Statistical methods of global air pollution modeling

University of Utrecht, The Netherlands

Meng Lu

Overview

- Overview

- Statistical learning methods:

  - Regularized linear regression
  - Regression trees and bagging

    random forest; stochastic boosting trees; extreme boosting trees; postprocessing

- Spatiotemporal epidemiology

- Global challenge of air pollution exposure assessment for health research.

- Global NO_2 mapping
  - Current methods used and status
  - Opportunities and the role of statistical learning techniques.

- R scripts, hands-on

Statisitcal learning

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

Enlarge This Image

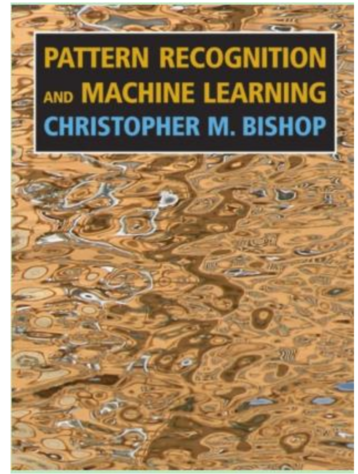
Thor Swift for The New York Times
Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

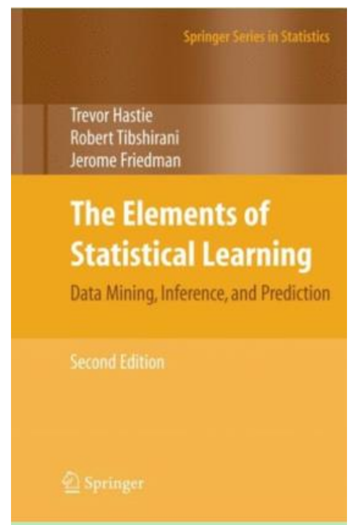
"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

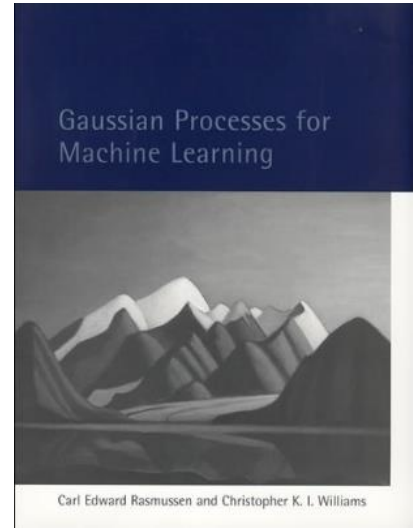
Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for

QUOTE OF THE DAY, NEW YORK TIMES, AUGUST 5, 2009
"I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding."
— HAL VARIAN, chief economist at Google.







Prediction problem:
Finding the best hypothesis

X: space of input values
Y: space of output values

Given a dataset $D \in X \times Y$, find a function (hypothesis)

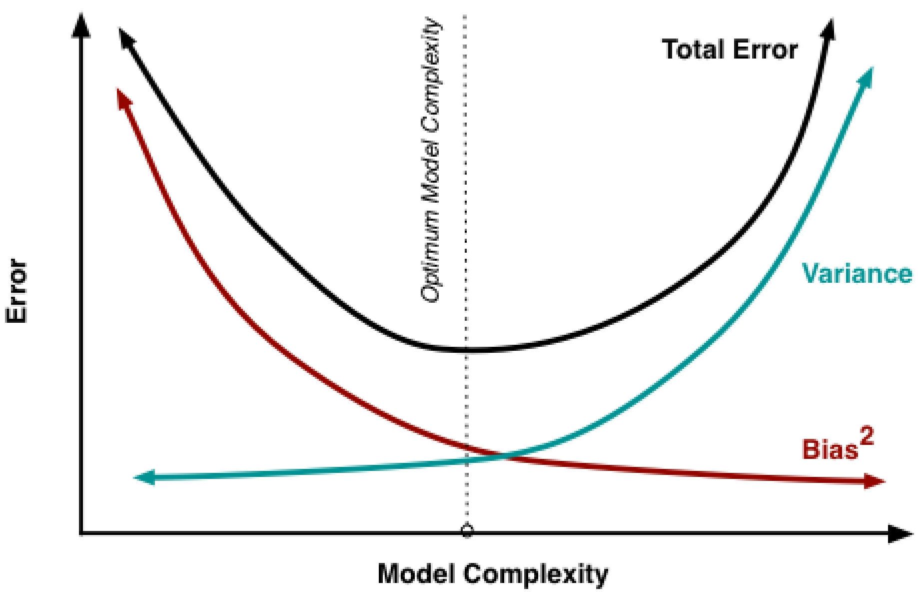
$$h: X \rightarrow Y$$

Y : categories; continuous data, graphic output

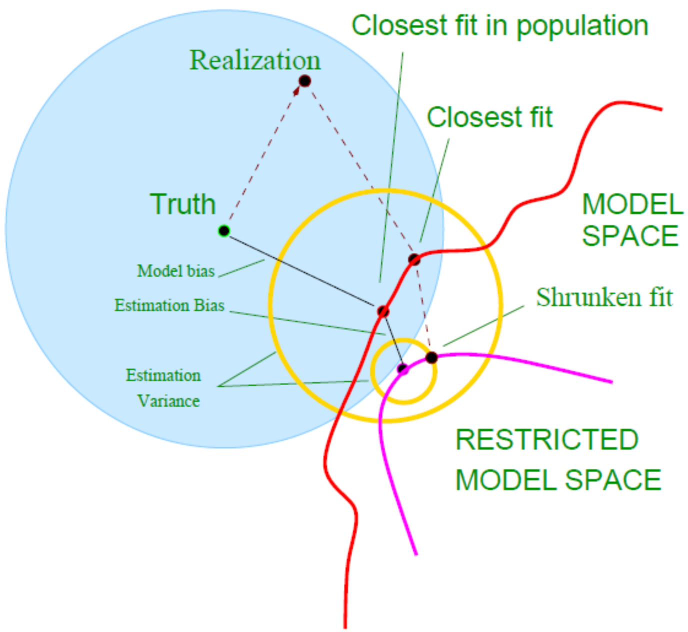
Bias-variance trade-off

$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2$$

$$Err(x) = \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{Irreducible\ Error}$$



All algorithms are affected by bias-variance trade-off

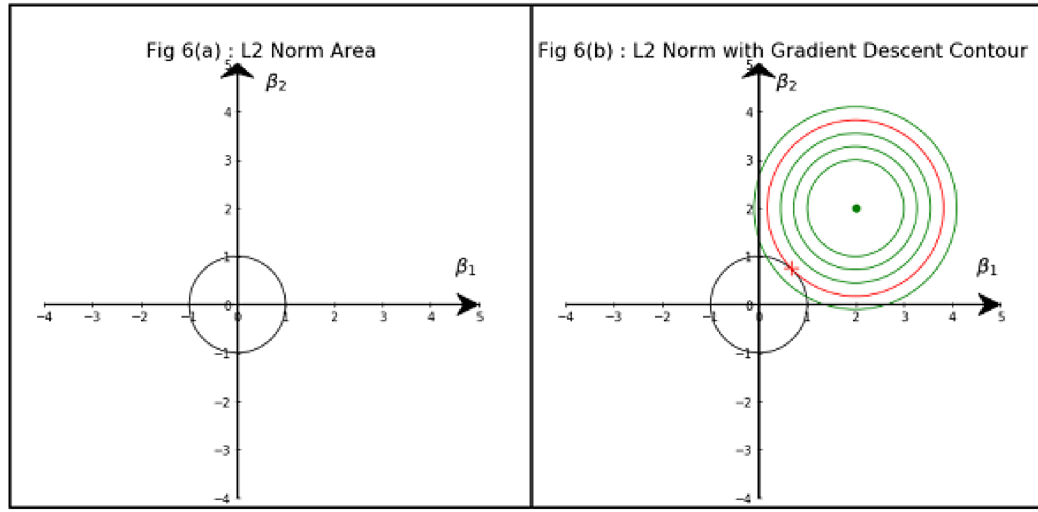


Schematic of the behavior of bias and variance.

Regularization

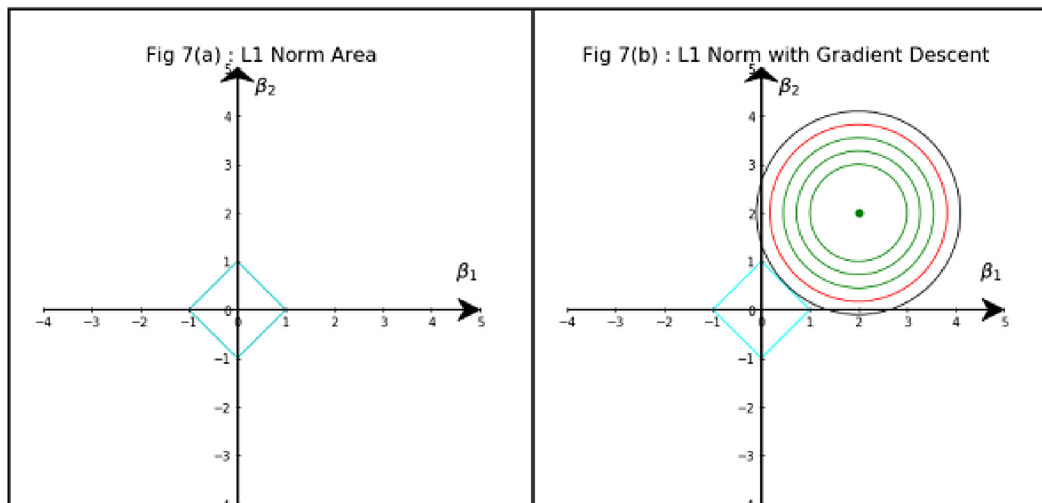
Ridge regression

$$L_{hridge}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{m} w_j \hat{\beta}_j^2 .$$

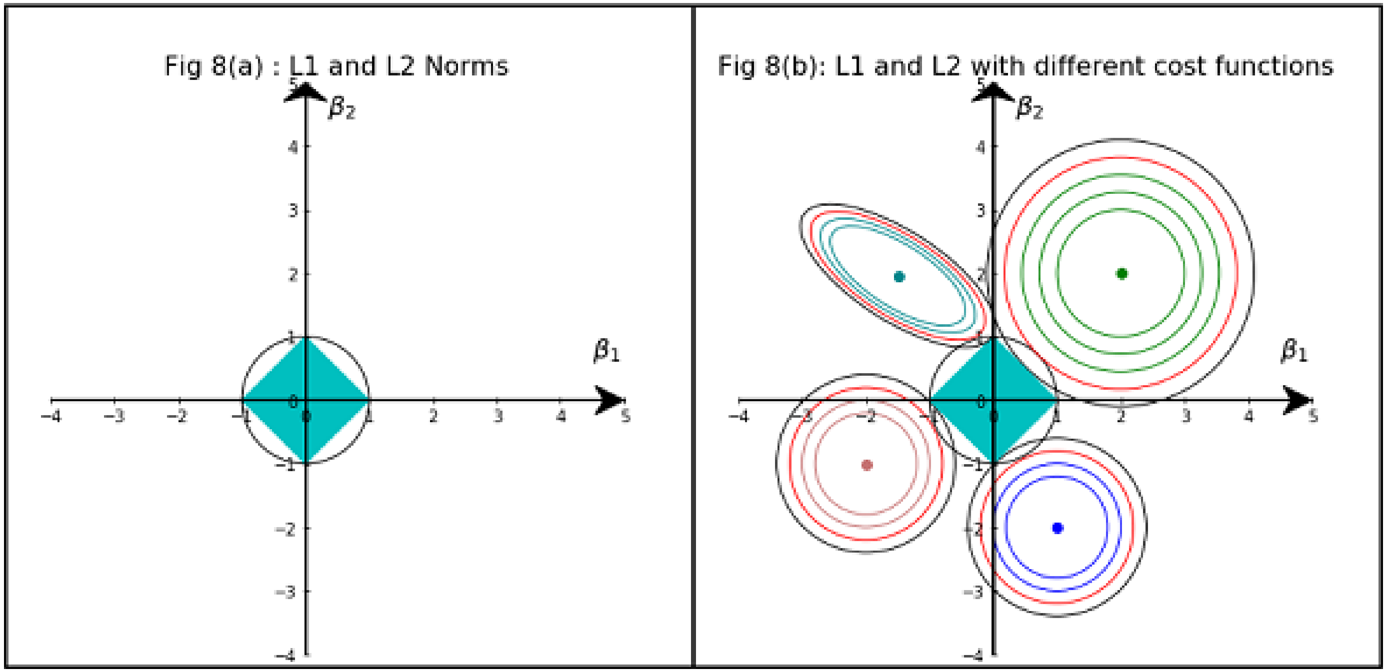


Lasso regression

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{m} |\hat{\beta}_j| .$$



Lasso vs. Ridge
ElasticNet



Regression trees

Features:

- Non-parametric
- Different kinds of variables
- Redundant variables are ignored
- Handle missing data
- Small trees are easy to interpret

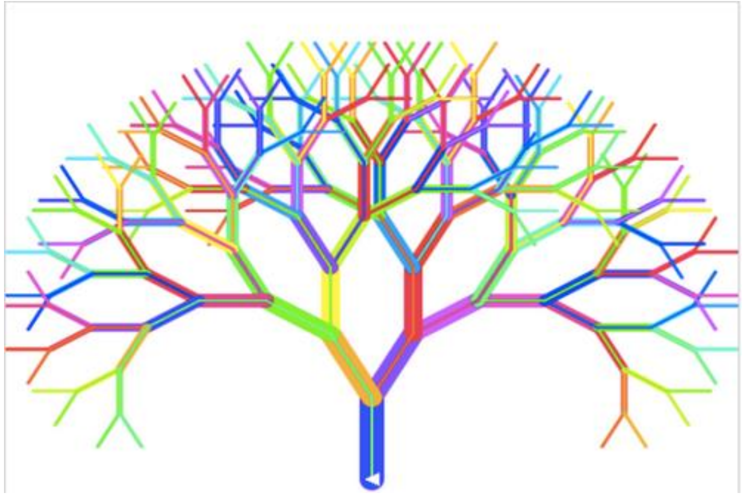
Leveraging trees to improve the performance:

- Bagging
- Boosting
- Random forest

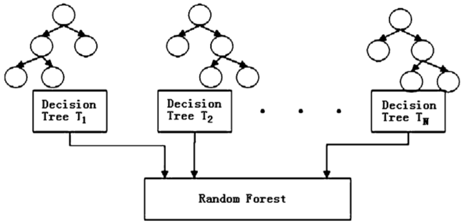
Dominance

Boosting > Randomforest > bagging > single tree

Is it true that boosting trees are always better than the randomforest?



Random forest



variance reduction

Identically distributed variables, each has variance σ
An average of B of i.i.d random variables has variance $\frac{1}{B}\sigma^2$

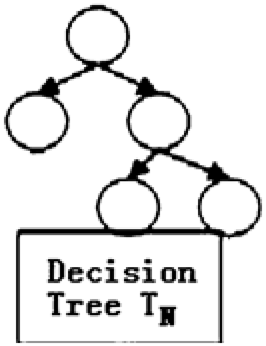
If the variables are not independent (but identically distributed) with positive pairwise correlation p , the variance of the average:

$$\mathrm{p}\sigma + \frac{1-p}{B}\sigma^2$$

"the more uncorrelated, the more you bringing down the variance".

(tunning parameter: number of trees, tree depth)

More details



For each tree:

1. Bootstraping sample D* from the training data D

2. Draw **m*** variables randomly from all variables m, pick the best split-point (variable), split the node.

Limitation:
Bias towarding variables with many splits or missing variables, does not assess uncertainty

Variations:

Recursive partition trees:
Hypothesis testing of dependency between variables and resursively fitting the splitting weight for 2

Baysian based sampling and variable selection:
Baysian framework for 1 and 2

Quantile random forest:
Estimate quantiles (beyond the conditional mean)

Stochastic gradient Boosting (regression)

-- Reweight based on the previous trees, stage-wise fitting

Each successive tree is built for the prediction residuals of the preceding tree in an adaptive way to reduce bias.

-
```
initial:
r = y
fit a regression tree to r: g(x)

for each tree:
f(x) = e*g(x)
r = r - f(x)
```

(r: residual; e: learning rate)

Gradient boosting: Greedy Function Approximation: A Gradient Boosting Machine. Friedman

XGboost
Exteme gradient boosting

Idea
Not only impurity, but also model complexity

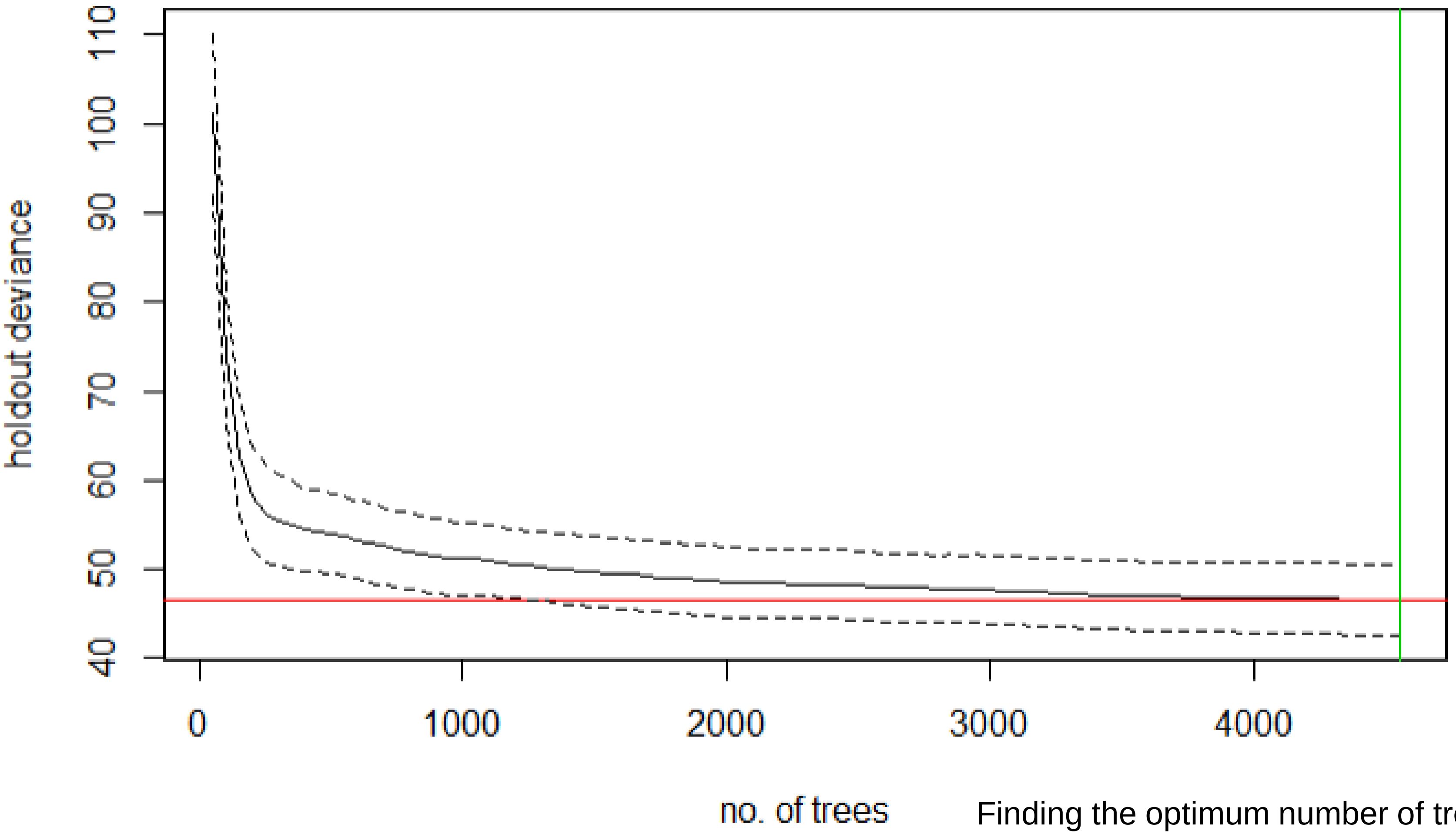
$$\text{obj}(\theta) = L(\theta) + \Omega(\theta)$$

Features

- Parallel computation
- Support dense and sparse matrix
- Can costomize objective fucntions

Cross validation

-- Automatically determine the tunning parameters:



Finding the optimum number of trees

Postprocessing

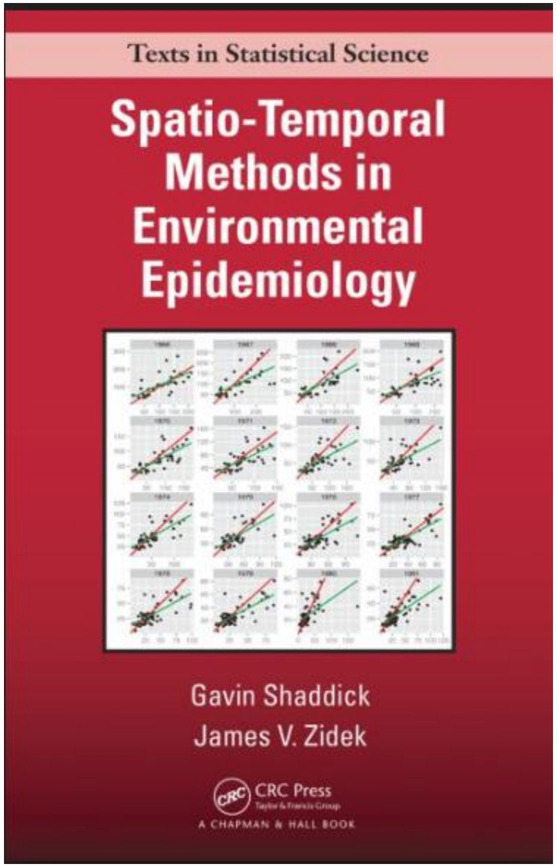
Lasso regularization of regression trees
--- discarding trees that are not useful

$$\alpha(\lambda) = \arg\min_{\alpha} \sum_{i=1}^{N} L[y_i, \alpha_0 + \sum_{m=1}^{M} \alpha_m T_m(x_i)] + \lambda \sum_{m=1}^{M} |\alpha_m|.$$

Spatiotemporal epidemiology

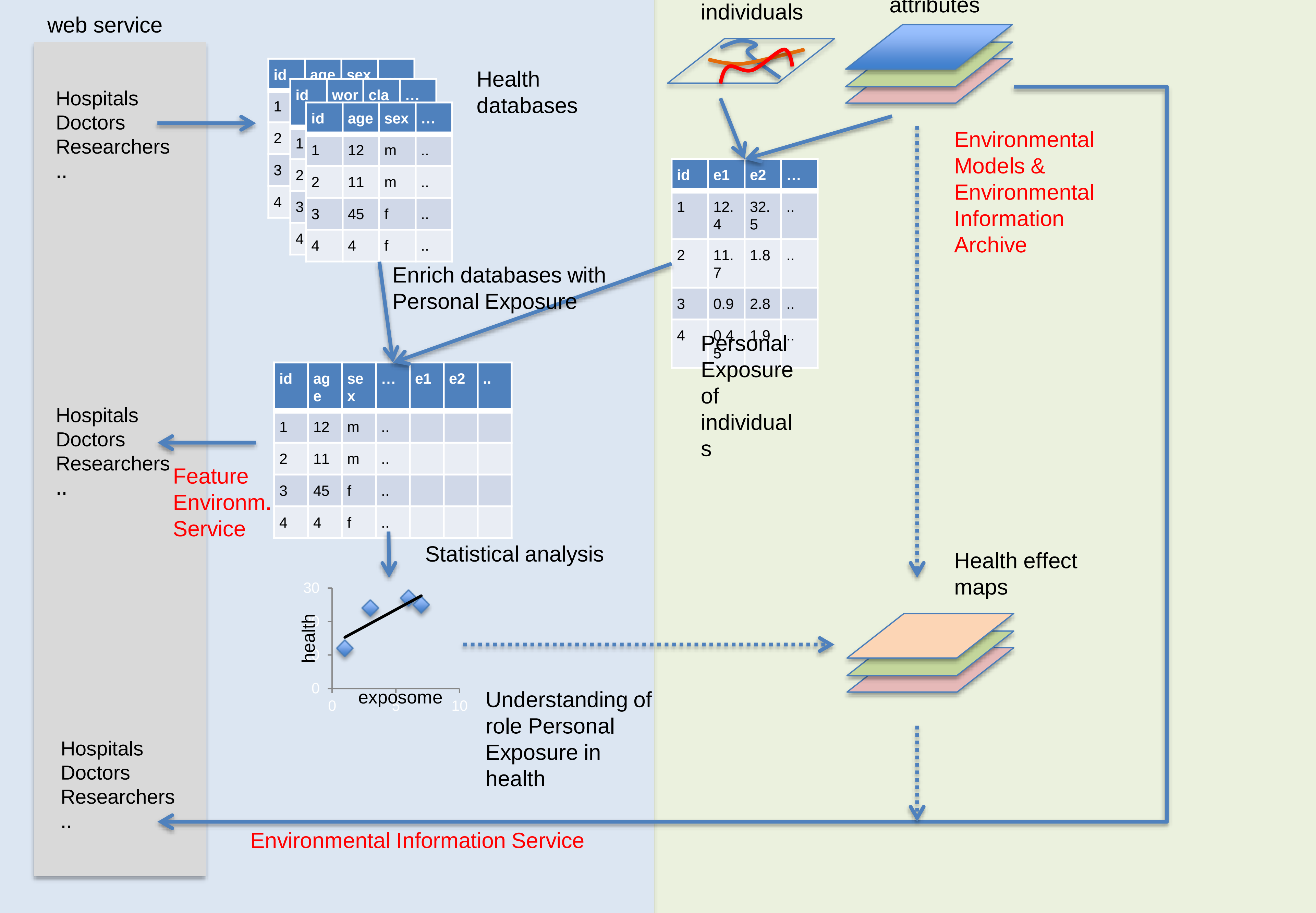
The description and analysis of geographical data, specifically health outcome data and factors that may explian variations in these outcome data over space.
Factors: demographic, environmental, genetic, infectious risk factors

1854 John Snow, cholera
Identify possible causes of outbreaks.



Bayesian framework, R-INLA

Environmental epidemiology



Air pollution

-- a major health risk factor and global challenging

Air pollution:

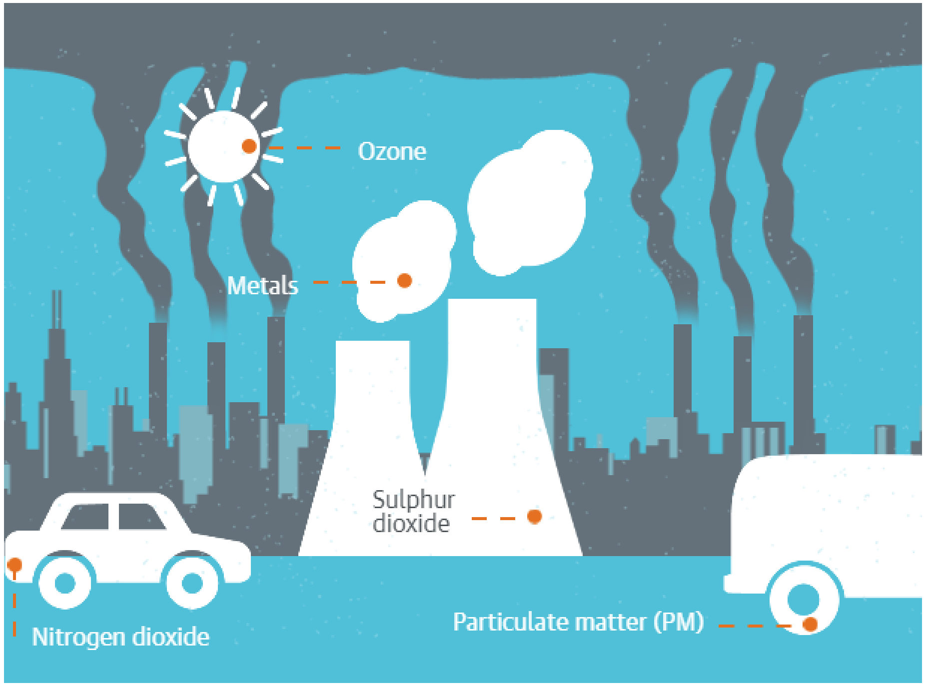
Consists of chemicals or particles in the atmosphere that poses health and environmental threats.

Mortality:

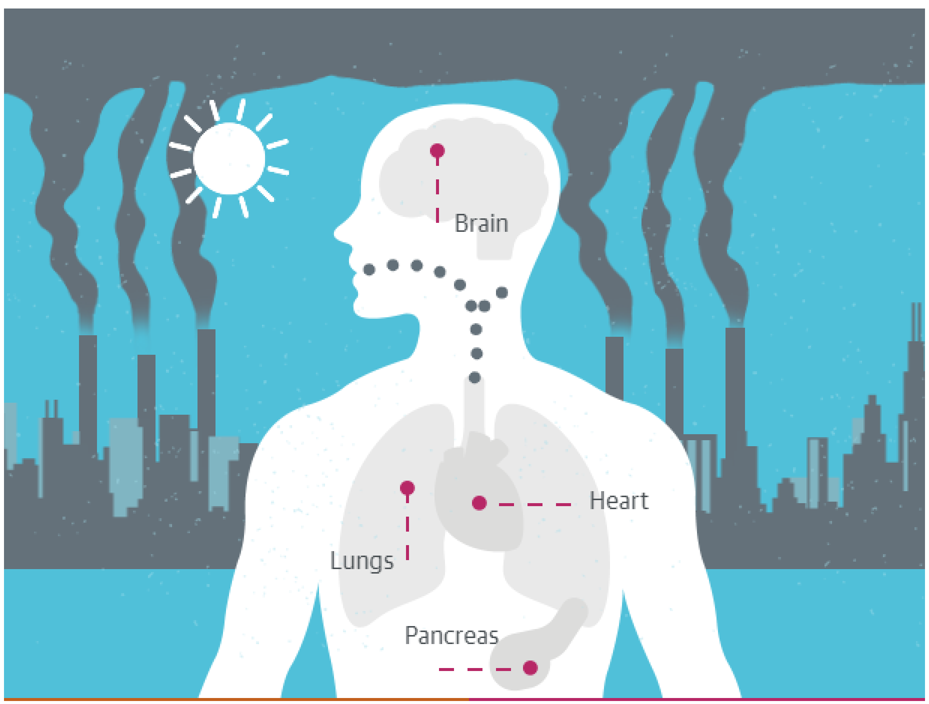
World: more than 3.2 millions death a year

Europe: 420,000 premature death every year

Most measured air pollutants and their health impacts

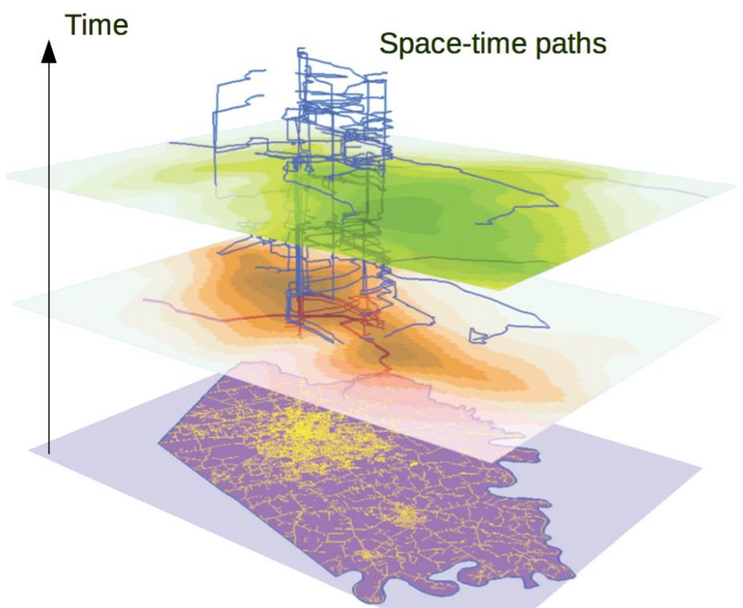


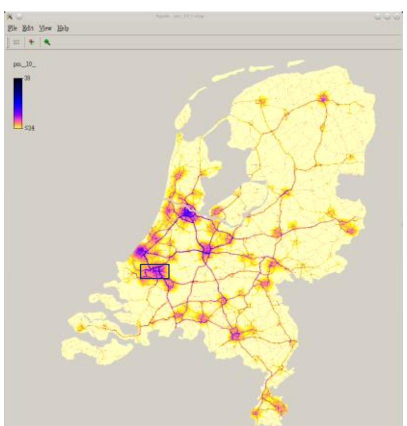
Source [2]

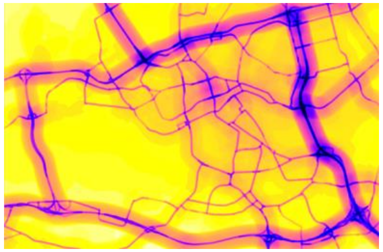


Source [2]

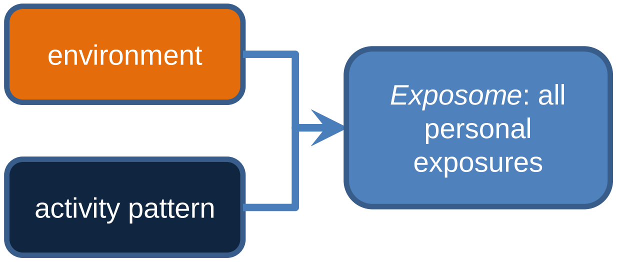
Quantifying the exposome







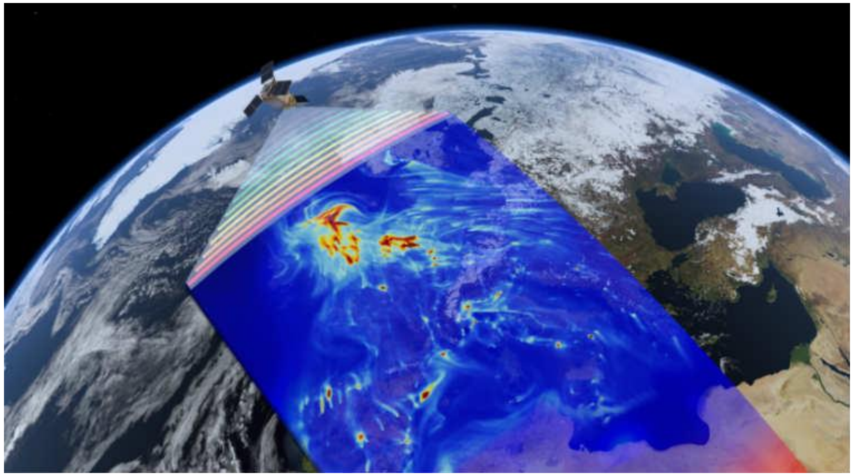
Aggregating exposure to the environmental variables along space-time paths



When detailed space-time paths are not available, exposure assessment techniques are used that assume a particular space-time behavior of a person.

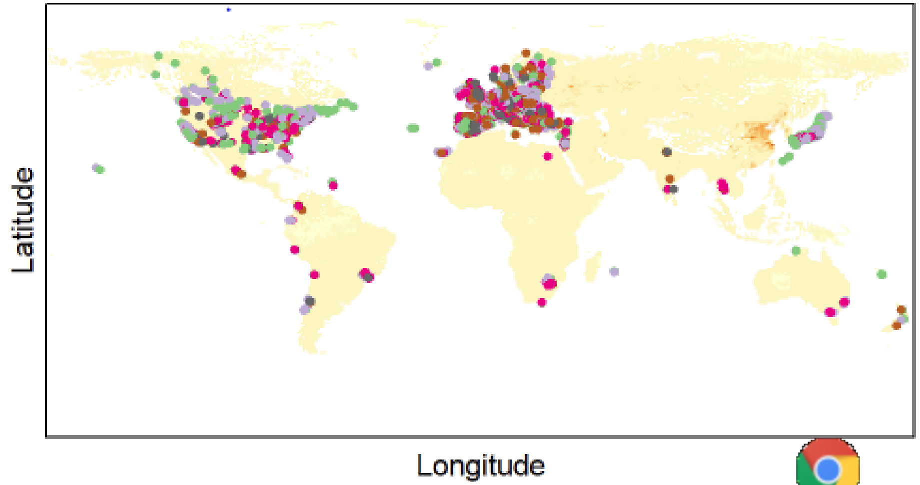
Global air pollutant measurements

- Remote sensing data:



Source[2]

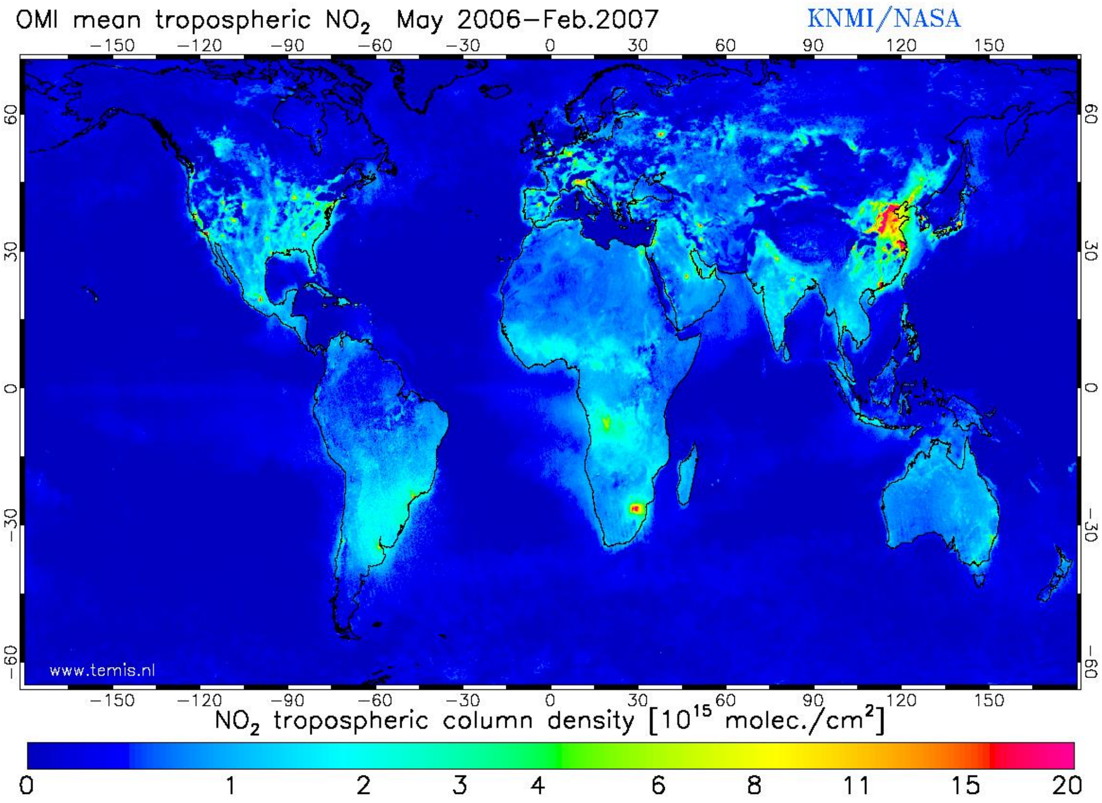
- Station measurements



NO2mean.html

	Remote sensing data	Station measurements
Data type	field (spaital continuous)	point (spatial discrete)
Spatial resolution	coarse	high
Temporal resolution	Low (trade-off with spatial resolution and coverage)	high
Global coverage	wide	limited

Remote sensing measurements:
OMI (Ozone Monitoring Instrument)



Date of Launch 15 July 2004

Swath Width 2600 km

Spectral bands: ultraviolet and visible (270 to 500 nm)

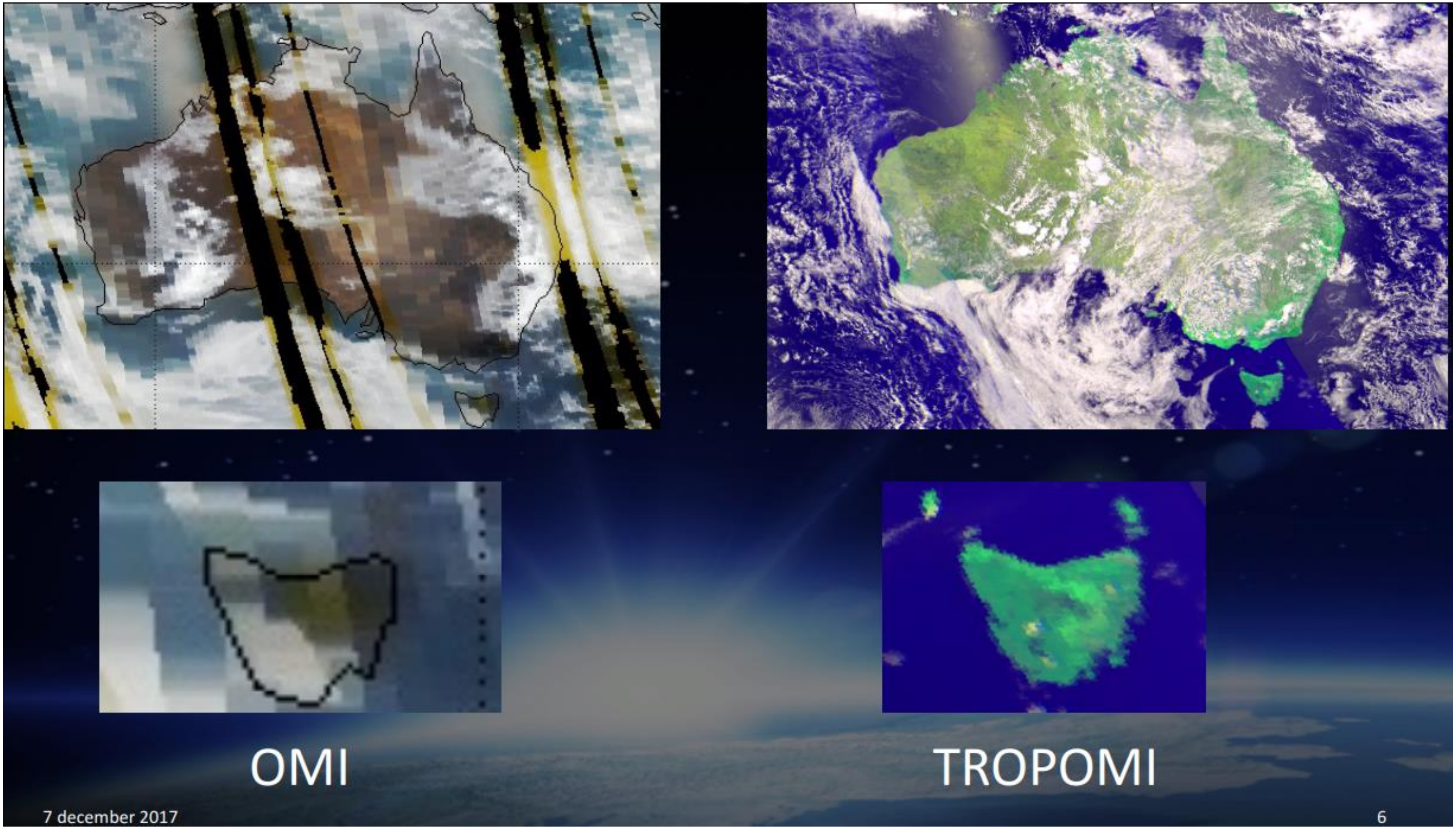
At nadir 13 km× 24 km

Zoom in mode 13 km× 12 km

Daily global coverage

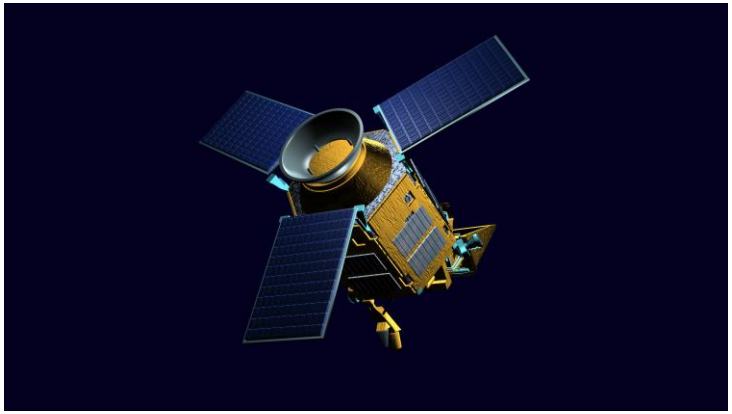
Tropomi

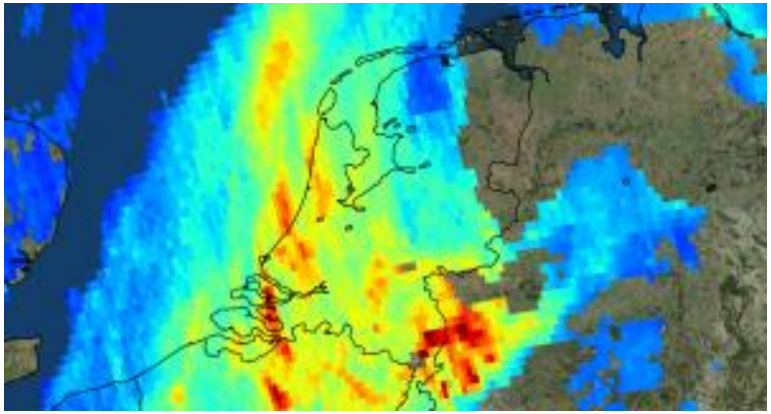
launched 2017, available from Feb 2018



Souce [3]

Tropomi





NO2, O3, SO2, methane and CO

Spectral bands:
ultraviolet and visible (270–500 nm), near-infrared (675–775 nm), shortwave infrared (2305–2385 nm) spectral bands.

Resolution:
7 km x 7 km
zoom in mode: 7 km × 3.5 km

Spectral bands of Tropomi

Product	Spectrometer	Application
Ozone	UV, UVIS	Ozone layer monitoring, UV-index forecast, Climate monitoring
NO_2	UVIS	Air quality forecast and monitoring
CO	SWIR	Air quality forecast and monitoring
CH_2O	UVIS	Air quality forecast and monitoring
CH_4	SWIR	Climate monitoring
SO_2	UVIS	Air quality forecast and monitoring, Climate monitoring, Volcanic plume detection
Aerosol	UVIS, NIR	Air quality forecast and monitoring, Climate monitoring, Volcanic plume detection
Clouds	UVIS, NIR	Climate monitoring
UV-Index	UVIS	UV index forecast

TROPOMI	UV		UVIS		NIR		SWIR	
Band	1	2	3	4	5	6	7	8
Spectral coverage [nm]	270 – 320		320 – 495		675 - 775		2305 – 2385	
Full spectral coverage [nm]	267 - 332		303 - 499		660 - 784		2299 - 2390	
Spectral resolution [nm]	0.49		0.54		0.38		0.25	
Spectral sampling ratio	6.7		2.5		2.8		2.5	
Spatial sampling [km^2]	7 x 28	7 x 3.5				7 x 3.5	7 x 7	

LUR modeling: Land use regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Sensor measurements:

Station measurements

Mobil sensors

Remote sensing measurements:

OMI (250 km)
Tropomi (8 km)
…

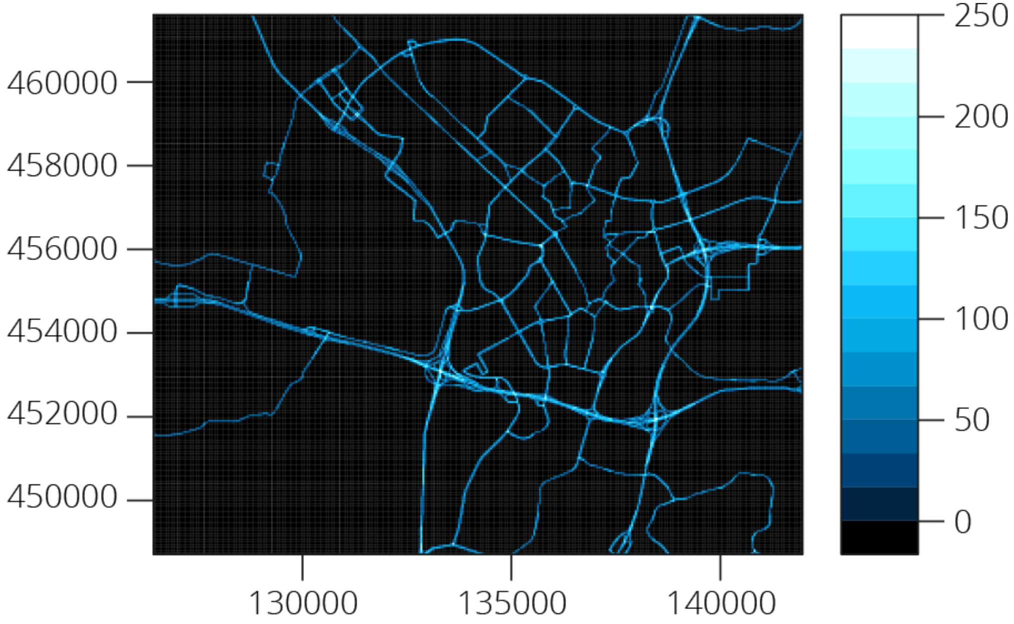
GIS predictors:

Population
Road length within a buffer
Distance to roads
Traffic load
…

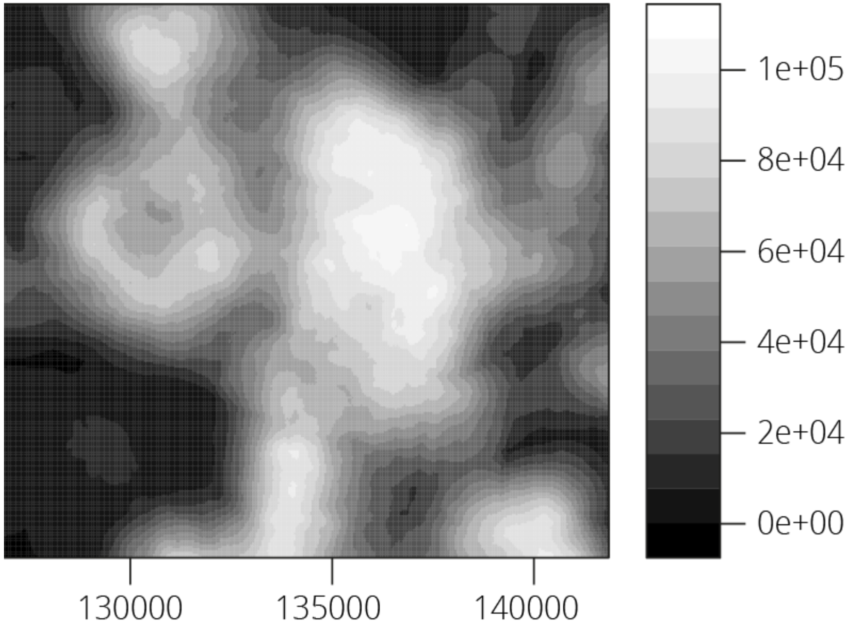
Example of predictors (independent variables): variables within different buffer sizes

Major road length

25m buffer

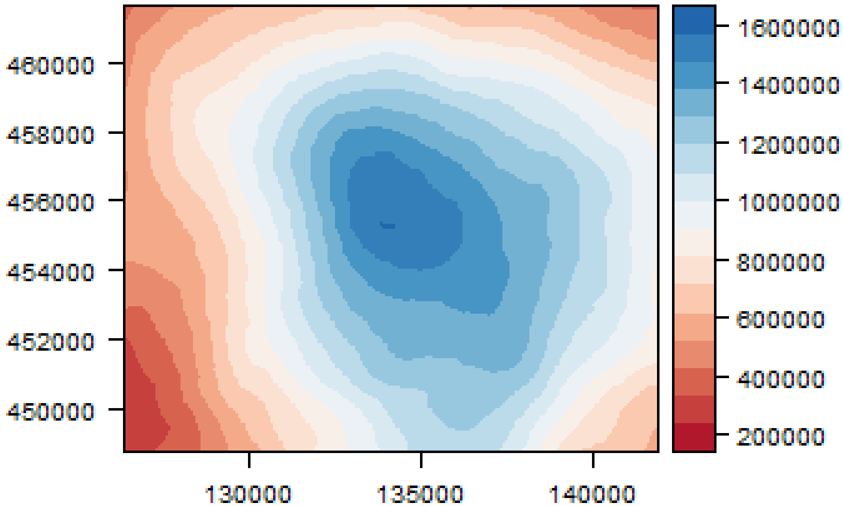


1000m buffer



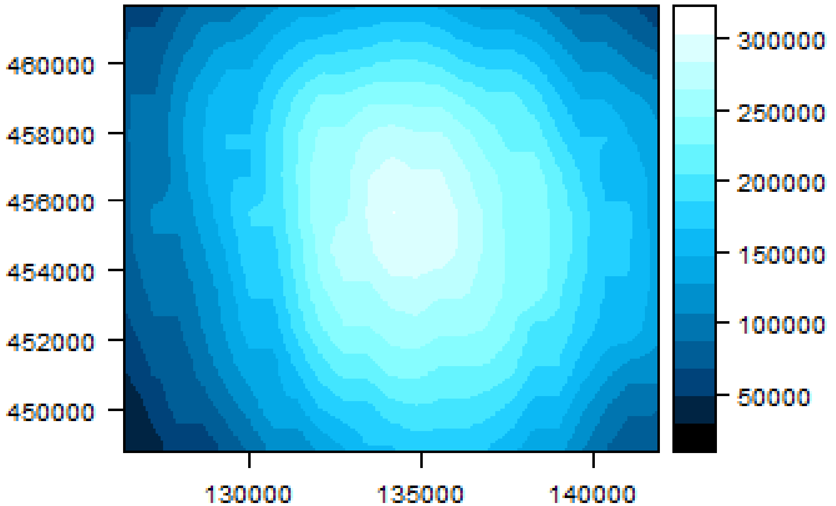
predictors: background information

Major road length
5000m buffer.

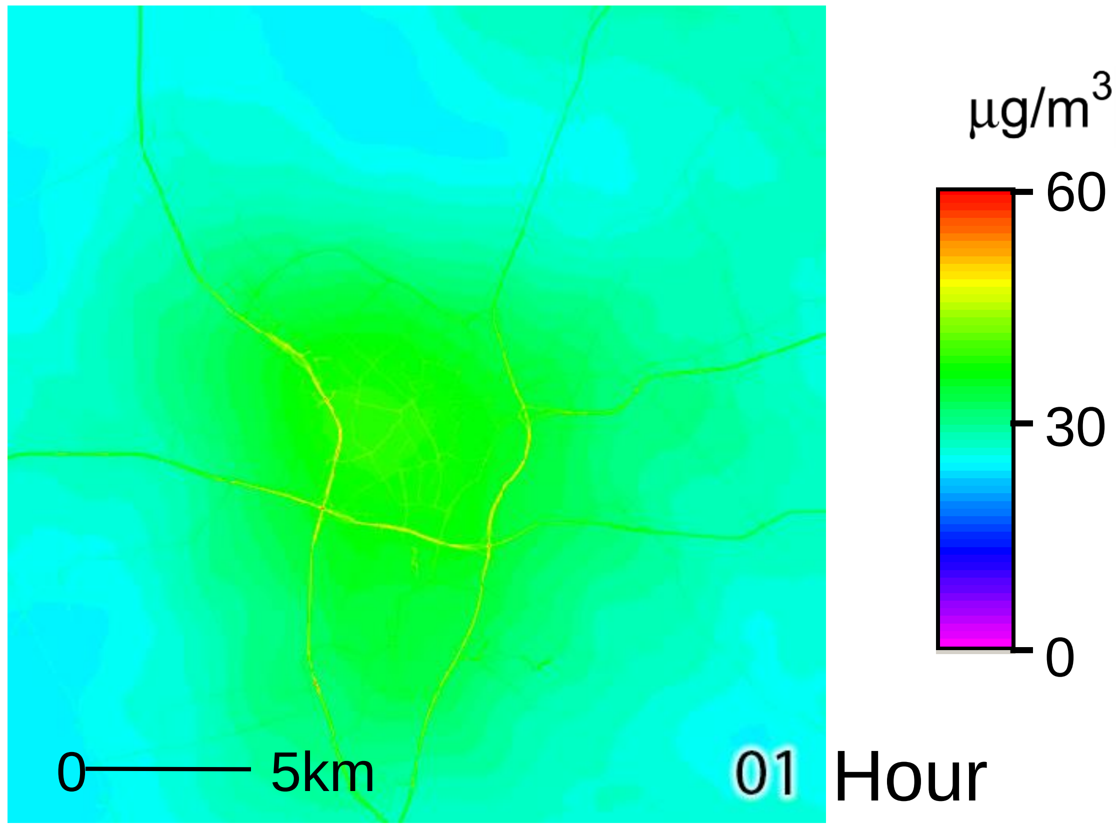


Correlation:
0.9679304

Population
5000m buffer.



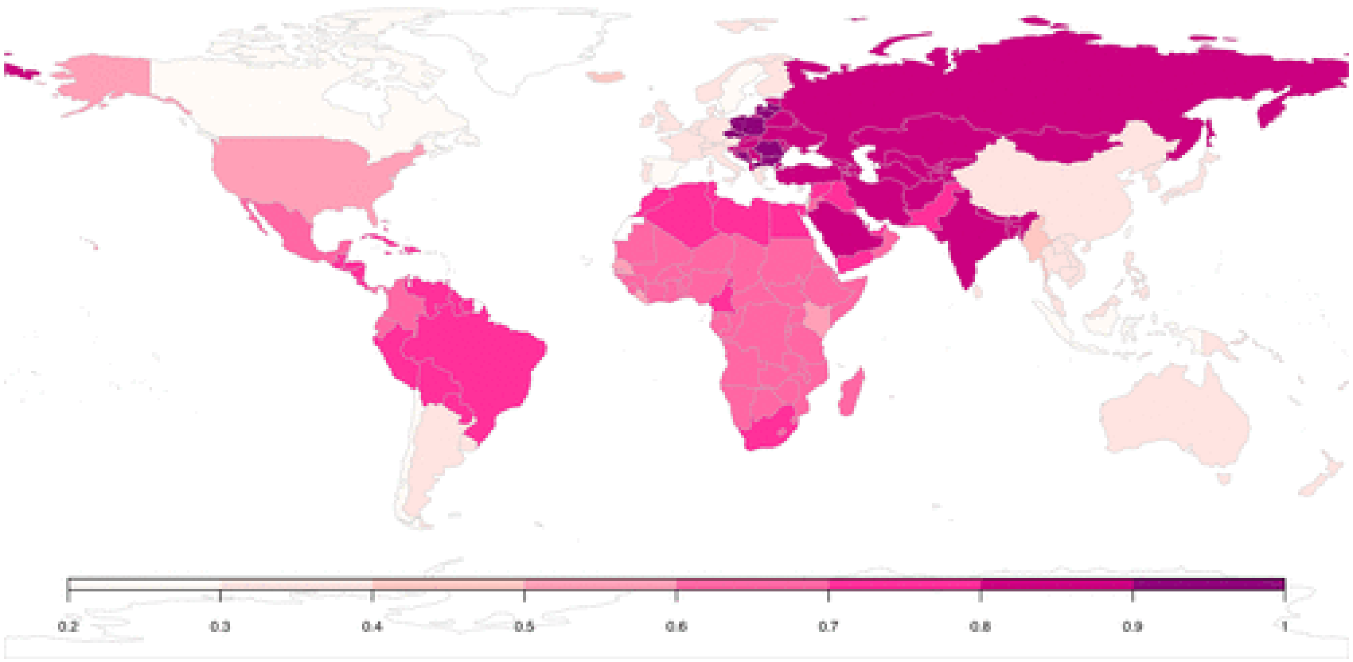
Spatiotemporal dynamic of air pollution showing road effects



5 m resolution

Problem: nonlinear relationship over global scale

spatially varying coefficients



Source: shaddock et al., 2018

GGHDC project: Global air pollution prediction

Station measuremnt

More than 3000 stations globally
2017, annual mean, separate day and night

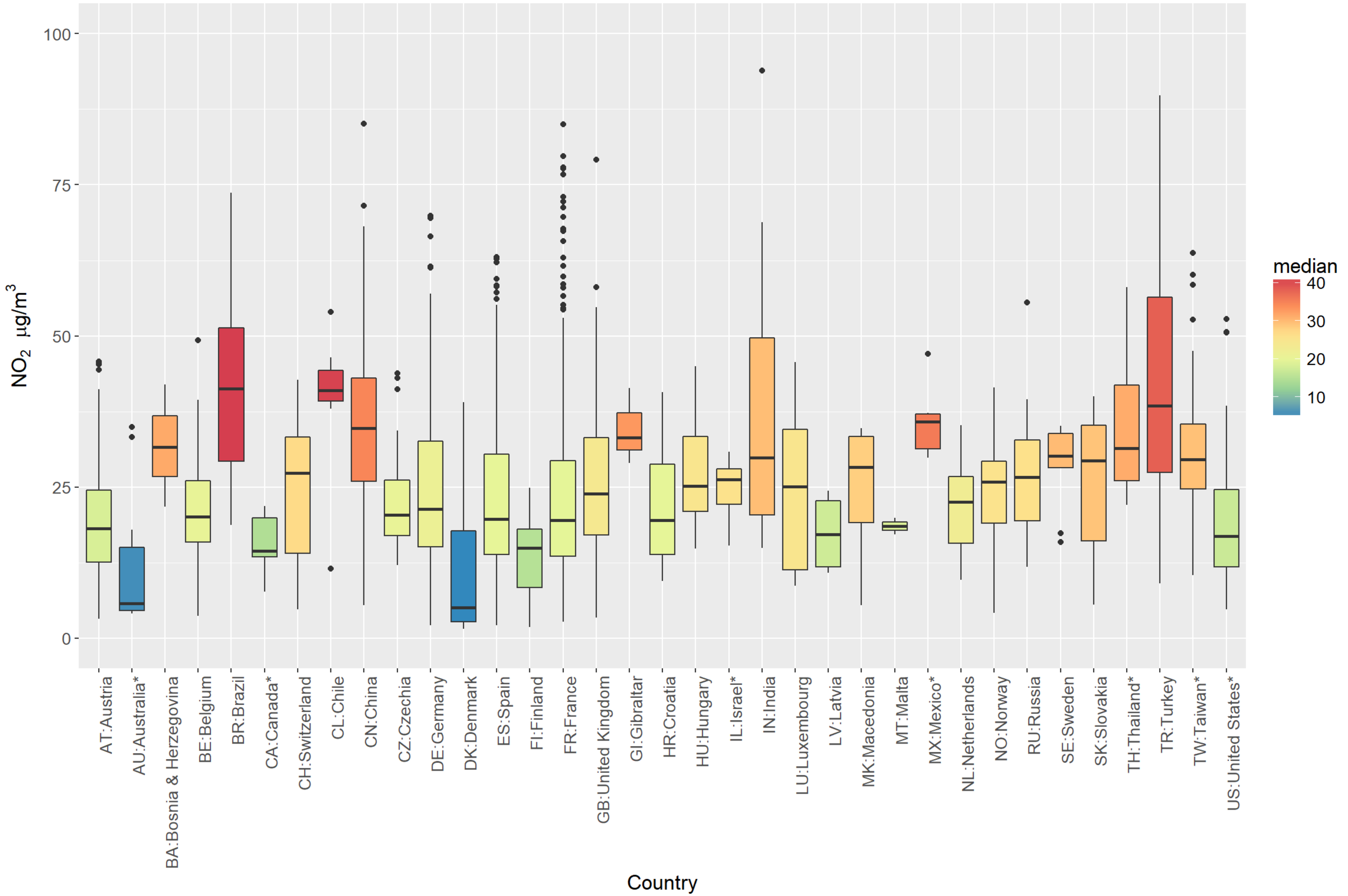
Predictors

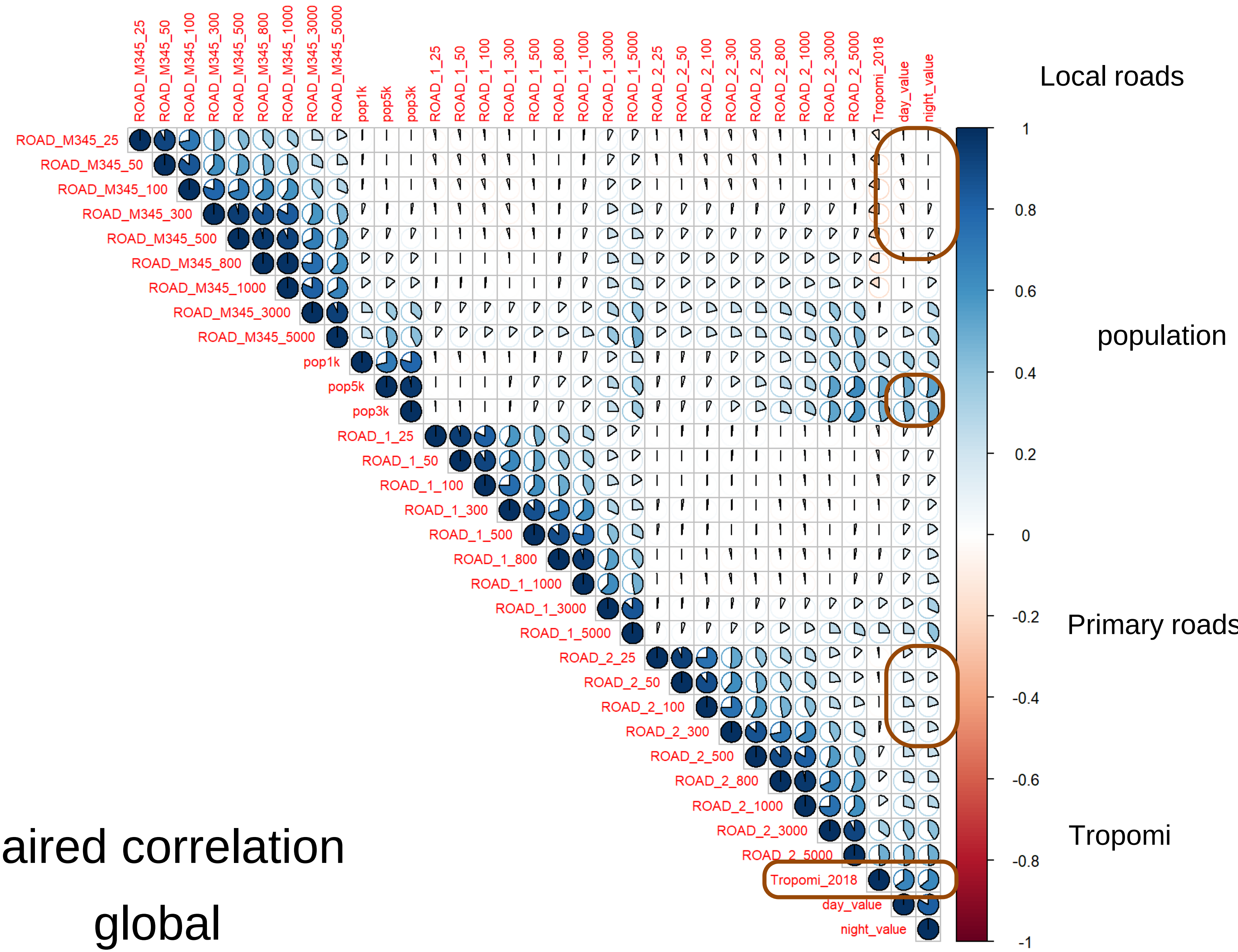
Variables in different buffers:

Road length: 25m - 5km
Highway, primary roads, secondary roads, local roads
Population: 1km, 3km, 5km
Industry area 25m - 5km

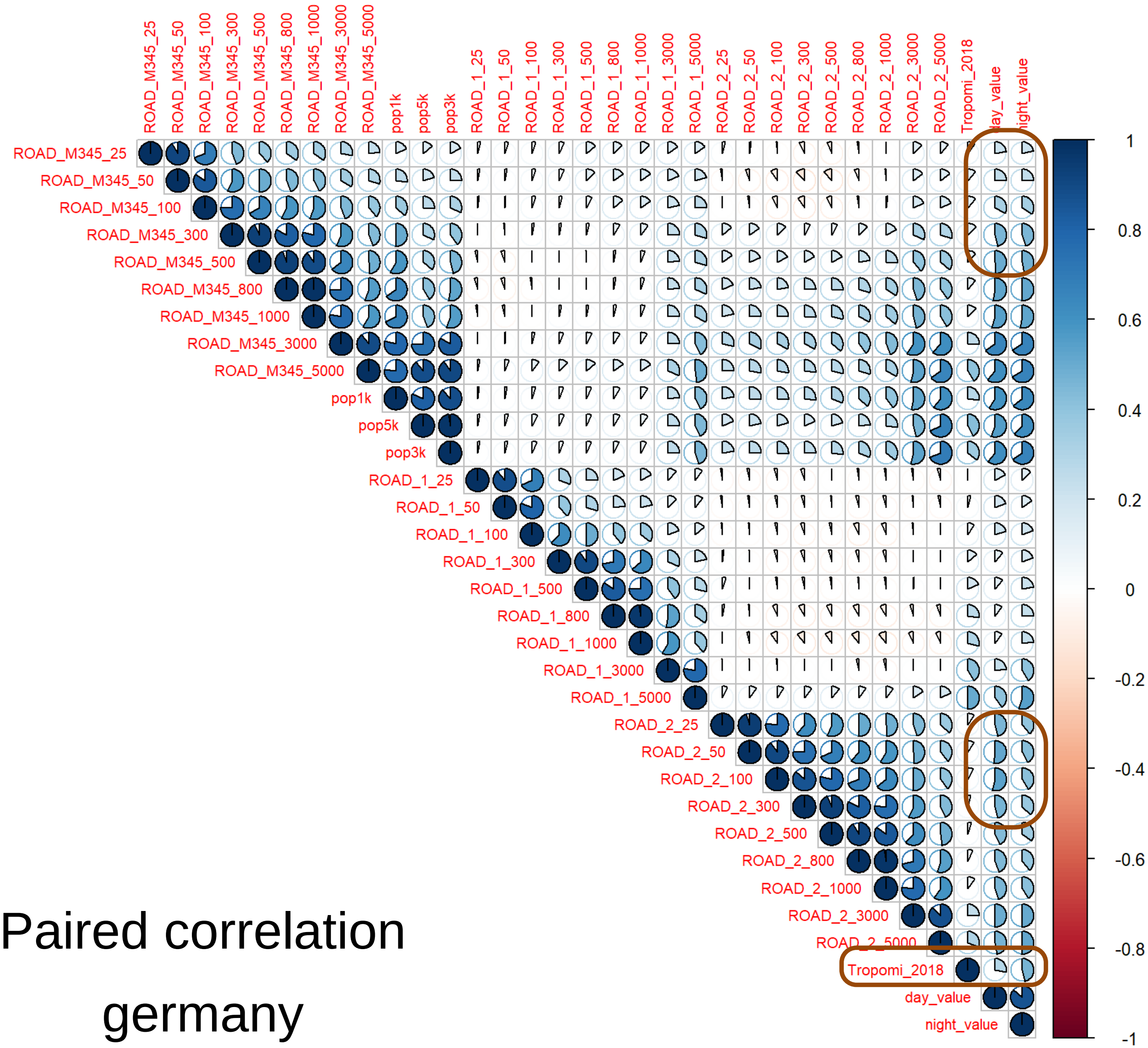
Points and Coverages
Monthly wind speed (0.5 degree)
Monthly temperature   (0.5 degree)
Surface concentration from Remote sensing products and physical models
Remote sensing measurements of NO2 column density
Distance to coast

NO2 of different countries





Paired correlation global



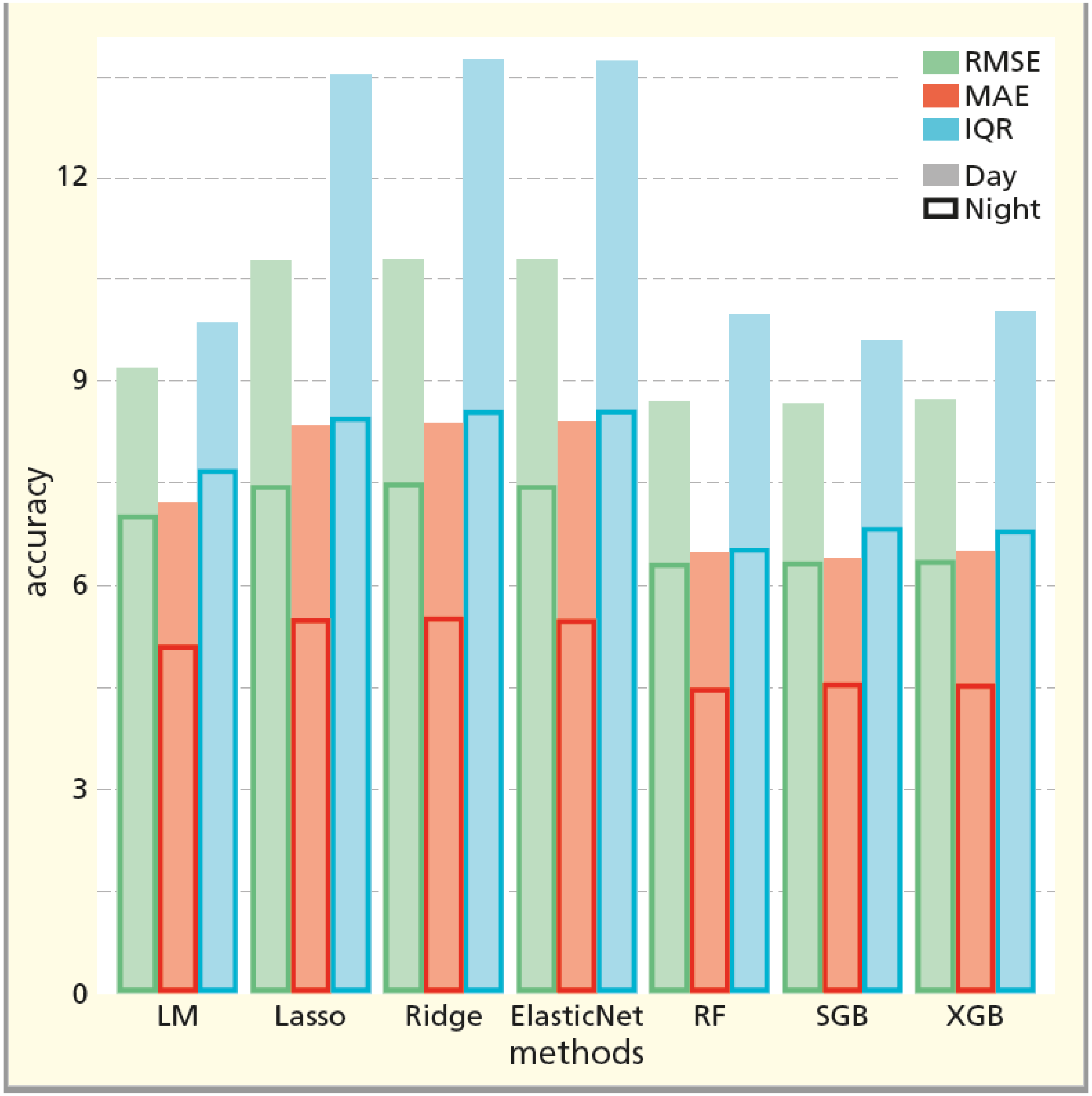
Local roads

Primary roads

Tropomi

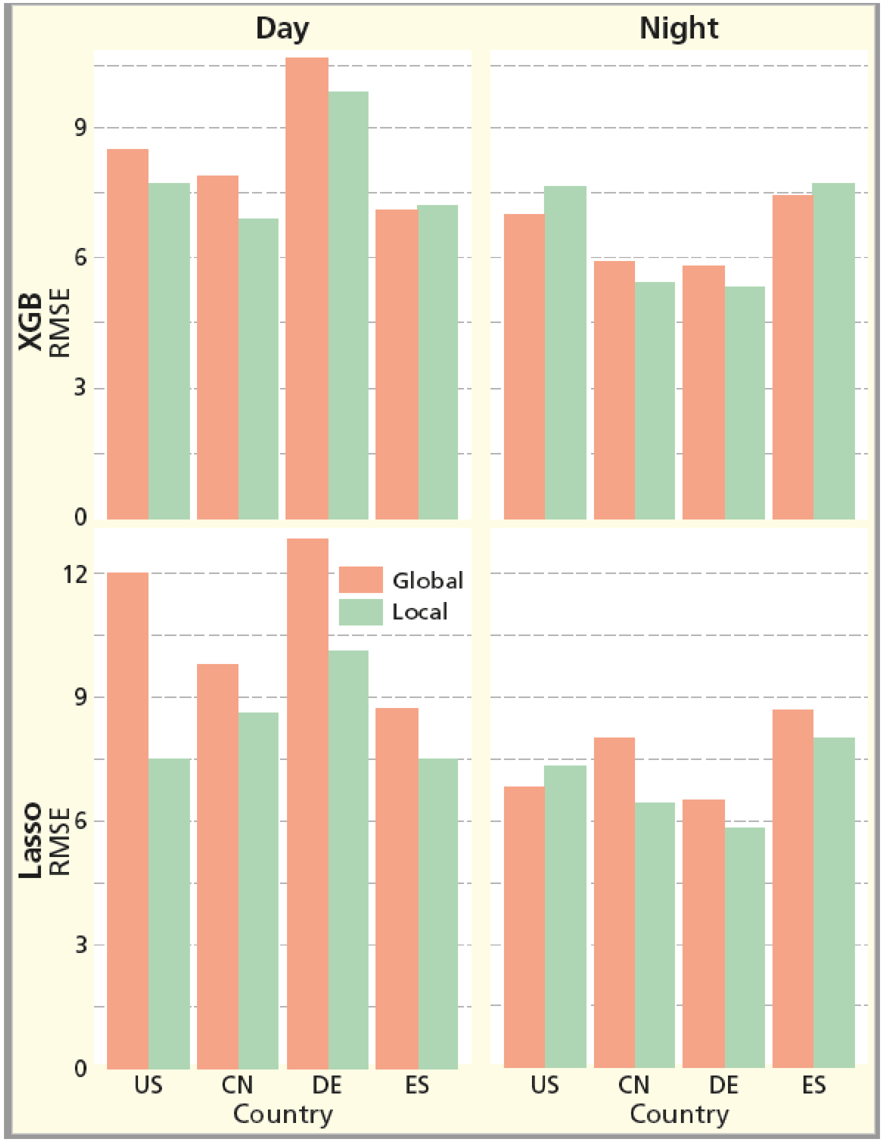
Paired correlation germany

Result

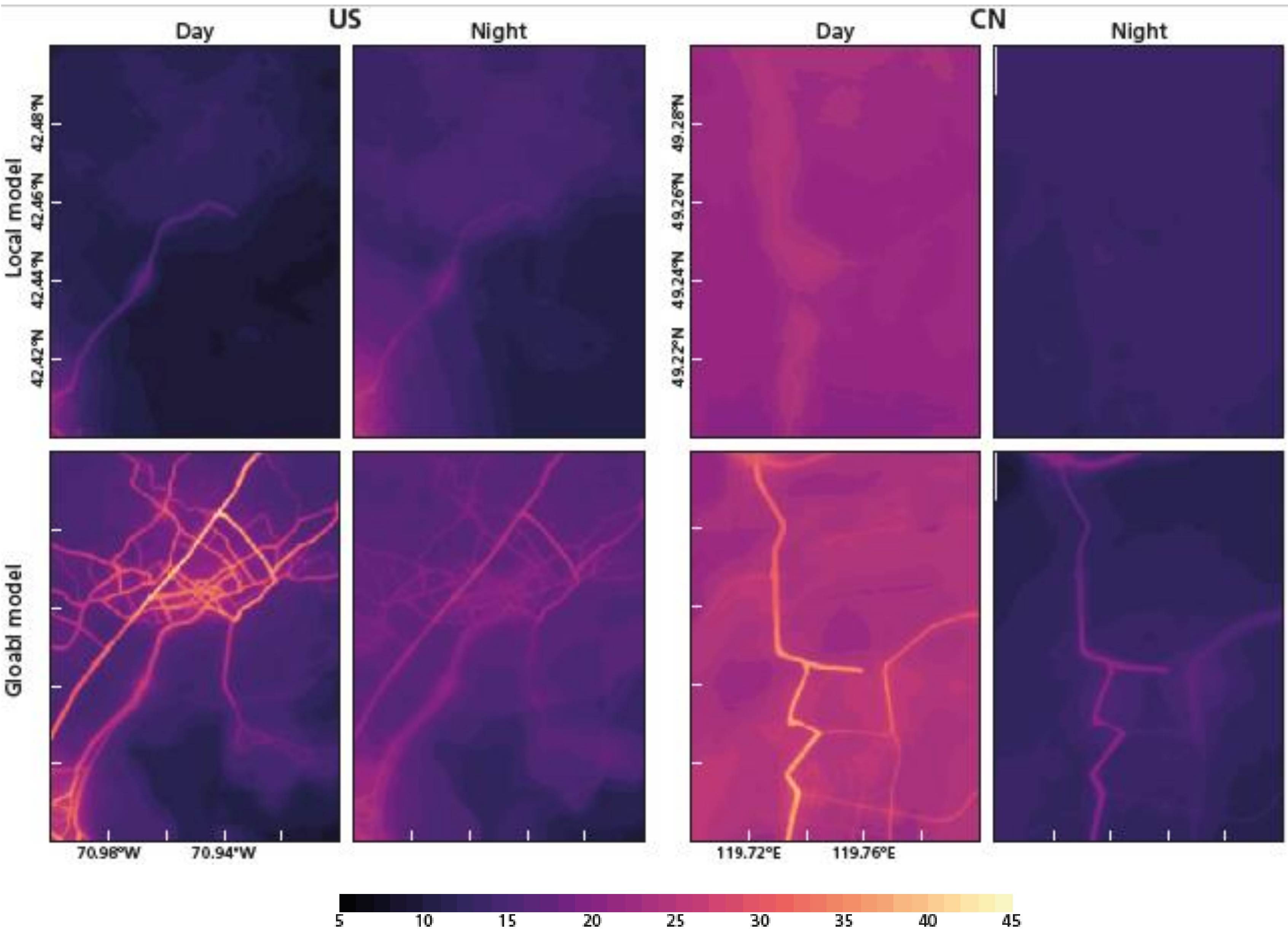


RMSE
MAE
IQR
Day
Night
accuracy
12
9
6
3
0
LM
Lasso
Ridge
ElasticNet
RF
SGB
XGB
methods

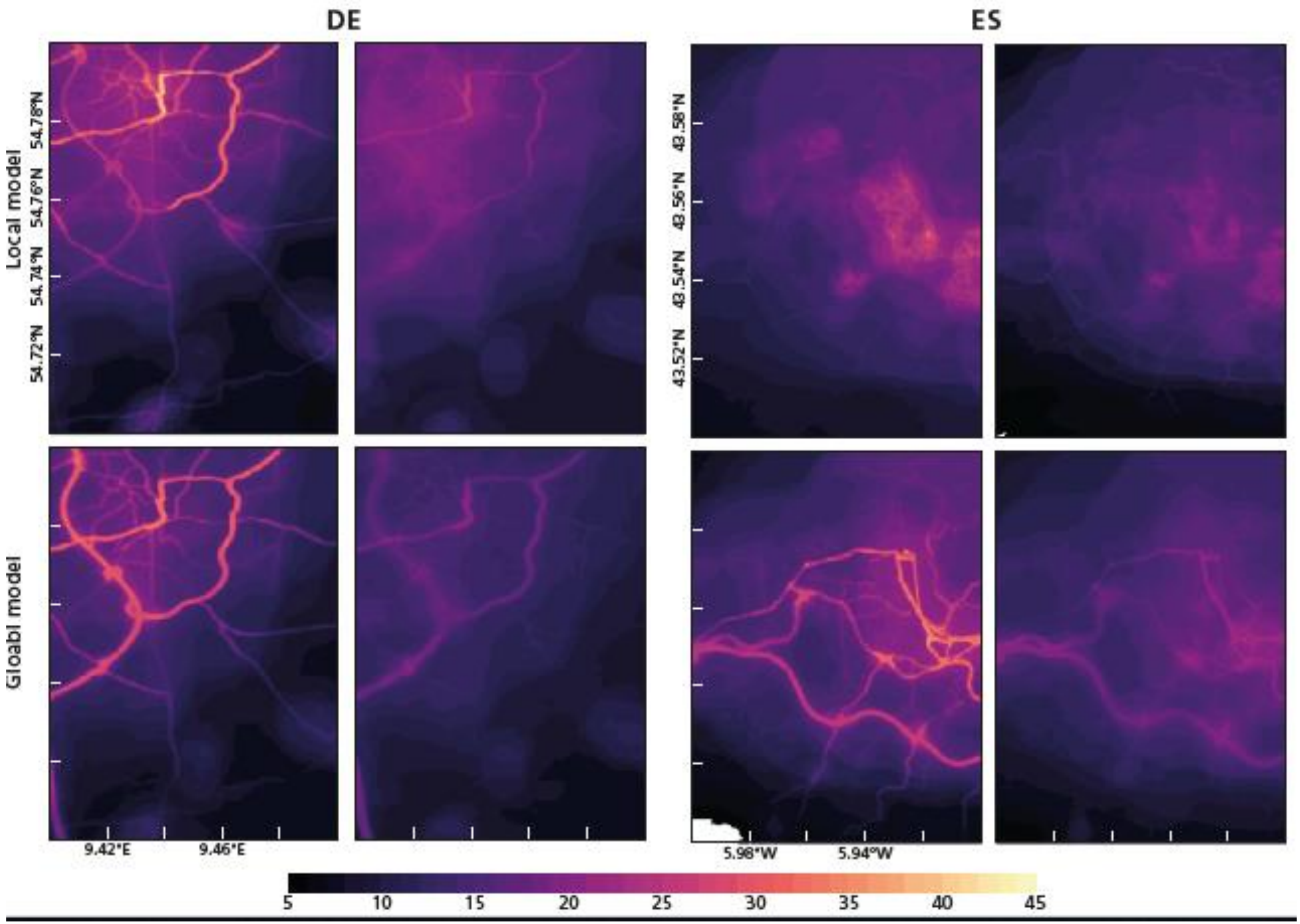
Xgboost vs. Lasso



Predicting using random forest



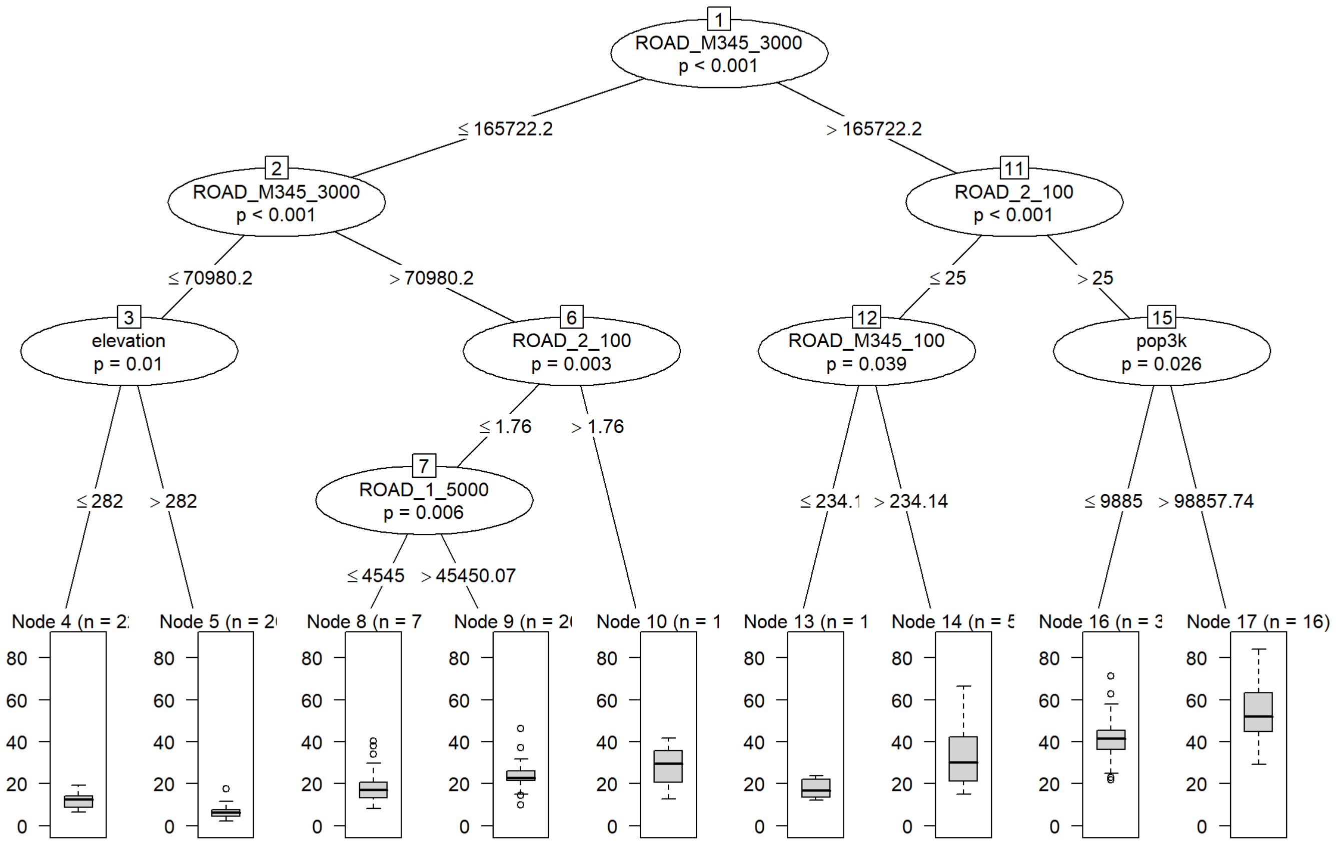
Predicting using random forest



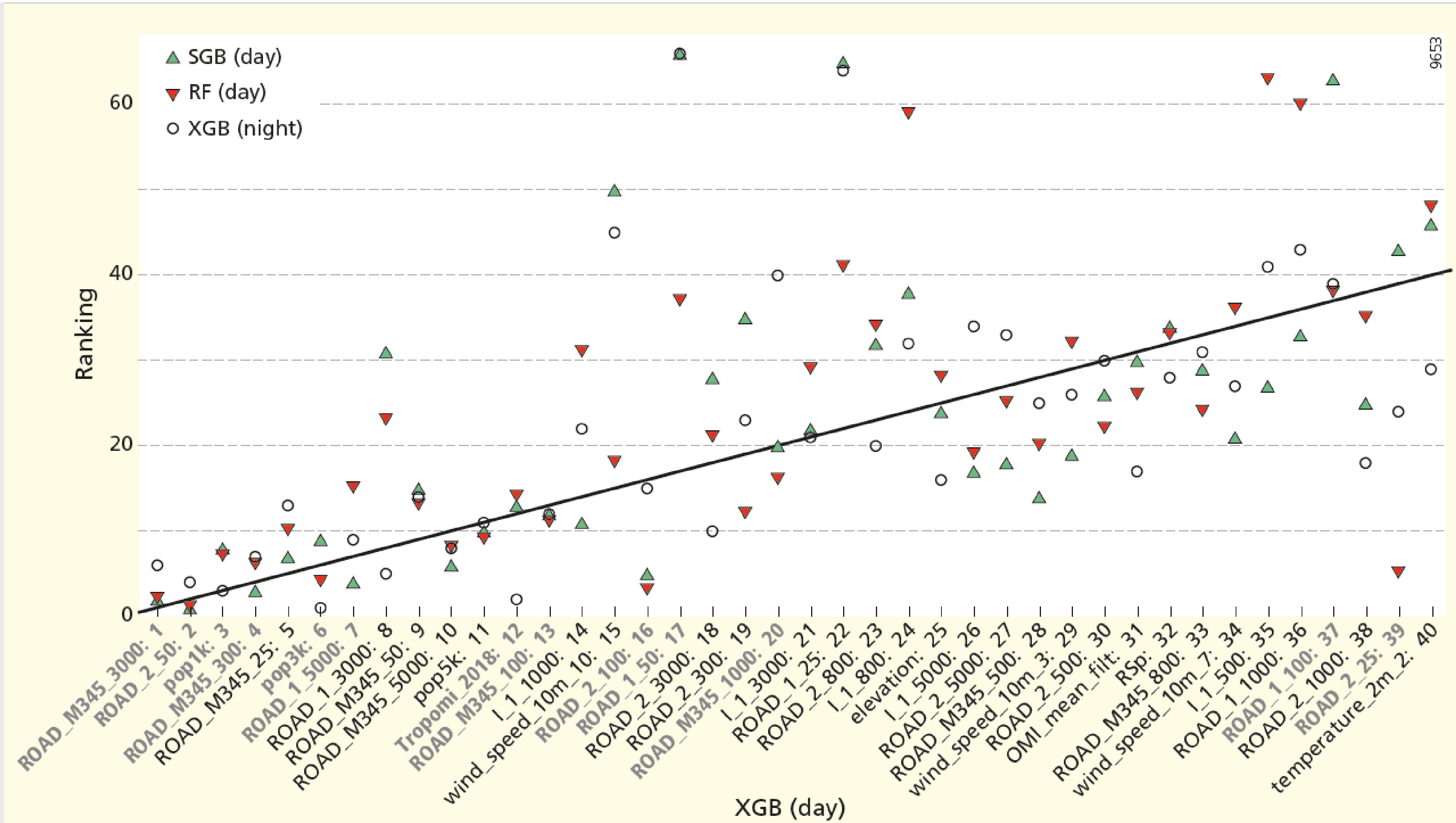
A closer look at the model

ROAD_M345: secondry and local roads
Pop_: population
ROAD_2: primary roads
ROAD_1: highway

Visualizing a tree



Variable importance



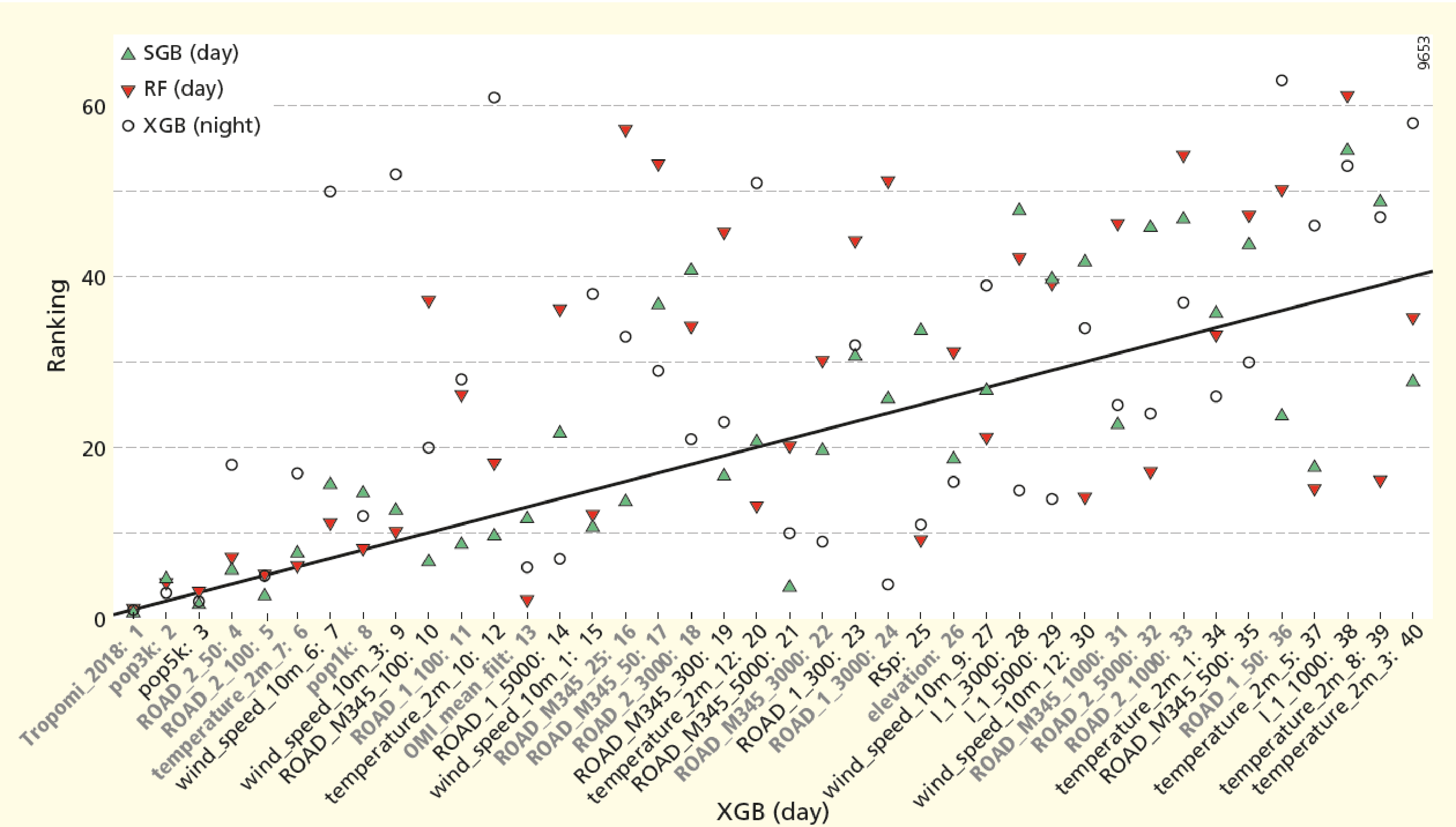
ROAD_M345: secondry and local roads

Pop_: population

ROAD_2: primary roads

Germany

Variable importance



ROAD_M345: secondry and local roads

World

Pop_: population

ROAD_2: primary roads

Partial dependence.

-- Shows the relationship between the target and a feature.

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C)$$

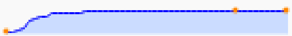
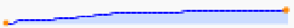
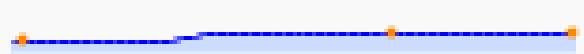
Xs : the features of the partial dependence function
Xc: the other features used in the machine learning model

Marginalizing the model output over the distribution of the features in set C,

Assumption: the features in C are not correlated with the features in S

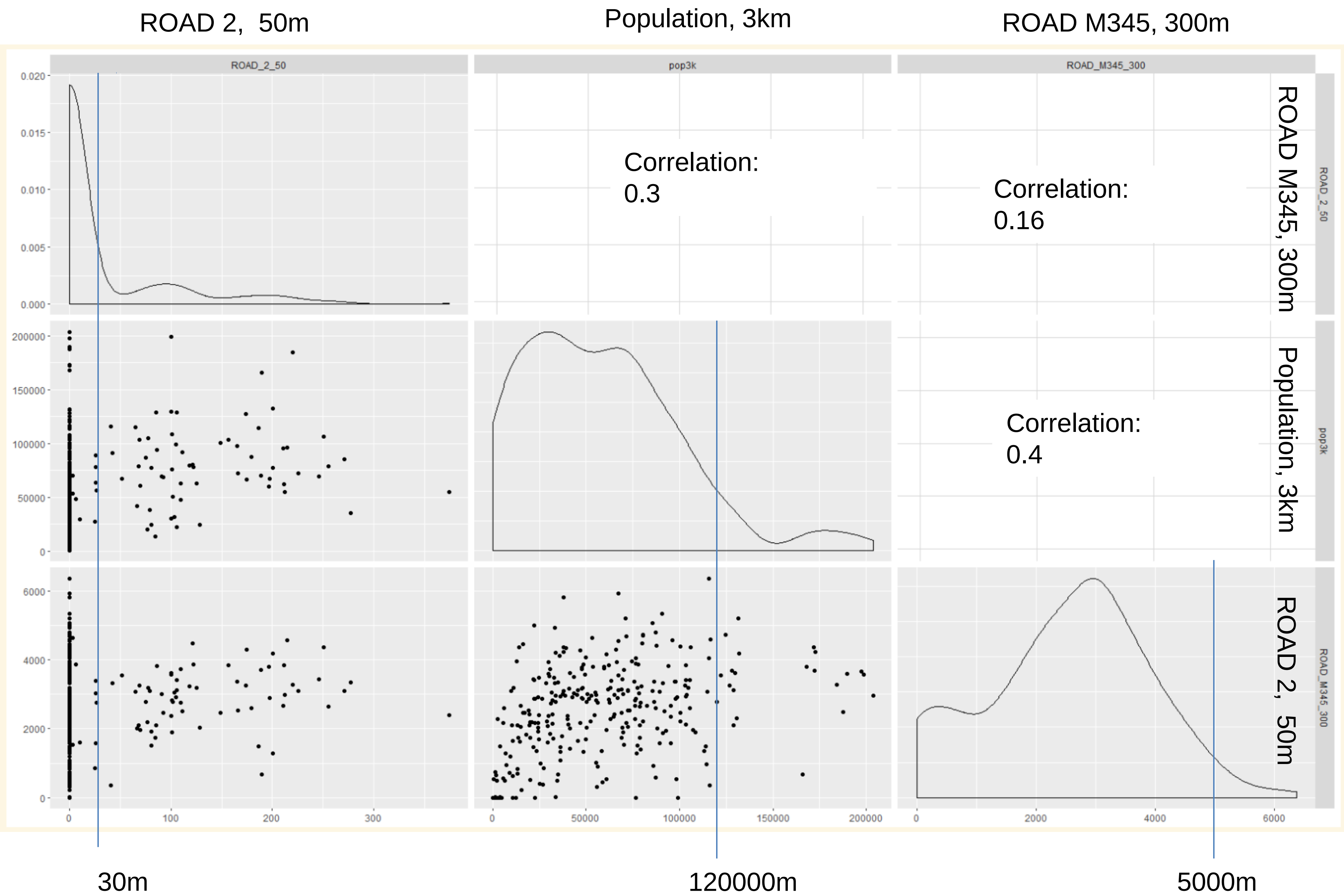
Show 10 entries

Search:

	Variable	Importance	Effect
1	ROAD_2_50	3.032	
2	ROAD_M345_3000	1.542	
3	pop3k	1.379	
4	ROAD_2_100	1.084	
5	ROAD_M345_300	1.058	
6	pop5k	0.840	
7	pop1k	0.756	
8	ROAD_M345_5000	0.674	
9	Tropomi_2018	0.654	
10	ROAD_M345_100	0.578	

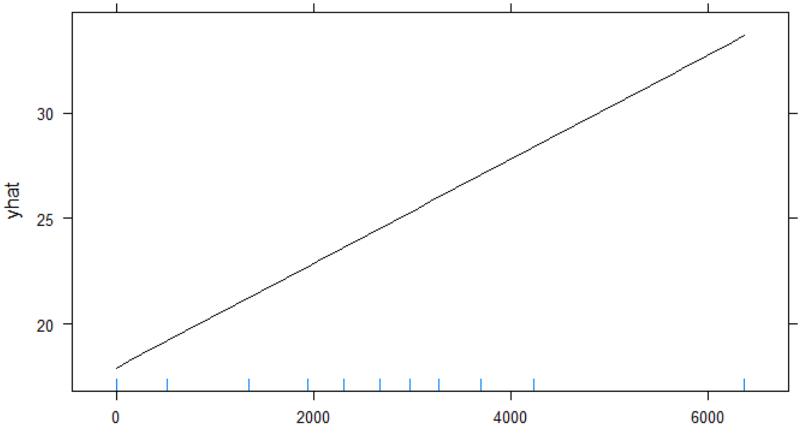
Showing 1 to 10 of 65 entries

Previous 1 2 3 4 5 6 7 Next

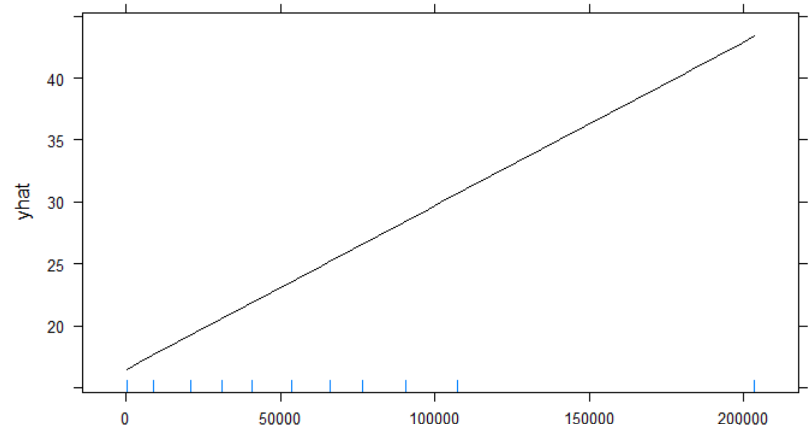


ROAD 2, 50m
Population, 3km
ROAD M345, 300m
ROAD_2_50
pop3k
ROAD_M345_300
Correlation:
0.3
Correlation:
0.16
Correlation:
0.4
ROAD M345, 300m
Population, 3km
ROAD 2, 50m
30m
120000m
5000m

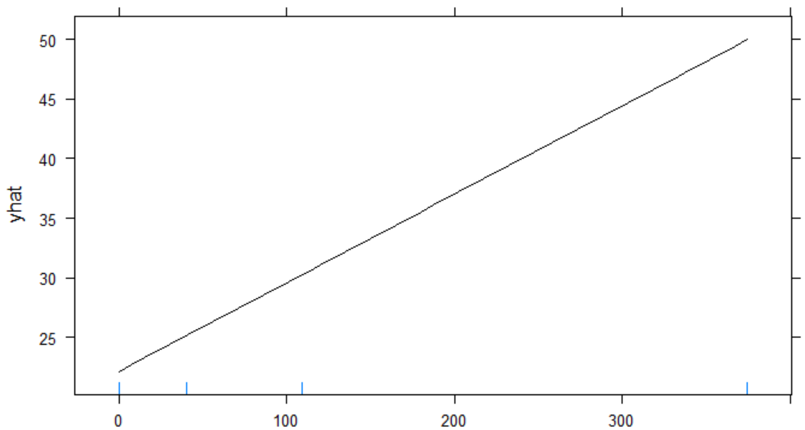
Partial dependent plots: Linear regression



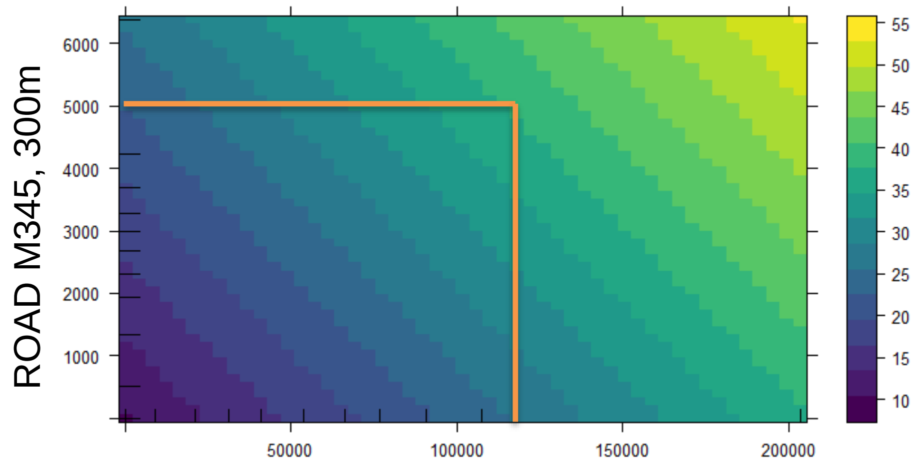
ROAD M345, 300m



Population, 3km

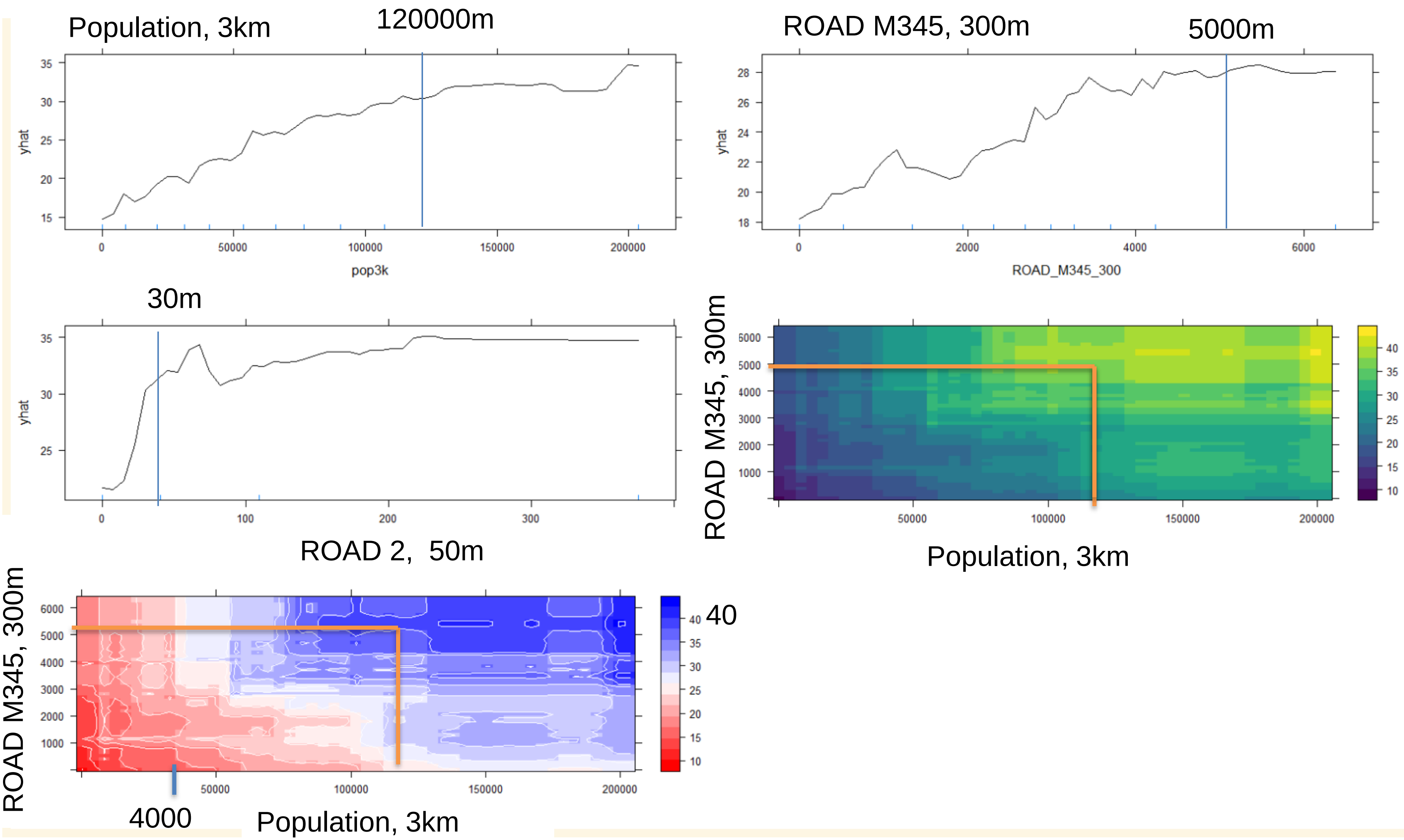


ROAD 2, 50m

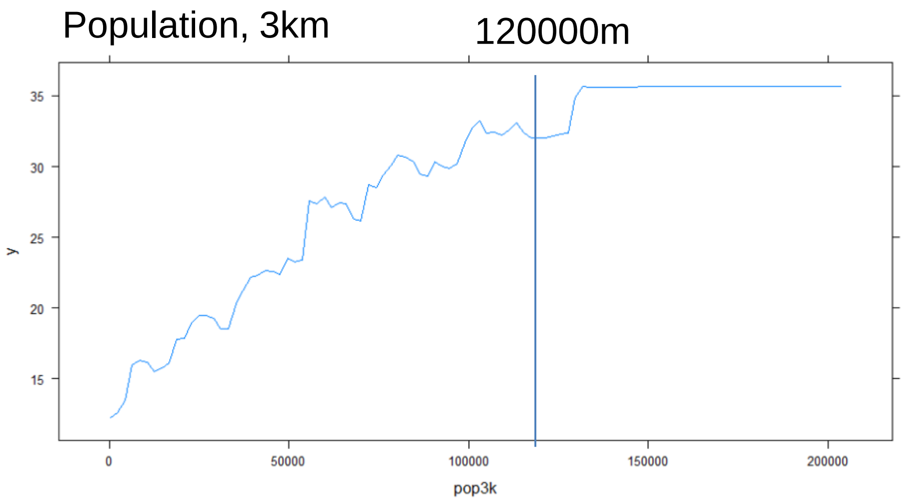


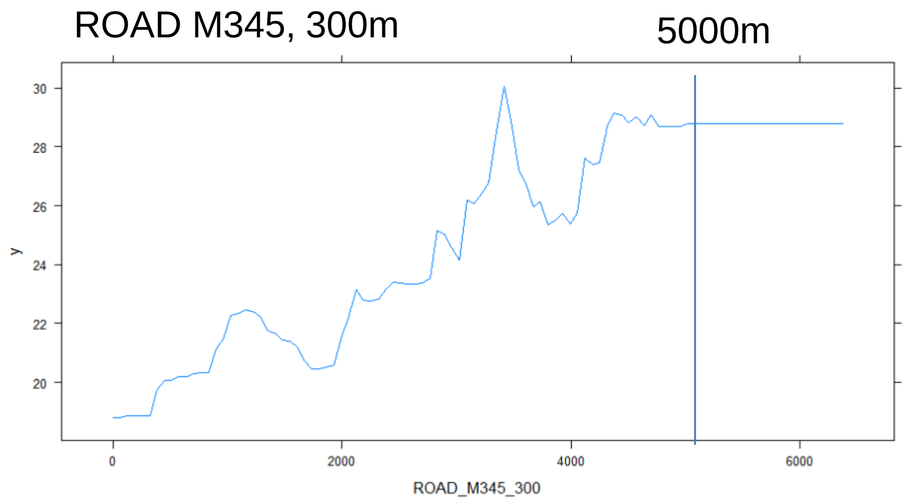
Population, 3km

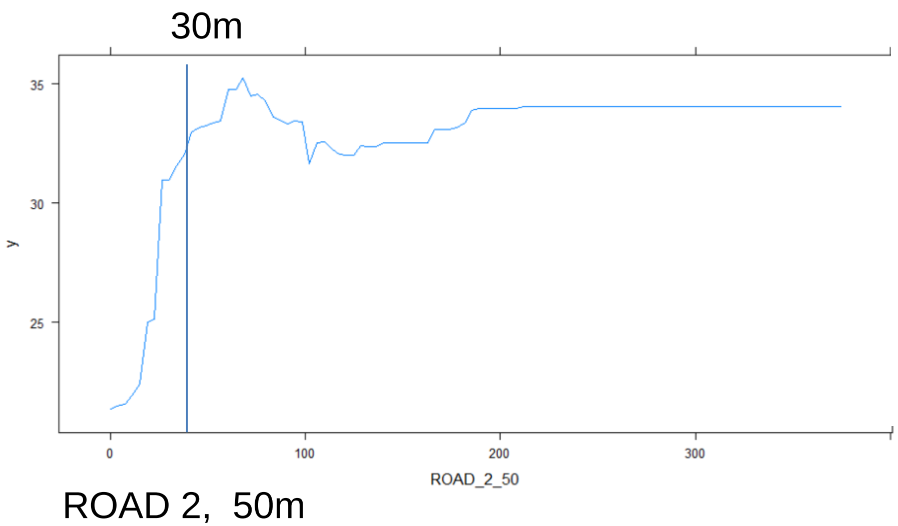
Partial dependent plots: Random forest

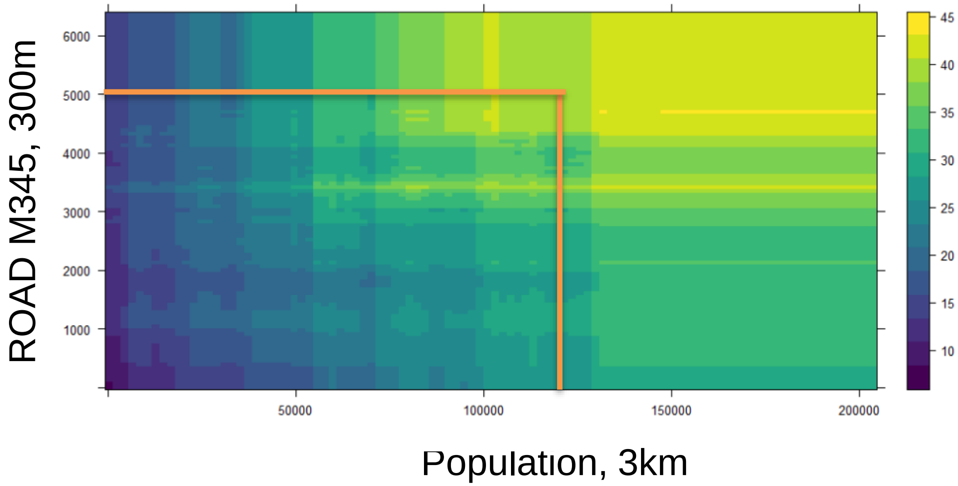


Partial dependent plots: boosted regression trees

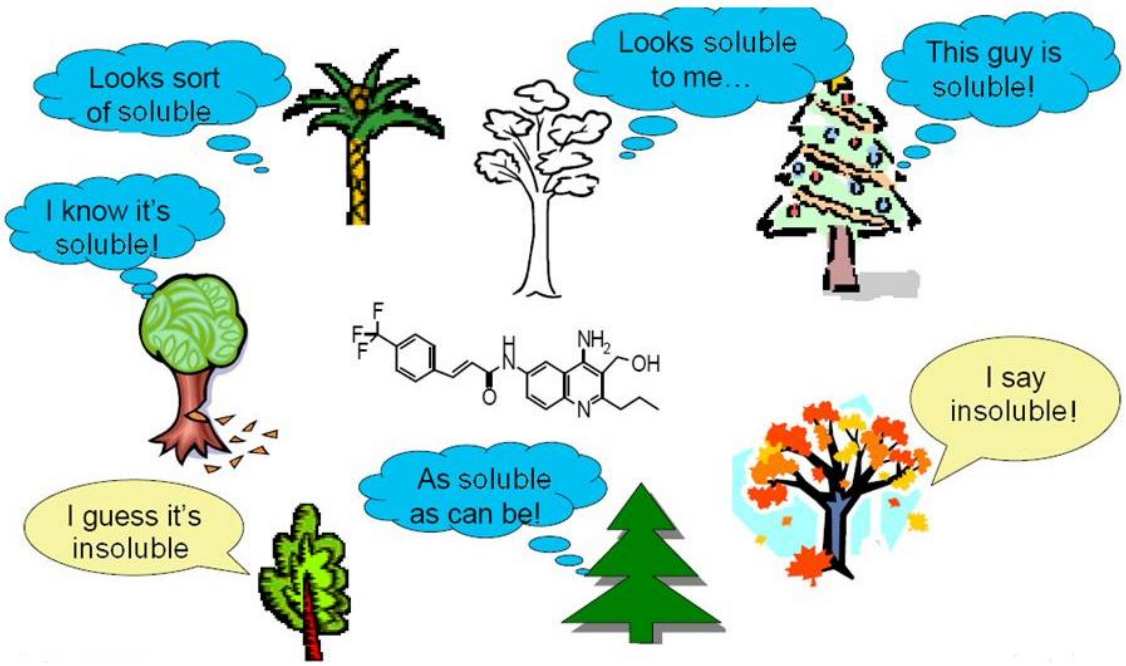








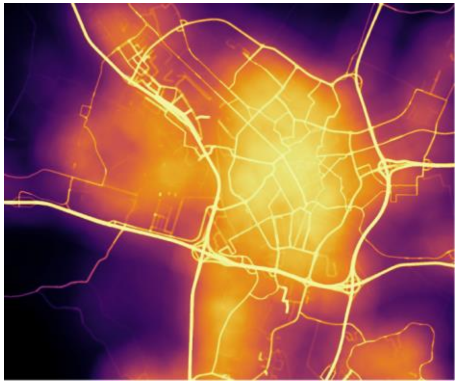
Questions



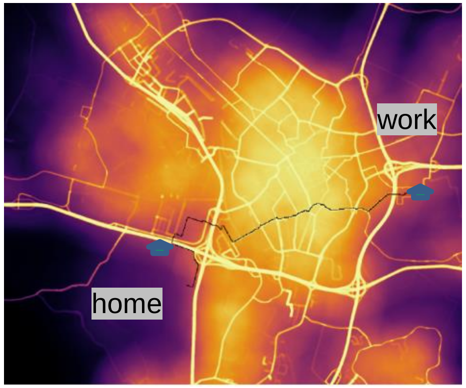
Looks sort of soluble
Looks soluble to me…
This guy is soluble!
I know it's soluble!
I say insoluble!
As soluble as can be!
I guess it's insoluble

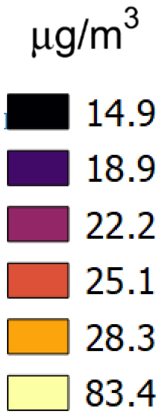
Personal exposure assessment

NO2 *of a single hour*



NO2 *exposure* assessed along the route from home to work

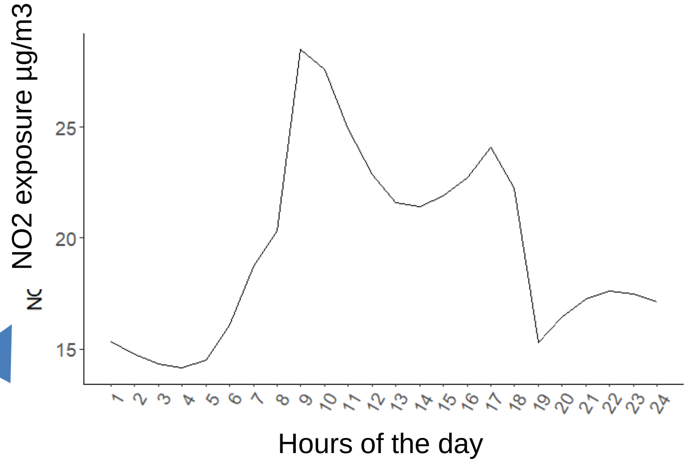




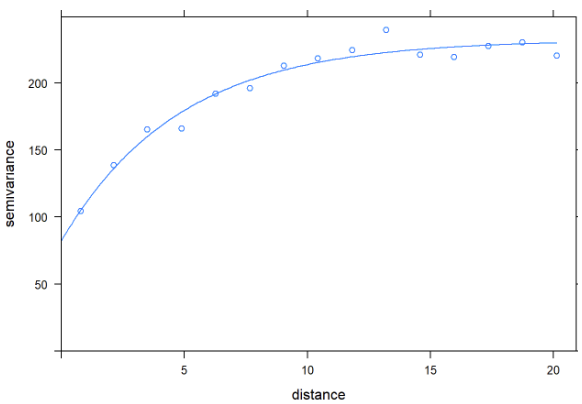
Predicting NO2 for each hour: land use regression model from sensor data

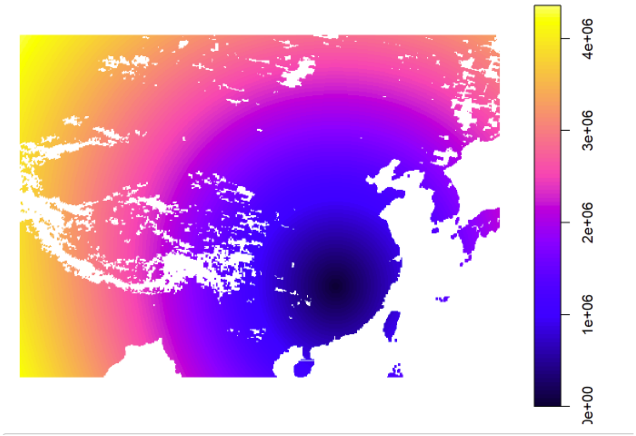
Assessing NO2 exposure according to a person's geographical location according to an activity schedule.

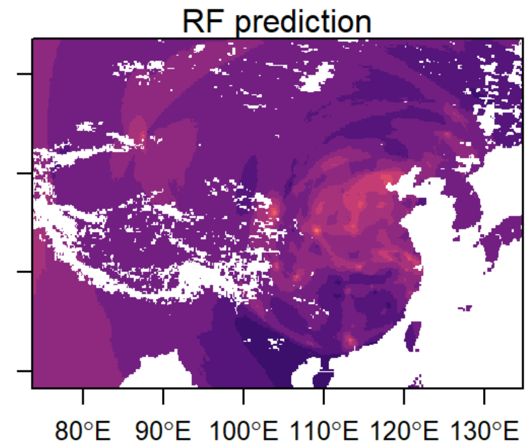
Personal NO2 exposure assessed for each hour of the day



Using random forest for Geostatistic-like interpolation







http://rpubs.com/menglu/473973

[1] Shaddock et al., 2018: *Environ. Sci. Technol.*201852169069-9078
[2] https://www.theguardian.com/sustainable-business/2016/jul/05/how-air-pollution-affects-your-health-infographic
[3] http://www.tropomi.eu/sites/default/files/files/agu_veefkind.pdf

Multiple linear regression model

$$Y = XB + e$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Y: $Station\_measurements$

X: variable matrix

e: error

B: coefficient matrix

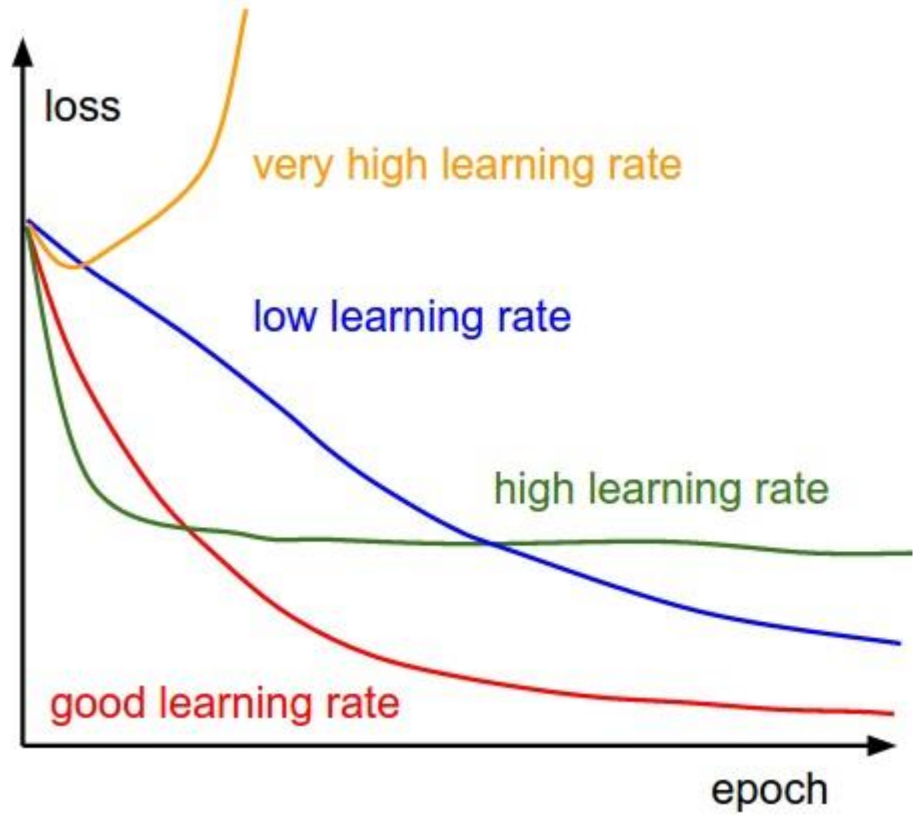
The variable matrix X

Consists of land use predictor variables, could be:

Measurements within a buffer: population, road length, traffic load, number of factories, ...

Point measurements: green space, metereological data, ...

Learning Rate



Stochastic Gradient Boosting

Each consecutive tree is built for the prediction residuals (from all preceding trees) of an independently drawn random sample

General boosting and gradient boosting

$$(\beta_m, \gamma_m) = \arg\min_{\beta,\gamma} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \beta b(x_i; \gamma)).$$

$$\text{Set } F_m(x) = F_{m-1}(x) + \epsilon \beta_m b(x; \gamma_m)$$

(Stochastic) Gradient Boosting

approach the gradient of the loss function (e.g. binomial, logistic, poison) by trees.

Each consecutive tree is built for the prediction residuals (from all preceding trees) of an independently drawn random sample