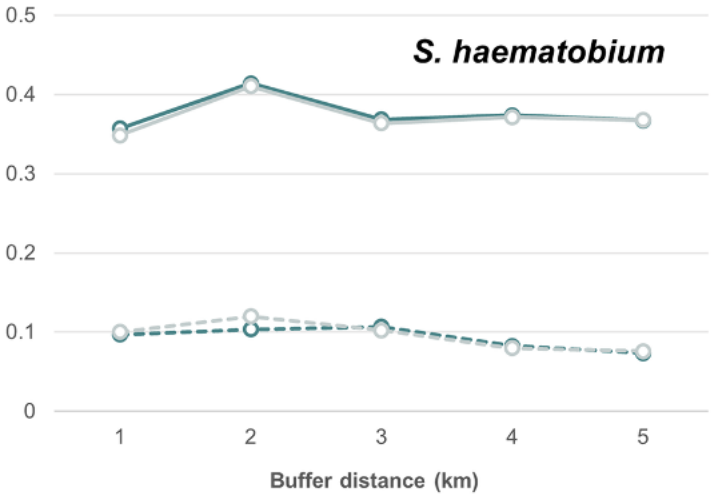
https://doi.org/10.1371/journal.pntd.0011424.t001

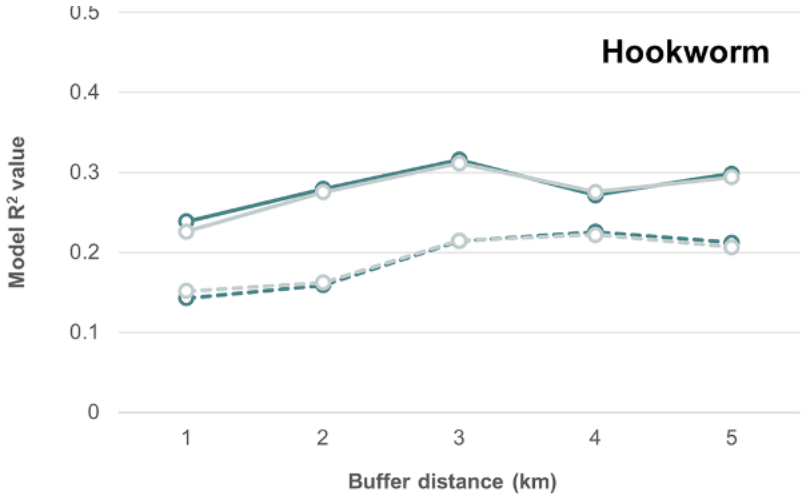
Packages used: caret, randomforest

Improvement due to chemotherapy clear

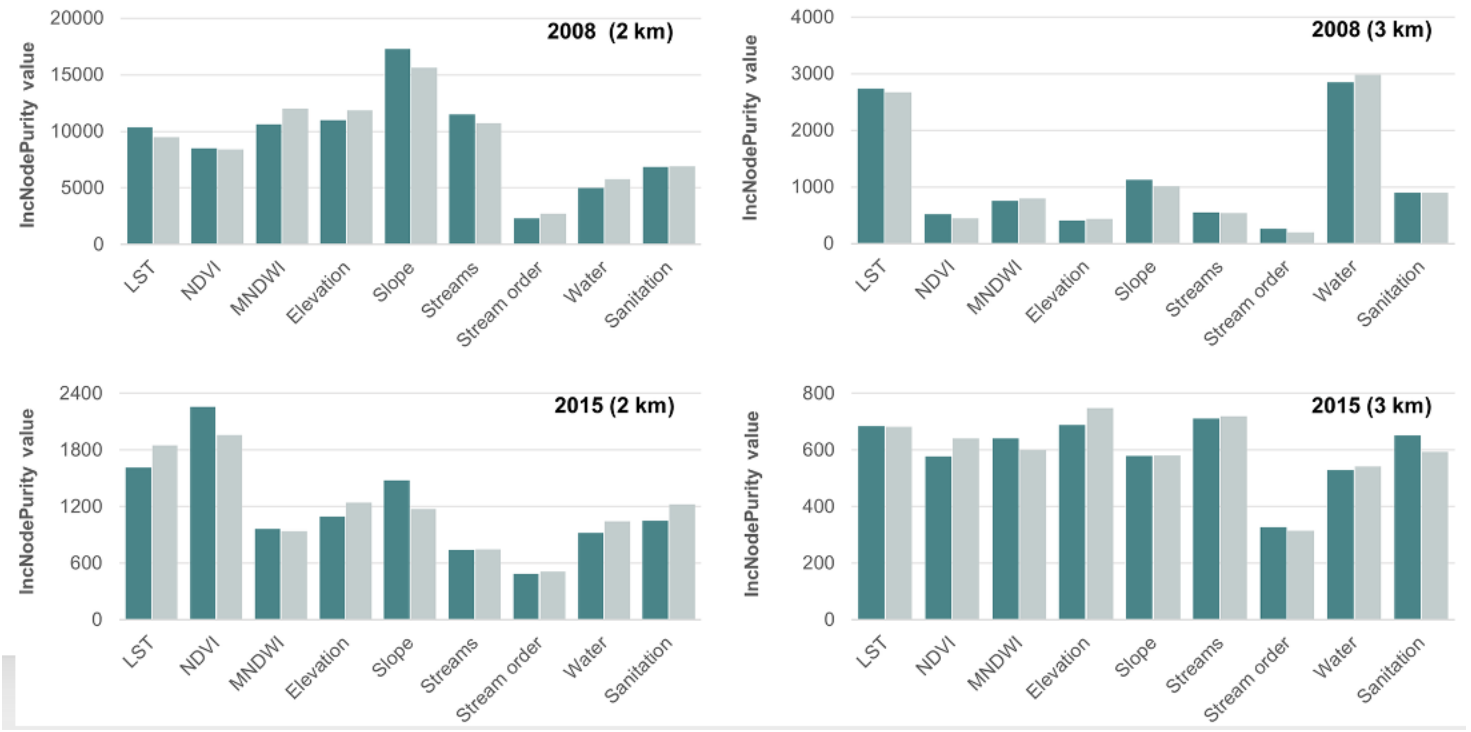


However, random forest models showed better fit for 2008 models as compared to 2015 for both *S. haematobium* and hookworm infections





The relative importance of different variables shifted- why?

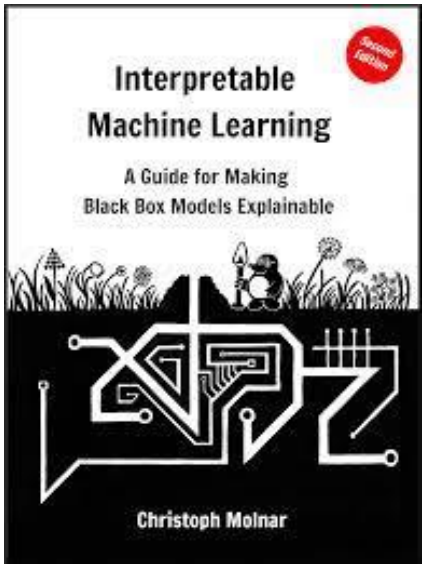


Techniques and algorithms for interpretable machine learning

- Decision trees
- LIME (Local Interpretable Model-Agnostic Explanations)
- SHAP (SHapley Additive exPlanations)
- ICE (Individual Conditional Expectation) and PDP (Partial Dependence Plot)

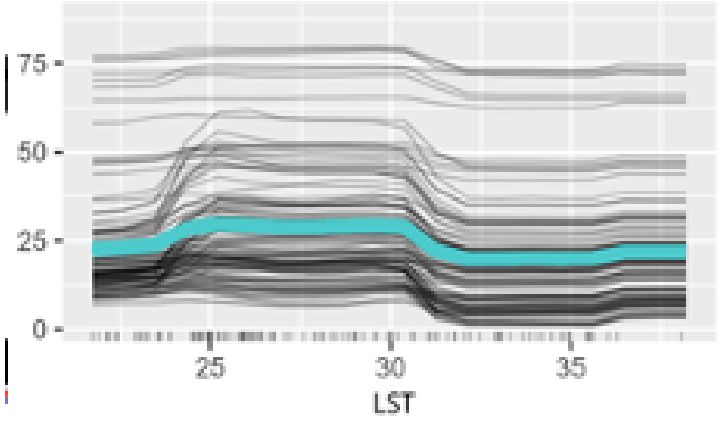
R Package:

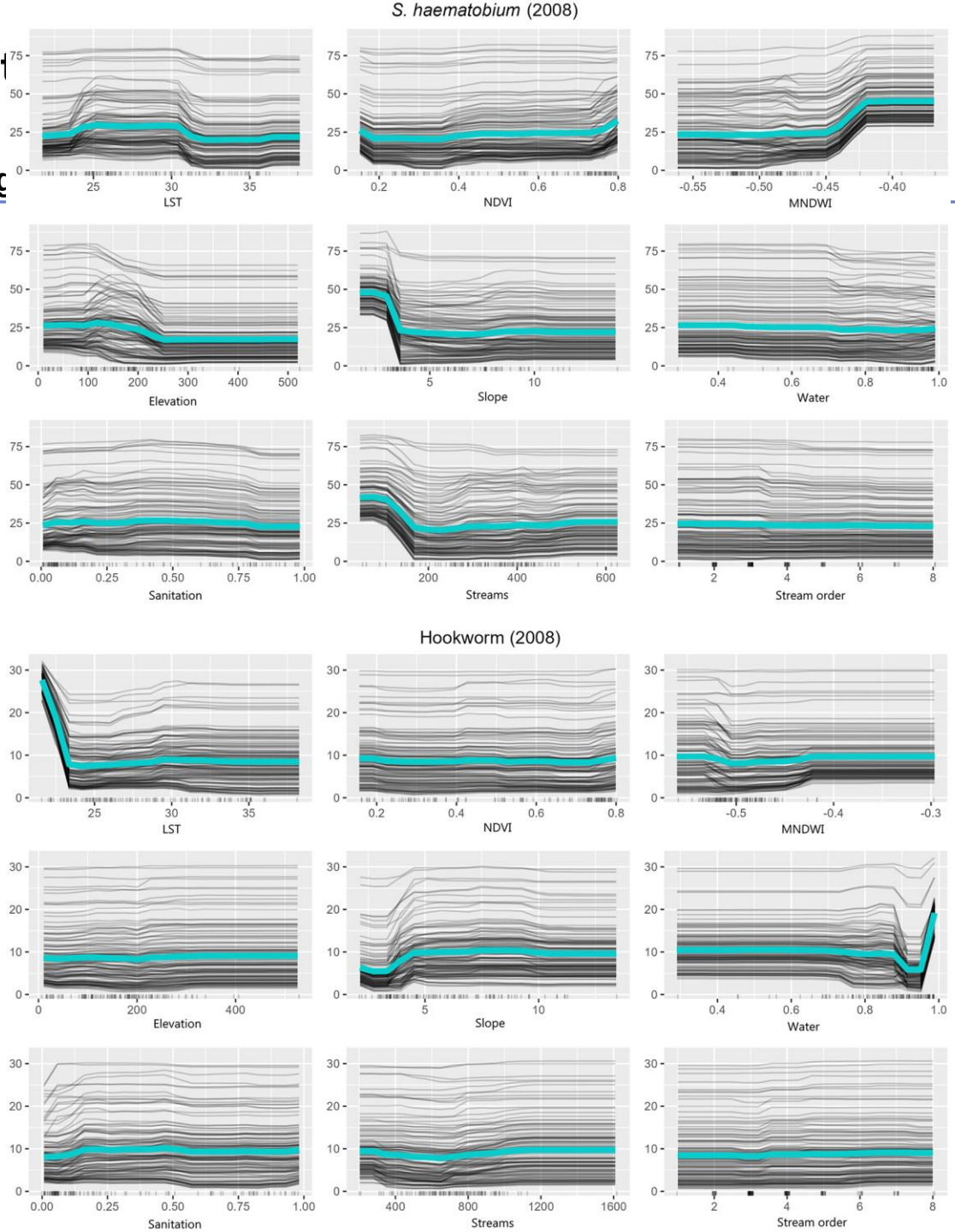
`iml: Interpretable Machine Learning`



Partial Dependence Plot (PDP)

- A PDP shows how a specific feature or variable influences the model's predictions while keeping all other features constant.
- It helps you understand how changes in that variable impact the model's predictions, making it useful for understanding the feature's importance or effect on the model's performance.




S. haematobium (2008)
LST
NDVI
MNDWI
Elevation
Slope
Water
Sanitation
Streams
Stream order
Hookworm (2008)
LST
NDVI
MNDWI
Elevation
Slope
Water
Sanitation
Streams
Stream order

R Code to produce

```
mod <- Predictor$new(SCH_base_2000, data = X, y = sch_base_2000_data$sch_base)


#lst only
eff <- FeatureEffect$new(mod, feature = "lst", method = "pdp+ice", grid.size = 50)
p1 <- eff$plot() + scale_color_brewer(palette = "GnBu") + xlab("LST") + ylab(NULL)

#for multiple features at once
eff <- FeatureEffects$new(mod, method = "pdp+ice")
eff$plot()


mytitle <- expression(paste(italic("S. haematobium"), " (2008)"))
plot(eff) +
  # Adds a title
  plot_annotation(title = mytitle)
```

Interpretable machine learning is key for dealing with complex public health data

IMPROVED DECISION-MAKING.

IDENTIFYING RISK FACTORS AND INTERVENTIONS.

ENHANCED TRUST AND COMMUNICATION.