Statistical methods of global air pollution modeling

University of Utrecht, The Netherlands

Meng Lu

S1 Air pollution mapping and data exploration
Introduction
Land use regression model
Data

Scripts
data exploration: from visualization to preliminary examination

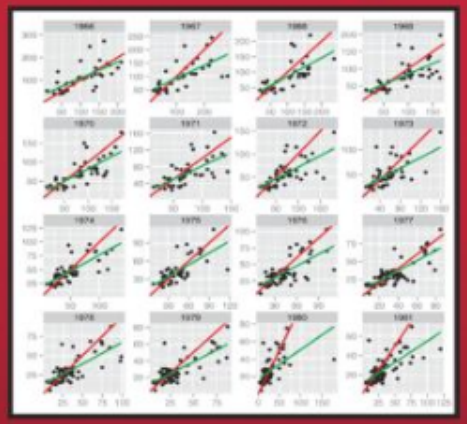
S2 Machine learning methods
Introduction
machine learning methods

Scripts
How to use them, modeling parameter tuning, partial dependence plot, variable importance

S3 Prediction (map) and result analysis
Scripts
Predicting using Lasso, random forest, xgboost
Results

Texts in Statistical Science

Spatio-Temporal Methods in Environmental Epidemiology



Gavin Shaddick
James V. Zidek

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Part 1

Land use regression and data used in the project

Land use regression (LUR)

Predicting air pollution and analyzing the sources.

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

Sensor measurements:

Station measurements

Mobil sensors

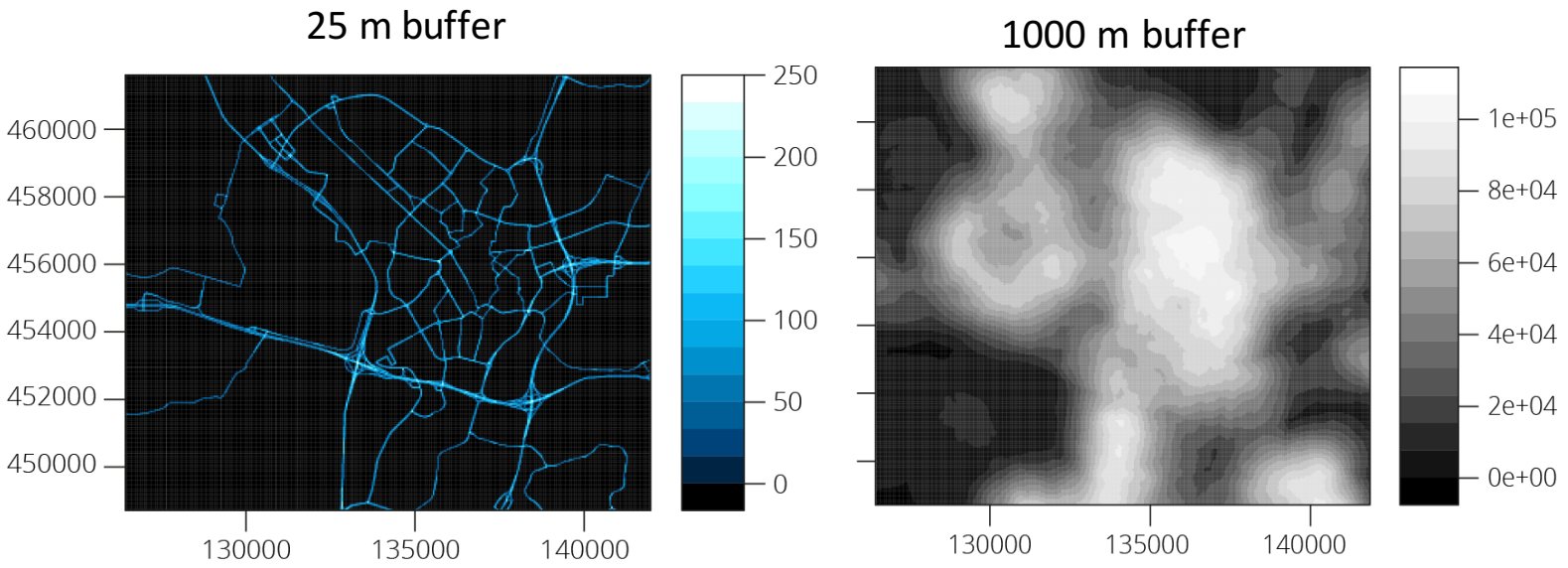
Remote sensing measurements:

OMI (250 km)
Tropomi (8 km)
…

GIS predictors:

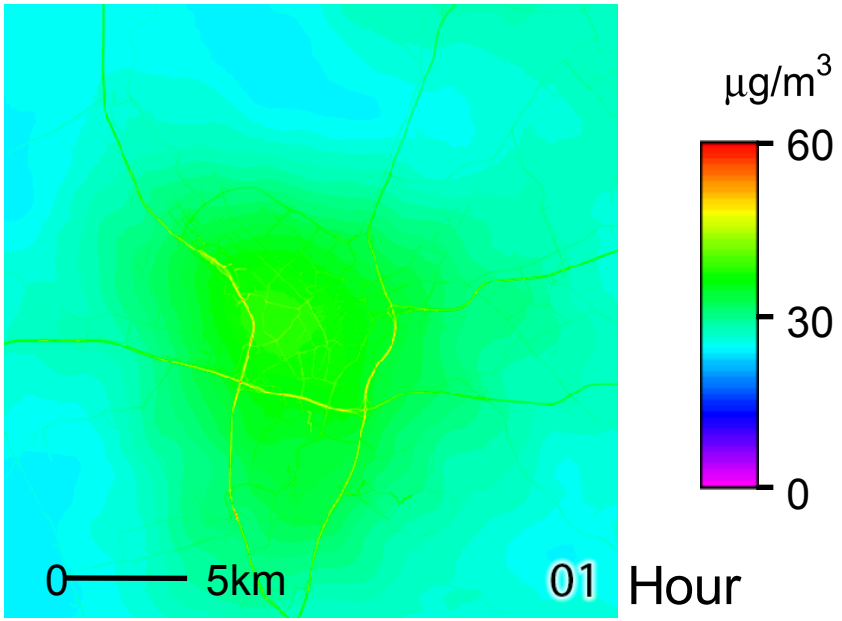
Population
Road length within a buffer
Distance to roads
Traffic load
…

Buffered predictors



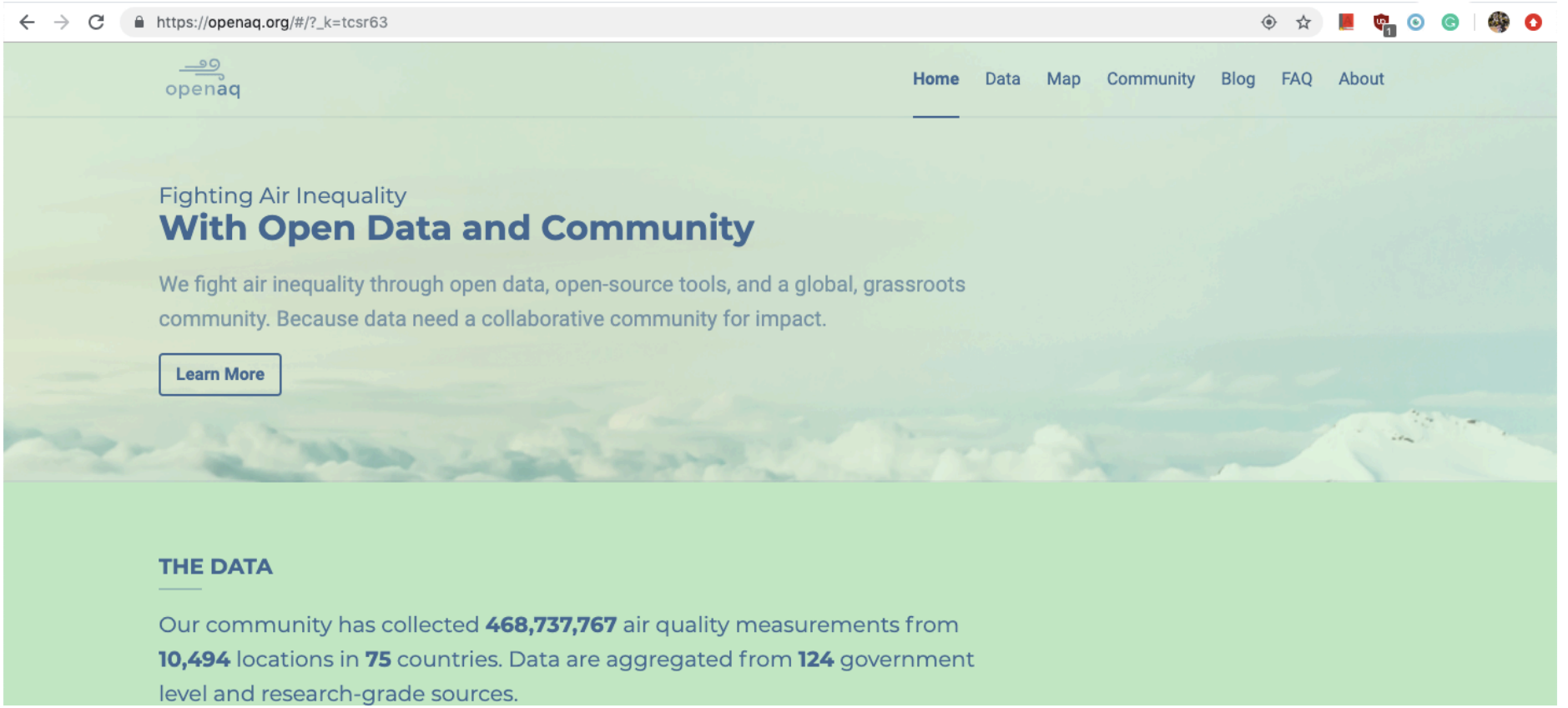
LUR prediction of NO_2

Spatiotemporal dynamic of air pollution showing road effects



5 m resolution

Ground monitor Data: OpenAQ (2017)



Predictors

Road types: Highway, primary roads, local roads

Emission-related
Road length and industry area within 25 m – 300 m radius ring

Background
Road length within 300 m - 5000 m radius ring
Population: 1 km, 3 km, 5 km
Industry area 300m - 5km
Monthly wind speed (0.5 degree)
Monthly temperature (0.5 degree)
Surface concentration product from Satellite products and the GEOS-CHEM
Satellite measured NO2 column density

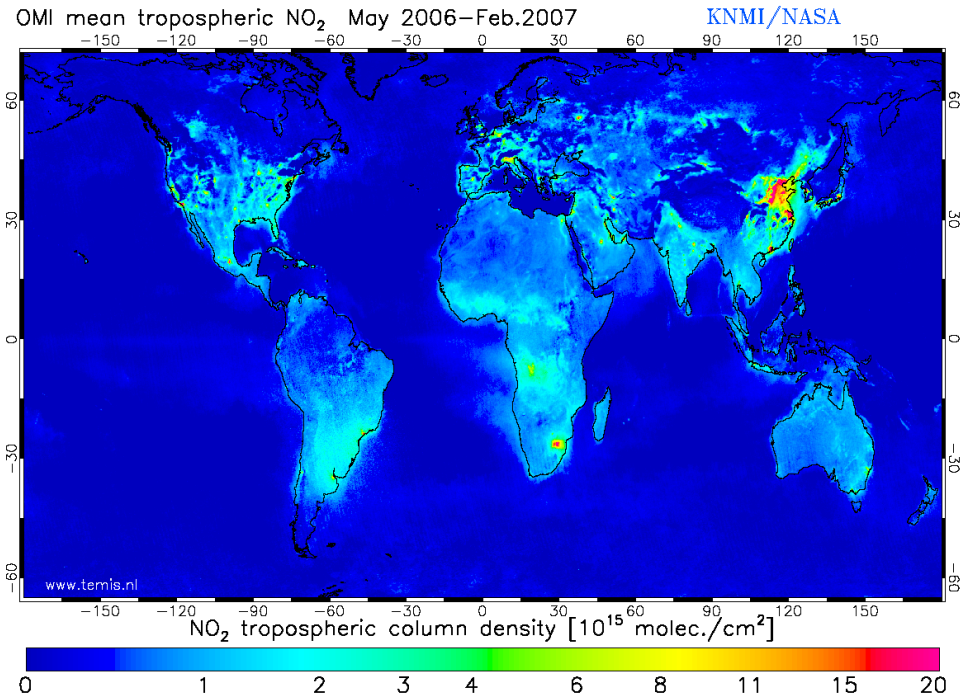
Tropomi

Product	Spectrometer
Ozone	UV, UVIS
NO_2	UVIS
CO	SWIR
CH_2O	UVIS
CH_4	SWIR
SO_2	UVIS
Aerosol	UVIS, NIR
Clouds	UVIS, NIR
UV-Index	UVIS

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	

Feb 2018 to Jan 2019

Remote sensing measurements:
OMI (Ozone Monitoring Instrument)



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

Zoom in mode 13 km× 12 km

Daily global coverage

Date of Launch: 15 July 2004

At nadir 13 km × 24 km

NO2, SO2, BrO, OClO, O3 (36 km × 48 km)

surface product based on remote sensing and GEOS-CHEM

- Chemical transportation models: GEOS-CHEM

- global prediction are of coarse resolution - > integrating RS data
- Surface NO2 product made using GEOS-CHEM together with GOME, SCIAMACHY, and GOME-2 satellite instruments. http://fizz.phys.dal.ca/~atmos/martin/?page_id=232
- The surface product is used in our model.

Road length	https://www.openstreetmap.org/
Industrial area	https://www.openstreetmap.org/
Population	http://www.integrated-assessment.eu/eu/index.html
Elevation	https://www2.jpl.nasa.gov/srtm/
Wind speed	https://confluence.ecmwf.int/pages/viewpage.action?pageId=74764925
Temperature	https://confluence.ecmwf.int/pages/viewpage.action?pageId=74764925
TROPOMI	http://www.tropomi.nl
NO2 Surface product	http://fizz.phys.dal.ca/~atmos/martin/?page_id=232
OMI	https://mirador.gsfc.nasa.gov/cgi-bin/mirador/presentNavigation.pl?tree=project&dataset=OMNO2d.003&project=OMI&dataGroup=L3_V003&version=003&CGISESSID=57642d429543194f1007d9b91bfab0ae

Part 2 (1)

Statistical learning

1980, computer science, machine learning, neural networks, zipcode recognision problem, bell lab

Statisitcal learning

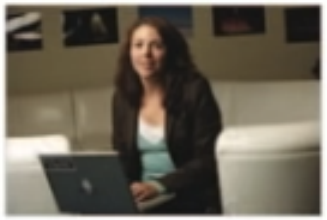
For Today's Graduate, Just One Word: Statistics

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.

SIGN IN TO RECOMMEND
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Enlarge This Image



Thor Swift for The New York Times

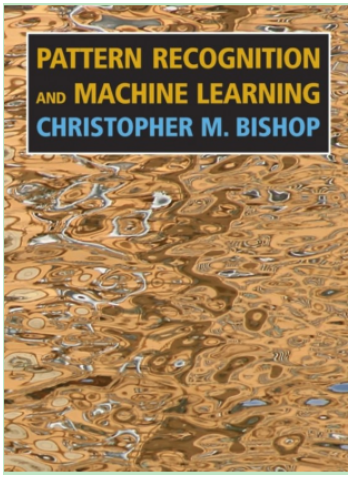
Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

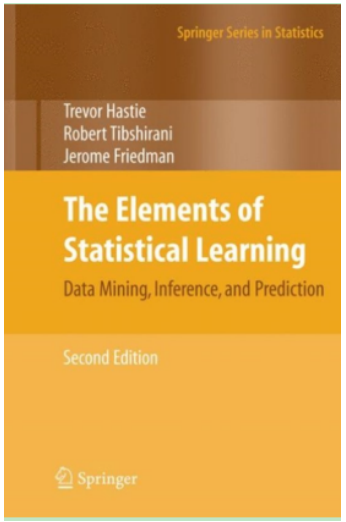
"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

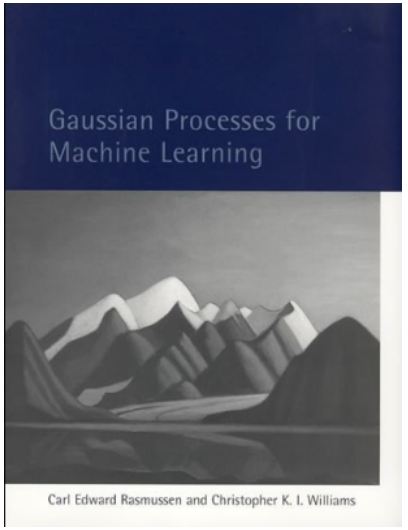
Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

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.



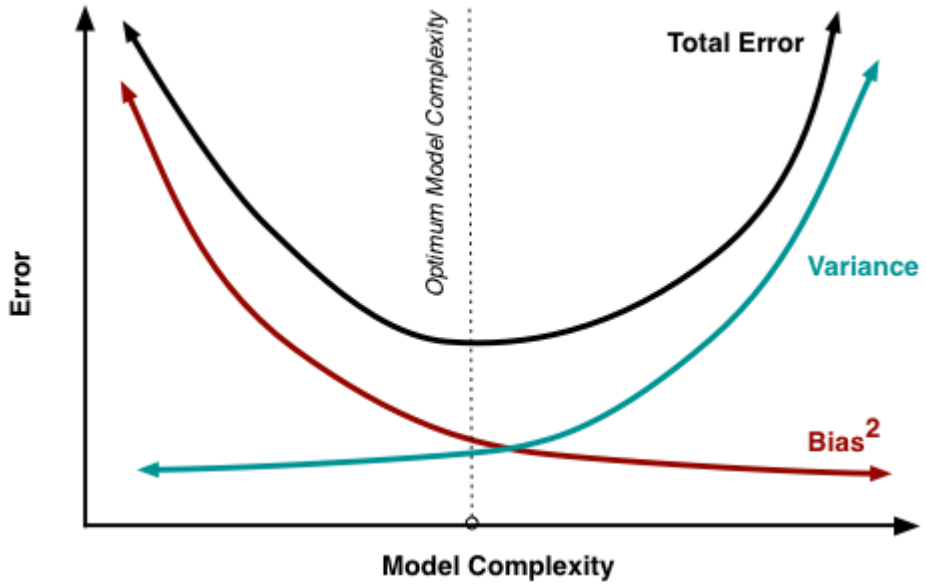




Bias-variance trade-off

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

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

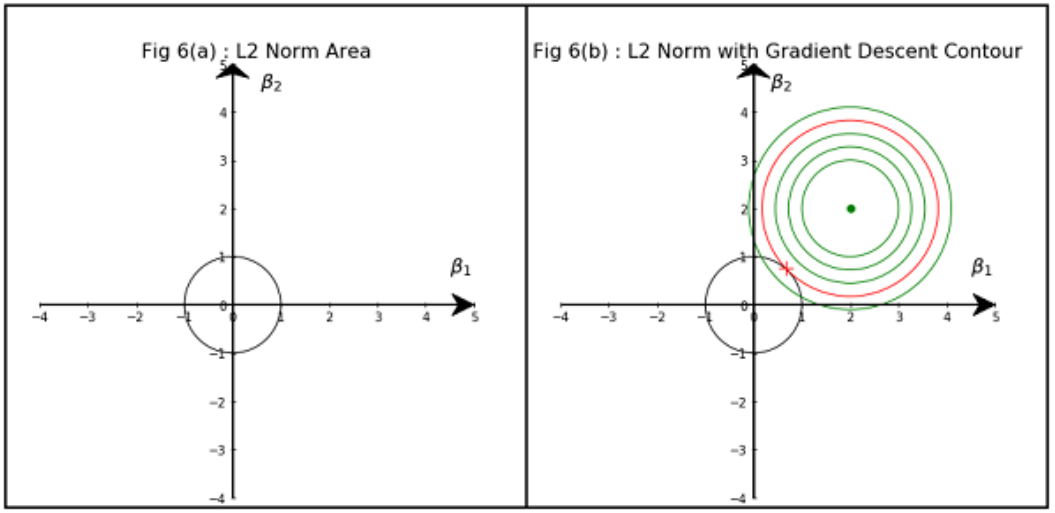


All algorithms are affected by bias-variance trade-off

Regularization

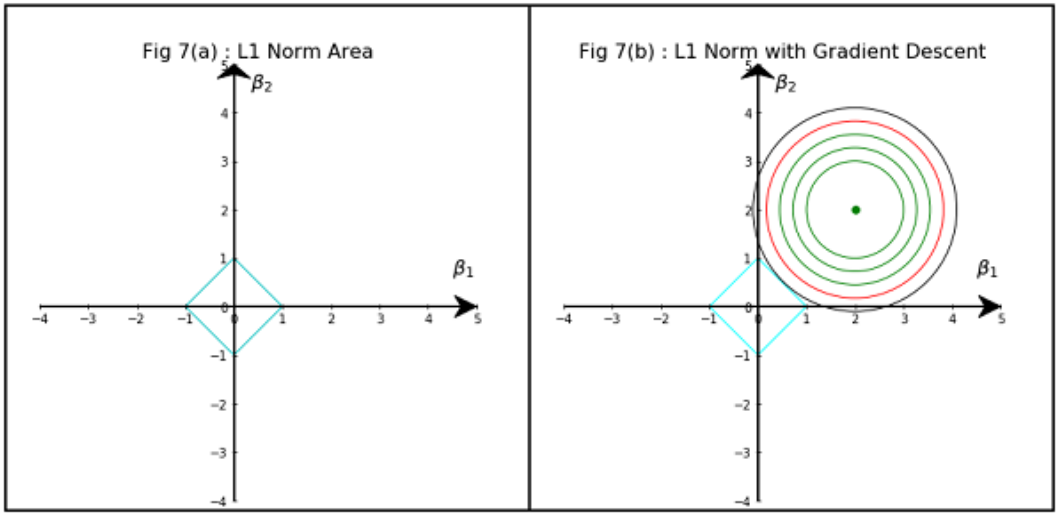
Ridge regression

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

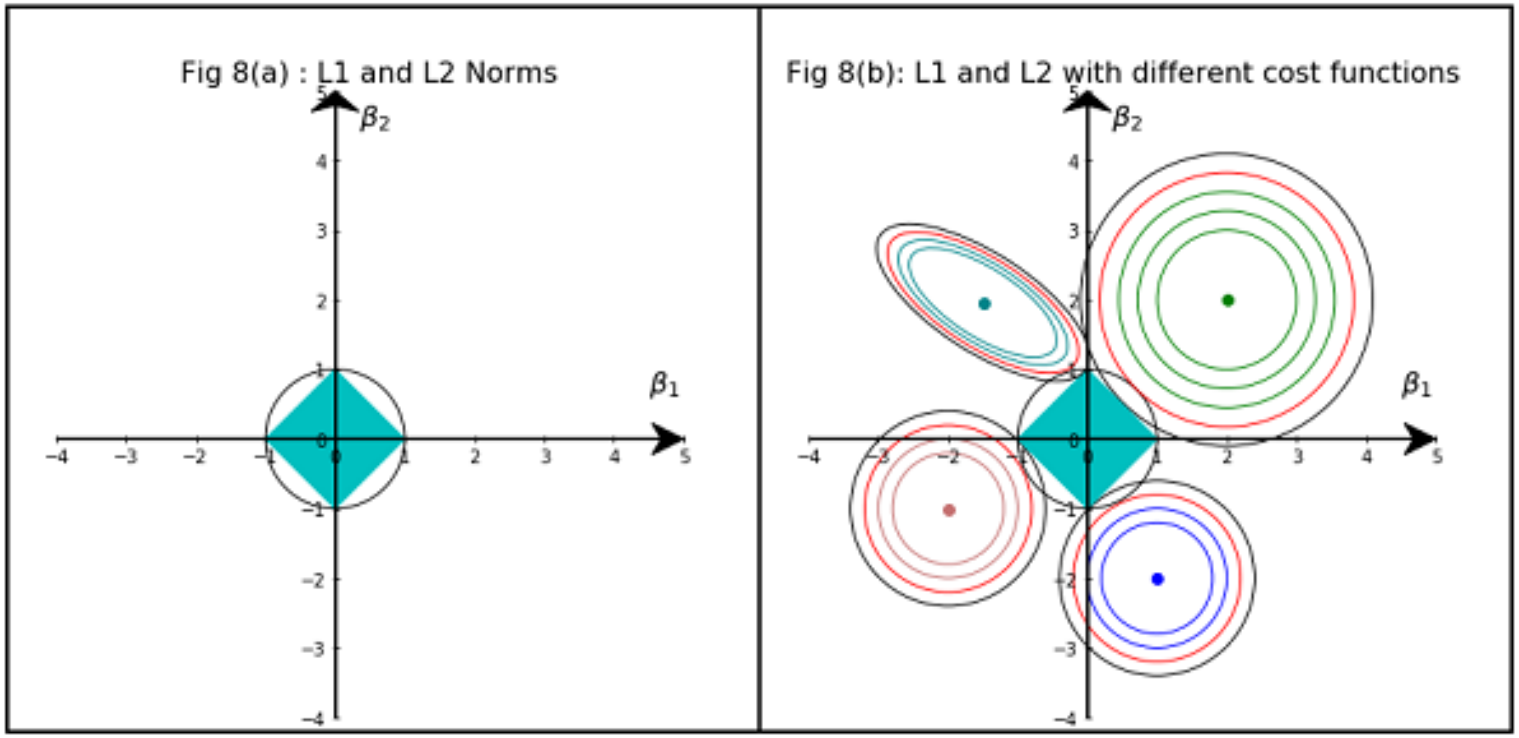


Lasso regression

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



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

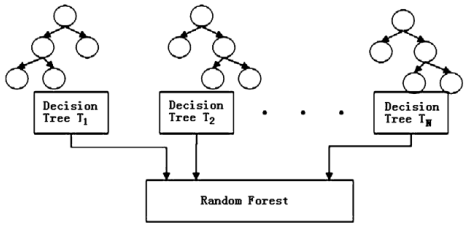
Leveraging trees to improve the performance:

- 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^2
An average of B of i.i.d random variables has variance $\frac{1}{B}\sigma^2$

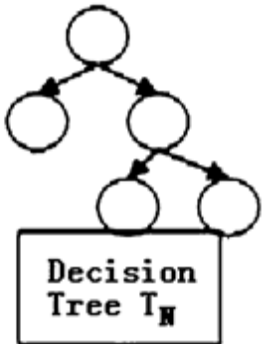
If the variables are not independent (but identically distributed) with positive pairwise correlation p , the variance of the average:

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

"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

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)
```

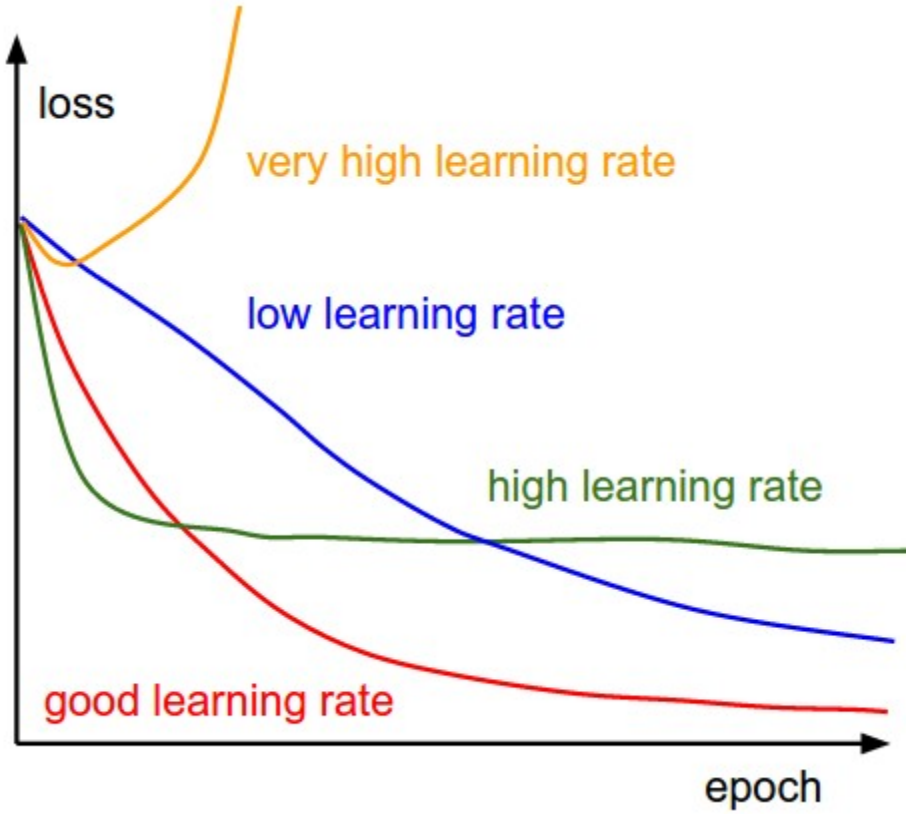
(Stochastic) Gradient Boosting

(r: residual; e: learning rate)

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

Learning Rate



Stochastic Gradient Boosting

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

XGboost
Exteme gradient boosting

Idea
Not only impurity, but also model complexity

$$obj(\theta) = L(\theta) + \Omega(\theta)$$

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

Hyperparameters

nrounds:

- Controls the maximum number of iterations.
- Tuned by CV

Eta:

- Controls the learning rate,
- It must be supported by increase in nrounds.

max_depth:

- It controls the depth of the tree. Larger data sets require deep trees to learn the rules from data.
- Tuned by CV

Gamma:

- Controls regularization (or prevents overfitting).
- Higher the value, higher the regularization. Regularization means penalizing large coefficients which don't improve the model's performance.
- If you see train error >> test error, bring gamma into action.

Norms:

- L1 norm (alpha) and L2 norm (lambda)

Postprocessing

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

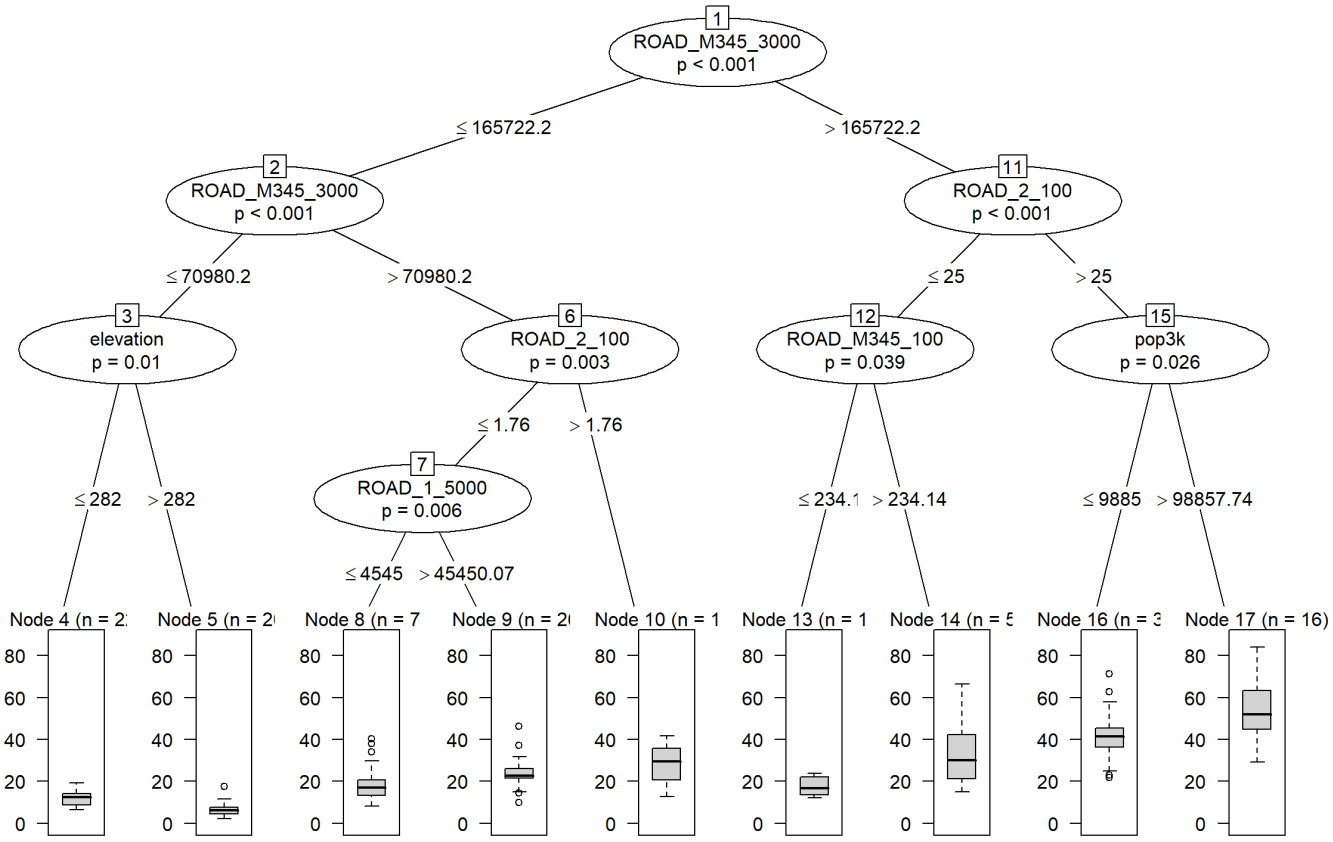
$$\alpha(\lambda) = \arg\min_{\alpha} \sum_{i=1}^{N} L[y_i, \alpha_0 + \sum_{m=1}^{M} \alpha_m T_m(x_i)] + \lambda \sum_{m=1}^{M} |\alpha_m|.$$

Part 2 (2)

What can the machine learning methods tell?
Partial dependence plot, trees, variable importance

A closer look at the model

Visualizing a tree



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

Partial dependence.

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

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

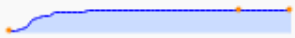
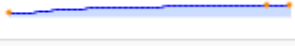
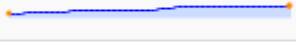
Xs : the features of the partial dependence function
Xc: the other features used in the machine learning model

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

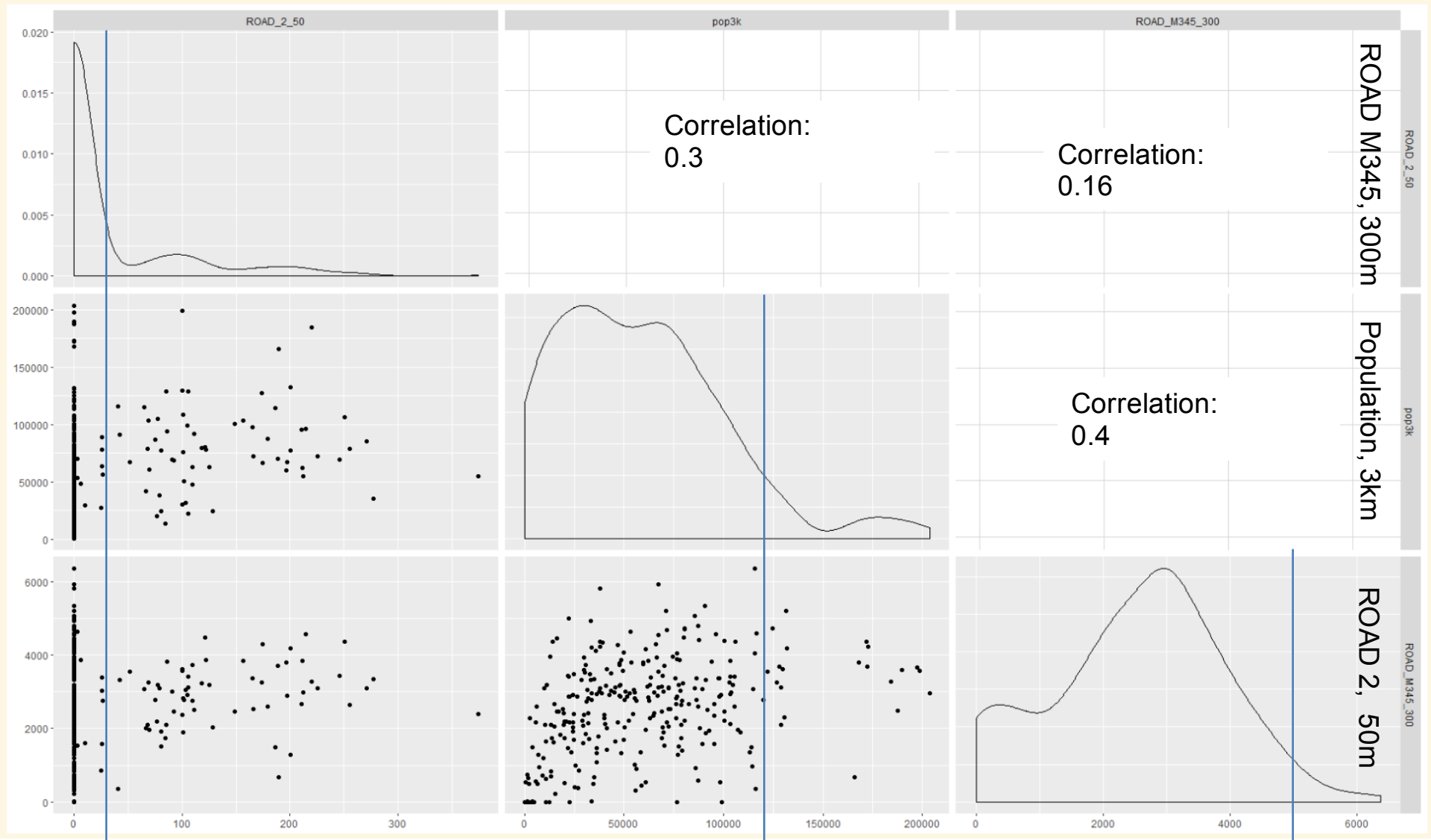
Show 10 entries

Search:

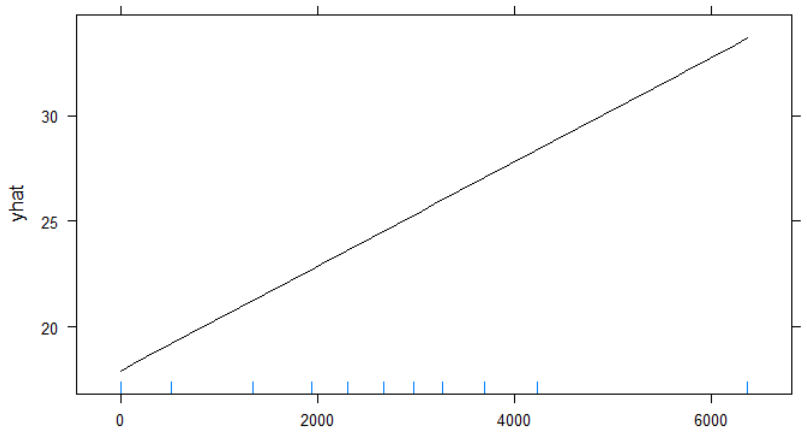
	Variable	Importance	Effect
1	ROAD_2_50	3.032	
2	ROAD_M345_3000	1.542	
3	pop3k	1.379	
4	ROAD_2_100	1.084	
5	ROAD_M345_300	1.058	
6	pop5k	0.840	
7	pop1k	0.756	
8	ROAD_M345_5000	0.674	
9	Tropomi_2018	0.654	
10	ROAD_M345_100	0.578	

Showing 1 to 10 of 65 entries

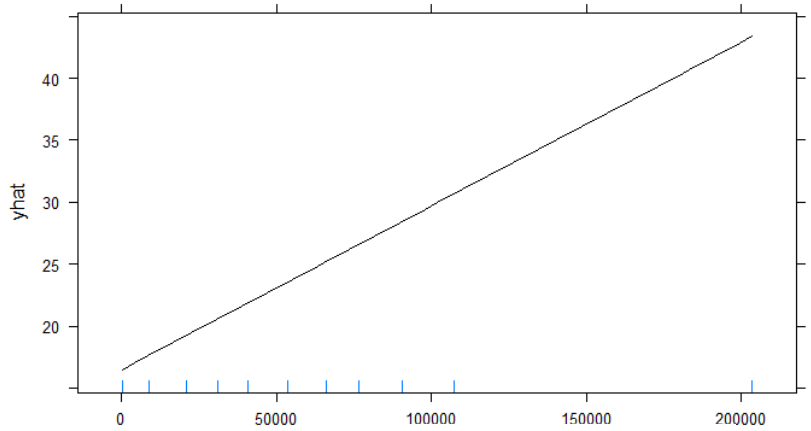
Previous 1 2 3 4 5 6 7 Next


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

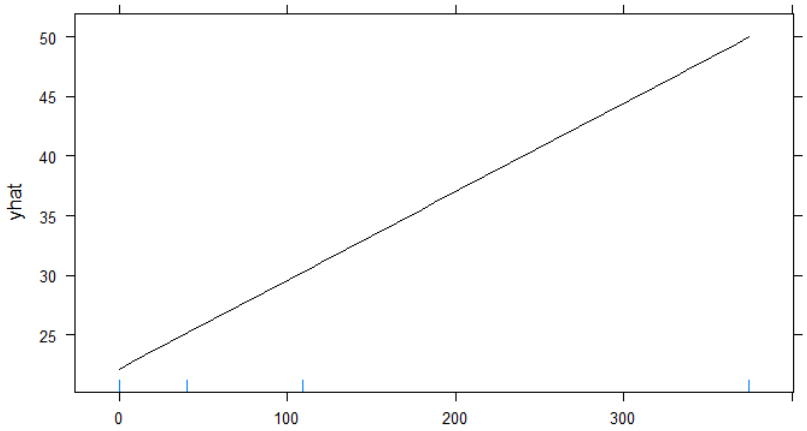
Partial dependent plots: Linear regression



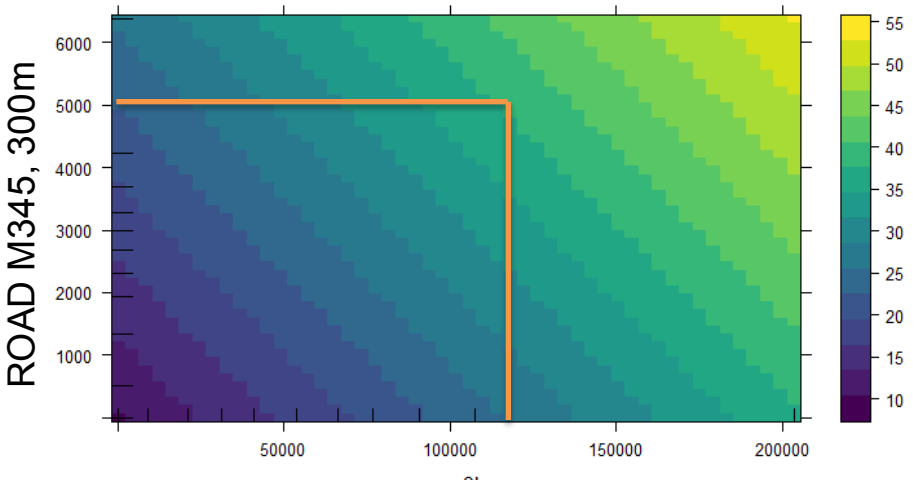
ROAD M345, 300m



Population, 3km

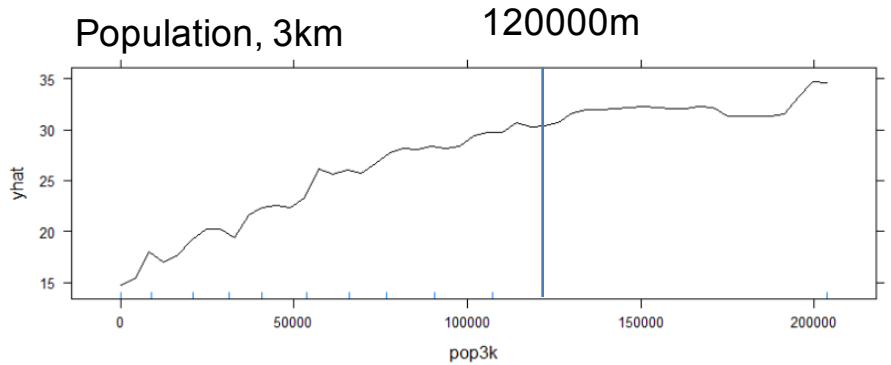


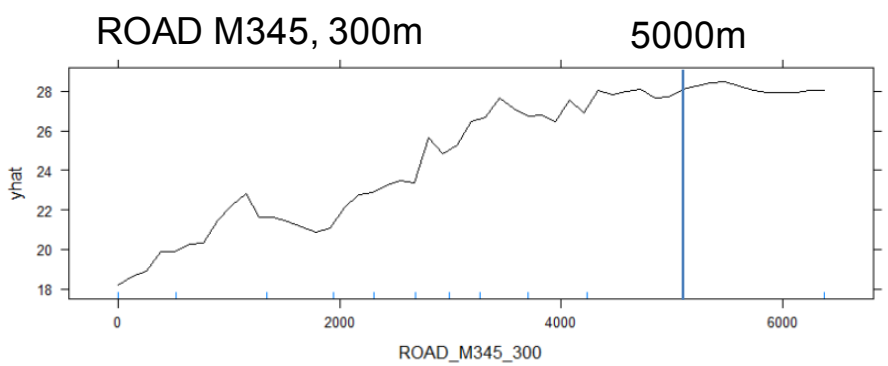
ROAD 2, 50m

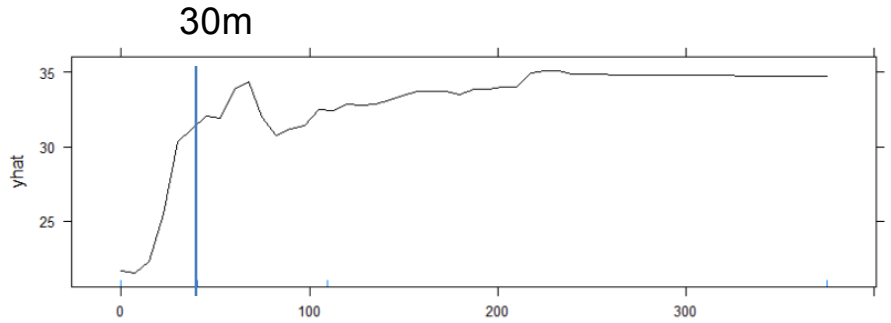


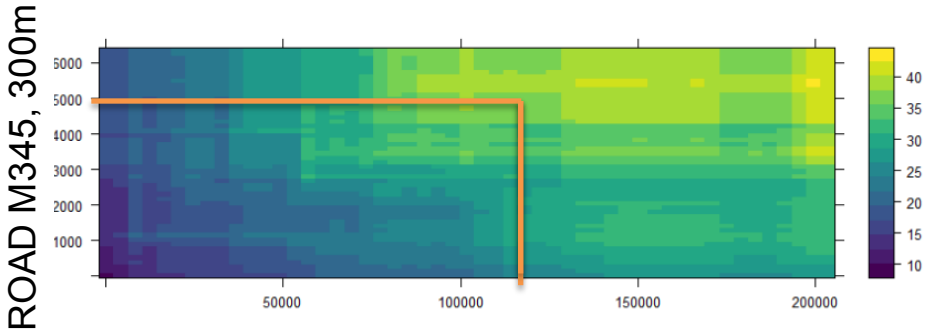
Population, 3km

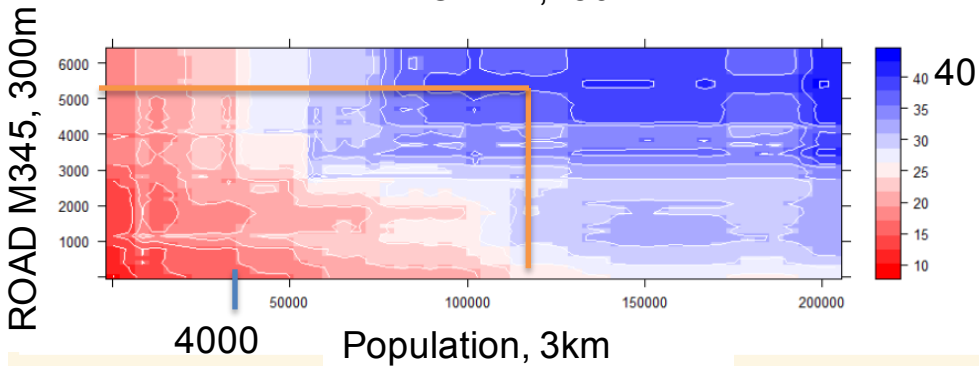
Partial dependent plots: Random forest



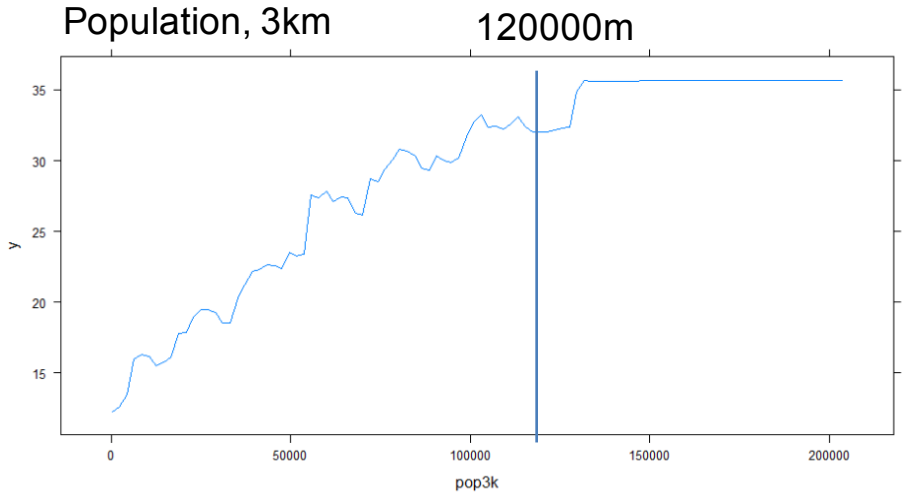


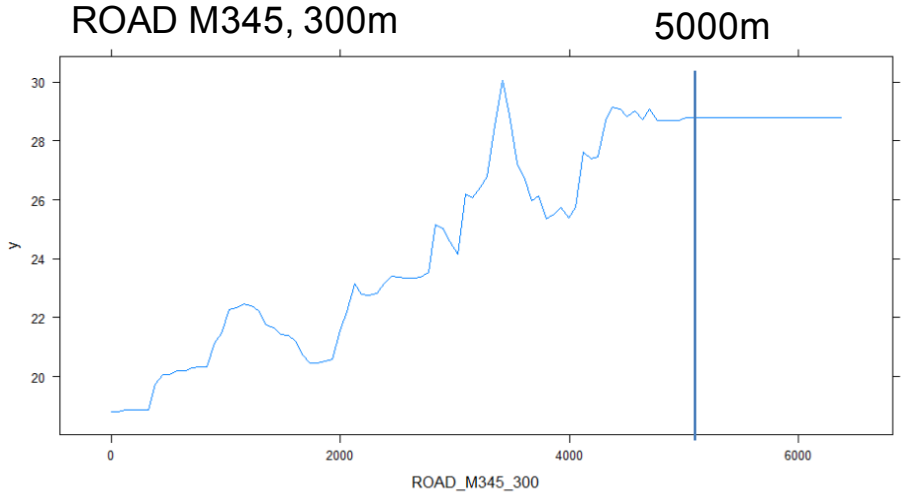


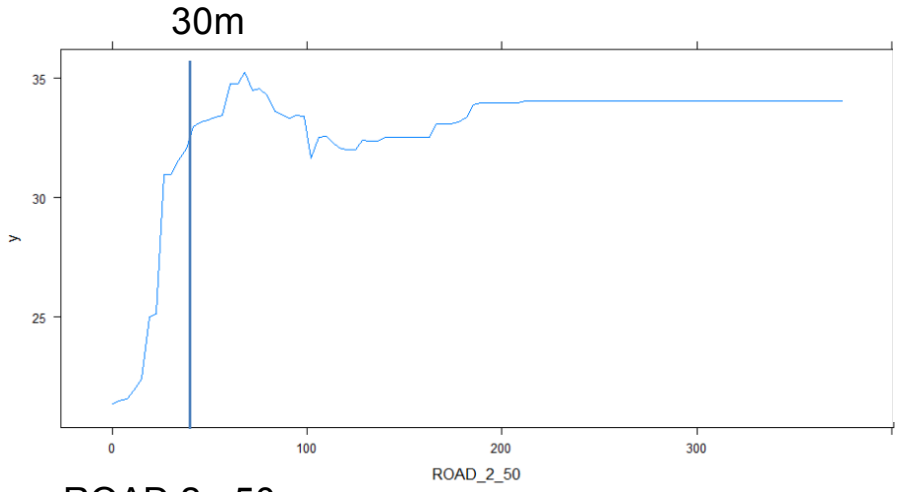




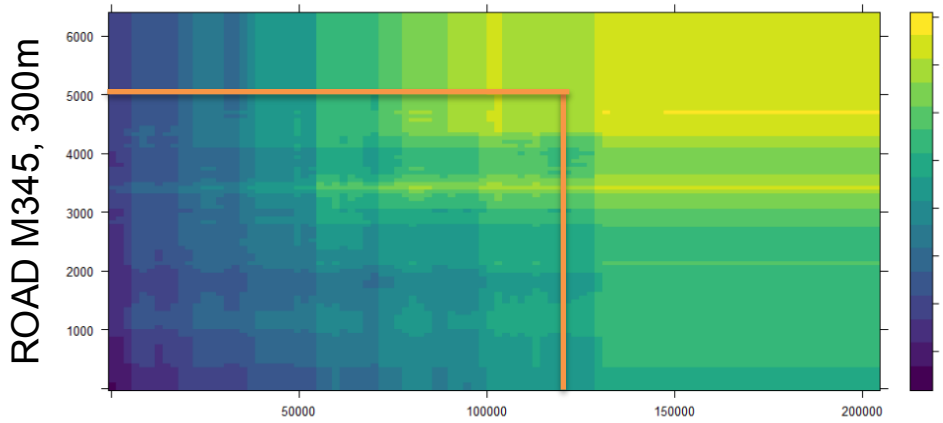
Partial dependent plots: stochastic gradient boosting







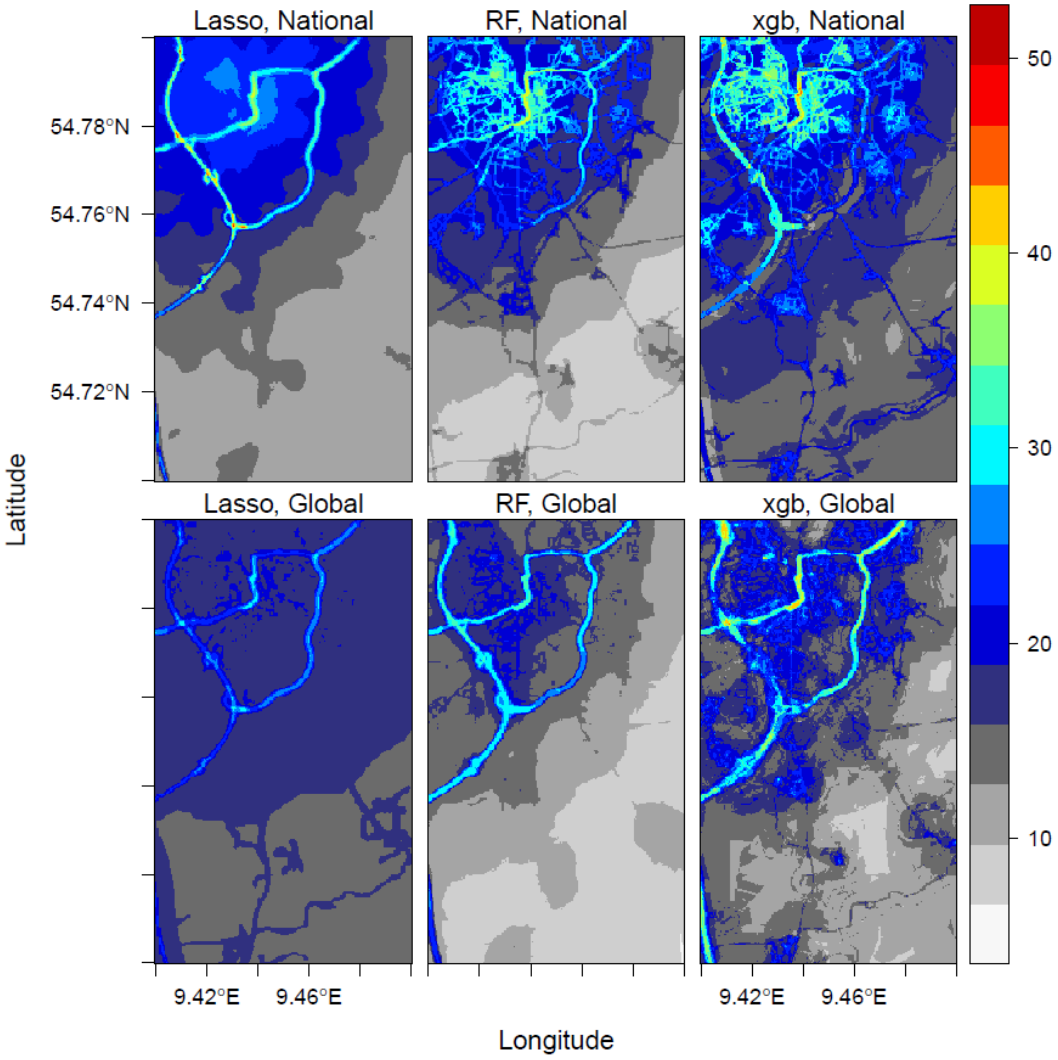
ROAD 2, 50m



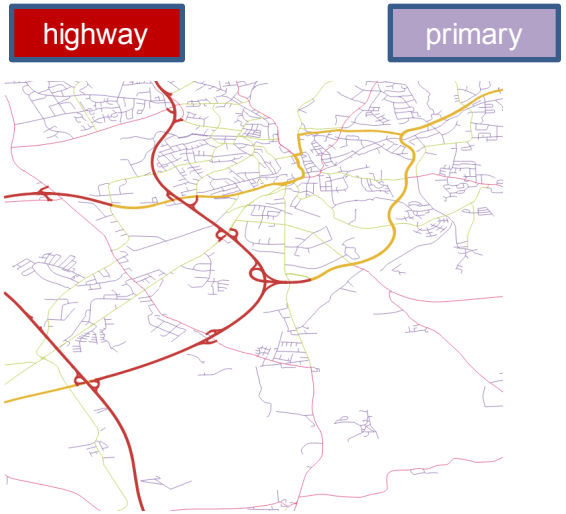
Population, 3km

Part 3

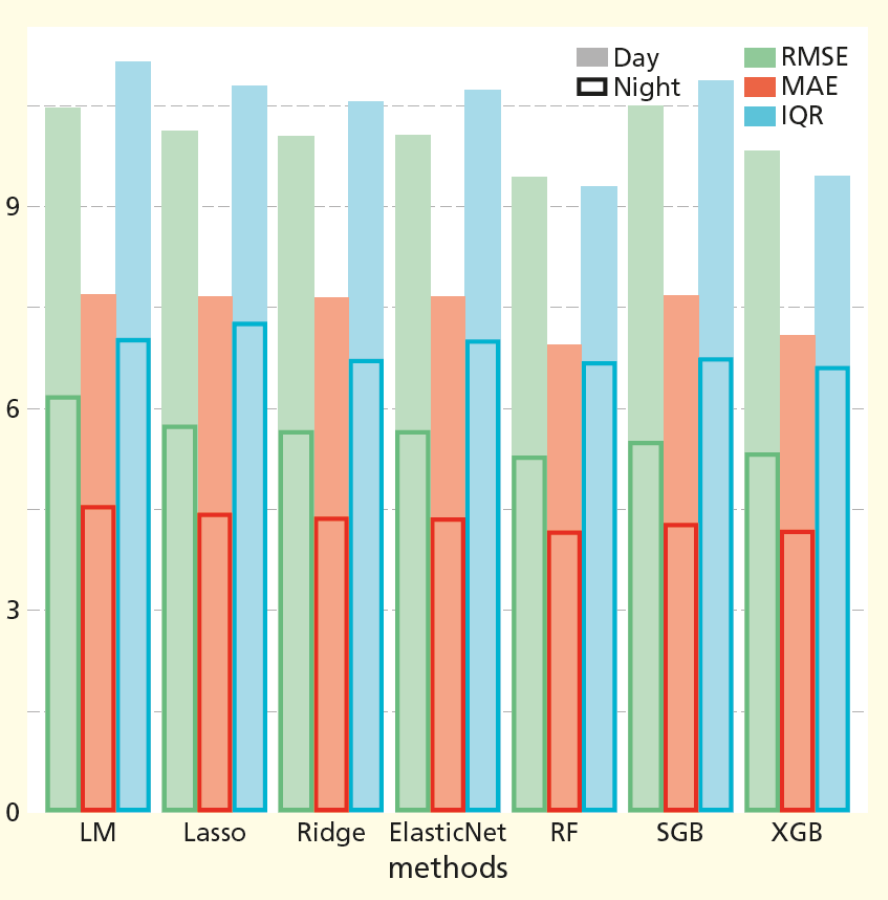
predicting and results


Lasso, National
RF, National
xgb, National
Lasso, Global
RF, Global
xgb, Global
54.78°N
54.76°N
54.74°N
54.72°N
9.42°E
9.46°E
Latitude
Longitude
50
40
30
20
10

Germany


highway
primary

Germany

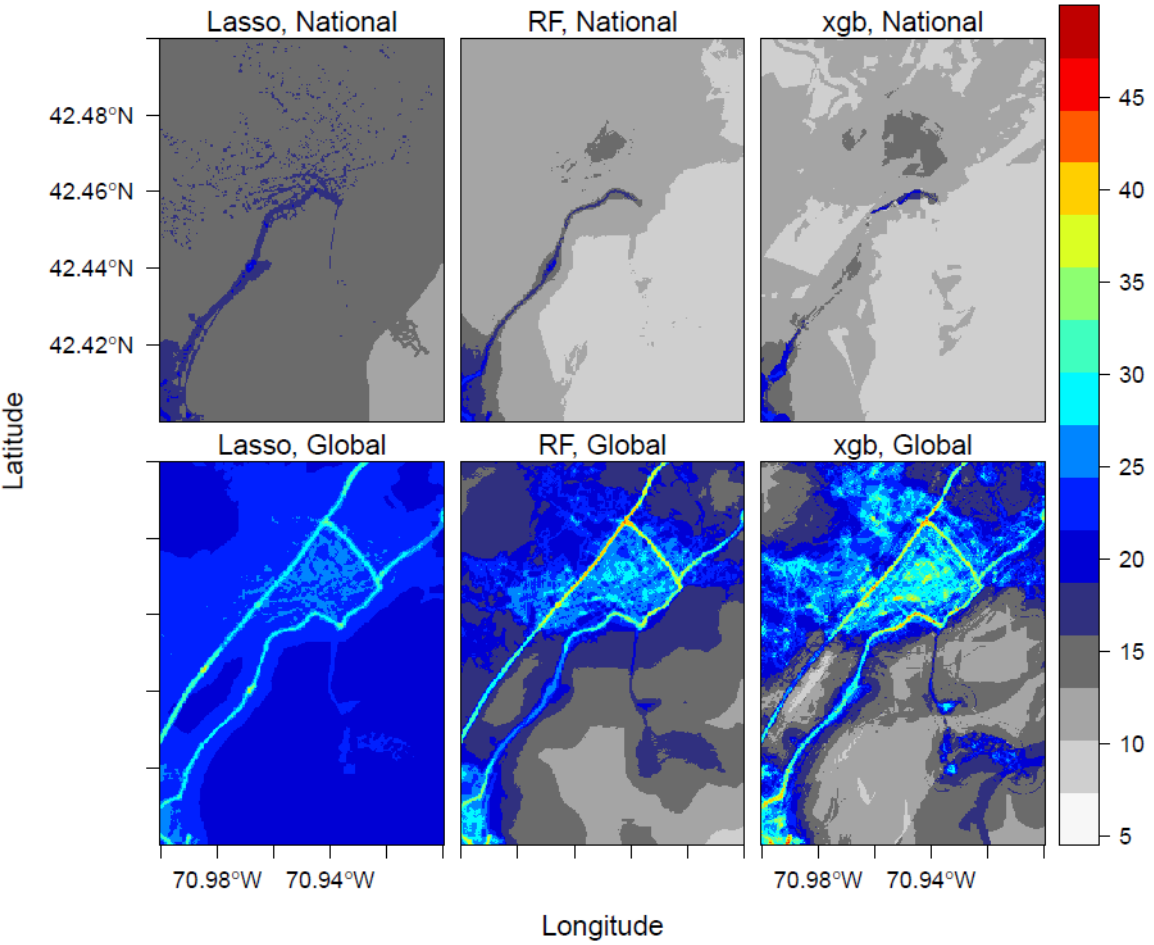


RMSE: root mean squared error
MAE: mean absolute error
IQR: interquartile range

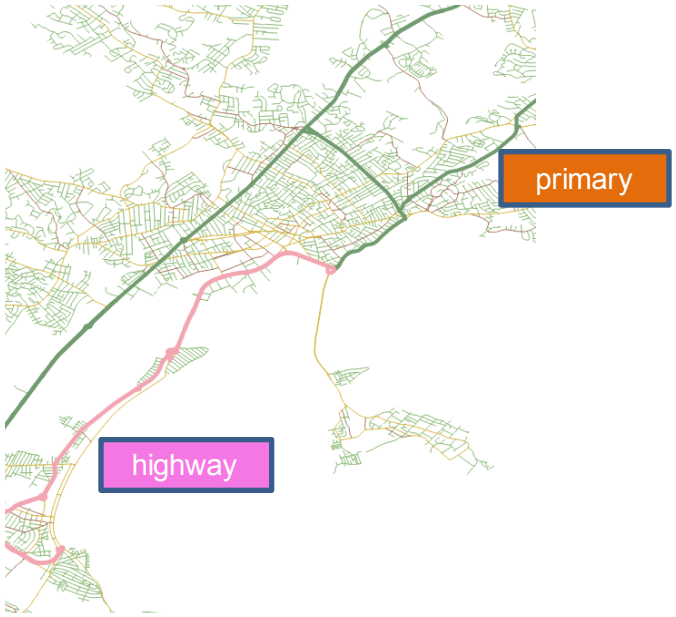
LM: Multiple linear regression
RF: random forest
SGB: Stochastic gradient boosting
XGB: xgboost

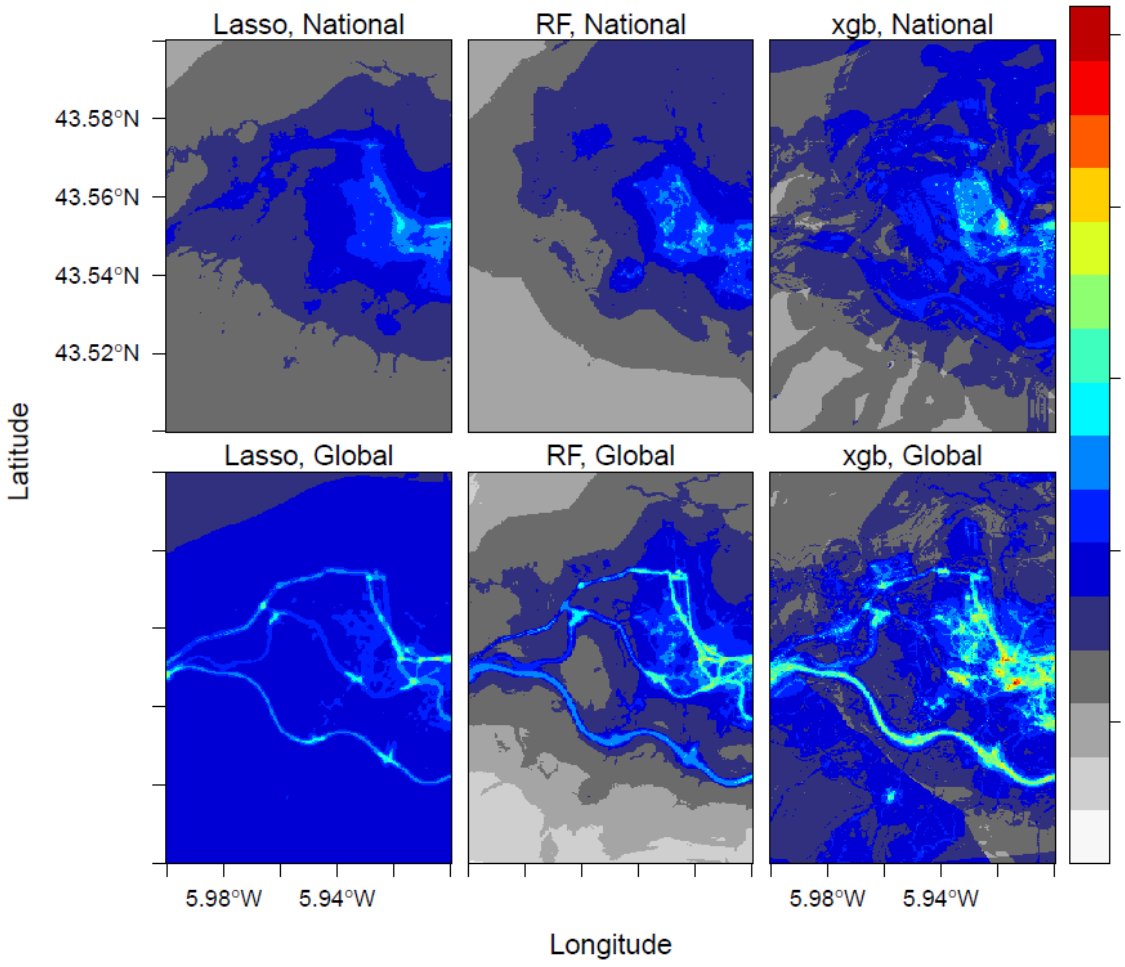
Important emition-related variables

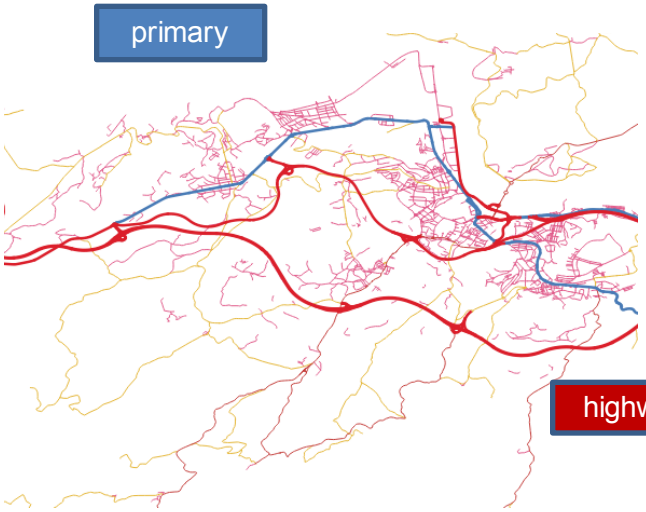
National model of Germany	Global model
Ranked top 20 by Random Forest • Primary road 25m, 50 m, 100 m • Local road 25m, 50 m, 100 m, 300 m	**Ranked top 20 by Random Forest*** • Primary road 50 m, 100 m *Highway 100 m ranked 26
Ranked top 20 by XGBoost • Primary road 50 m, 100 m • Highway 50 m • Local road 25m, 50 m, 100 m, 300 m	**Ranked top 20 by XGBoost** • Primary road 50 m, 100 m • Highway 100 m • Local road 25 m, 50 m, 100 m
Selected by LASSO • Primary road 25m, 50 m, 100 m • Highway 50m, 100 m • Local road 100 m, 300m	**Selected by LASSO** • Primary road 50 m, 100 m • Highway 50 m, 100 m • Local road 25 m, 50 m, 100 m


Lasso, National
RF, National
xgb, National
Lasso, Global
RF, Global
xgb, Global
42.48°N
42.46°N
42.44°N
42.42°N
Latitude
70.98°W
70.94°W
Longitude
45
40
35
30
25
20
15
10
5

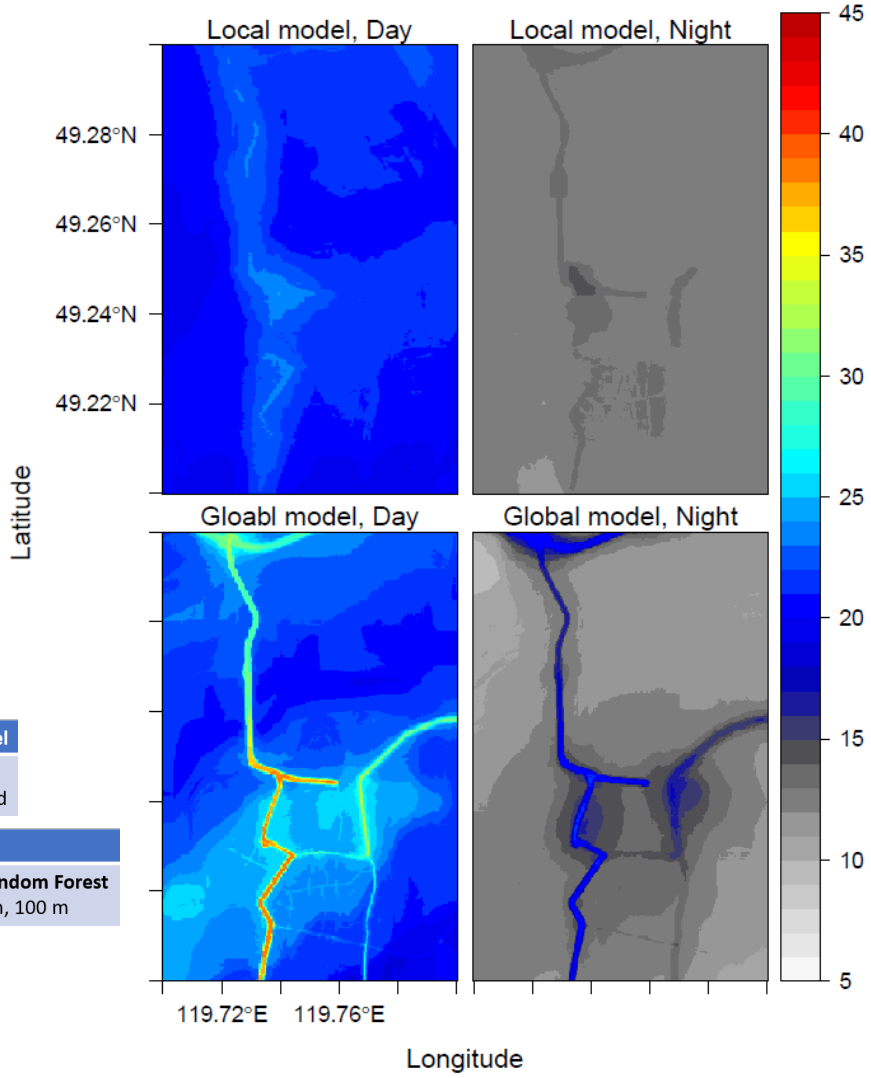
US


primary
highway


Lasso, National
RF, National
xgb, National
Lasso, Global
RF, Global
xgb, Global
43.58°N
43.56°N
43.54°N
43.52°N
5.98°W
5.94°W
Latitude
Longitude
50
40
30
20
10


Spain
primary
highway

Random Forest Prediction: China



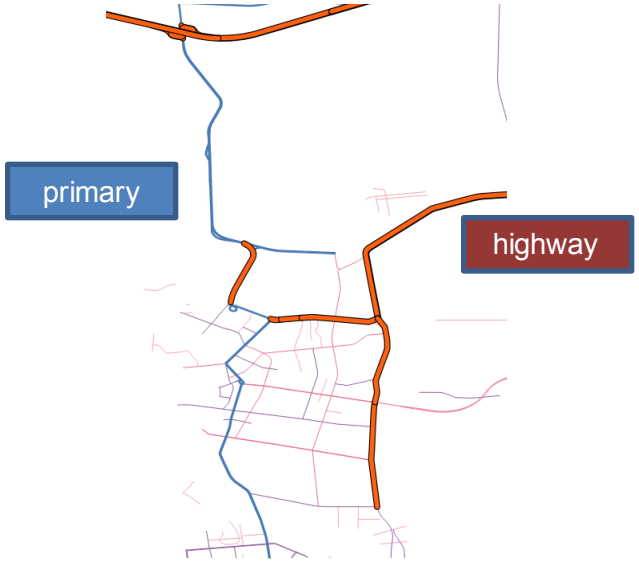
China national model

Only background variables are selected

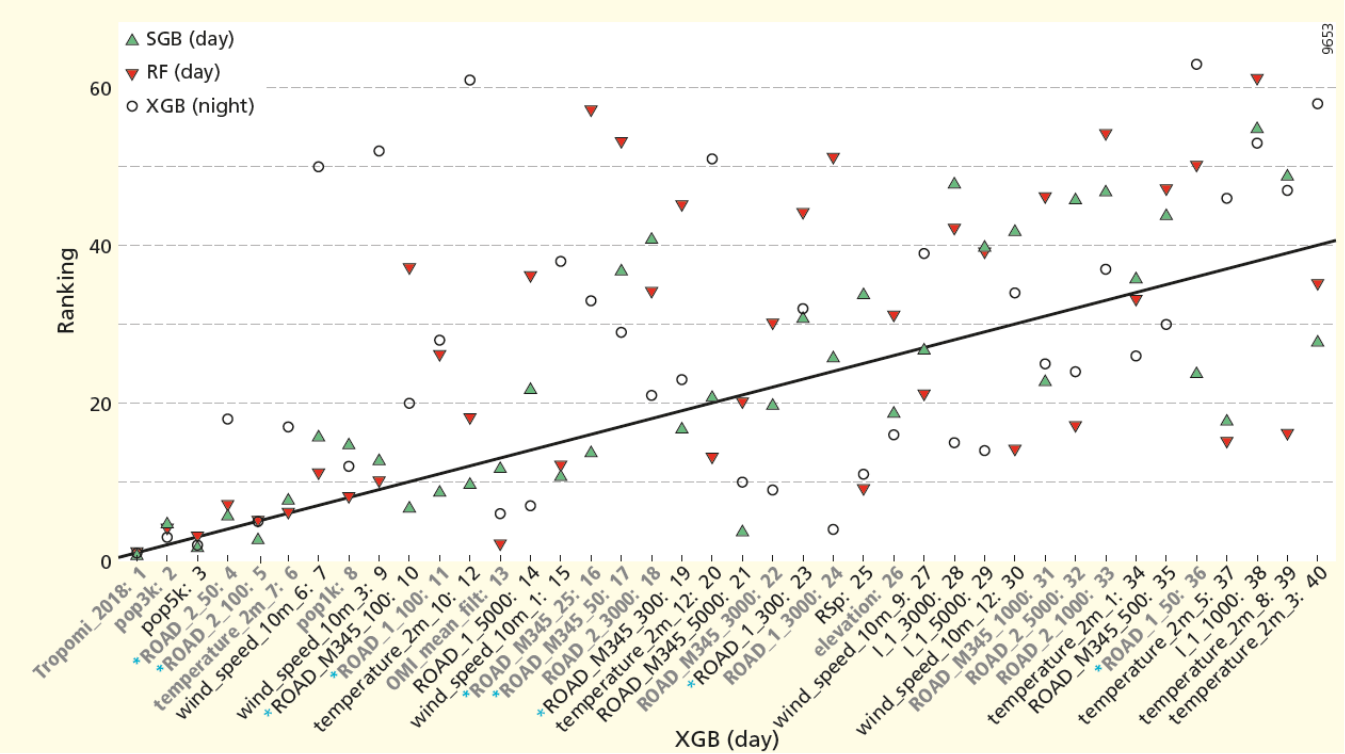
Global model

Ranked top 20 by Random Forest

- Primary road 50 m, 100 m



Variable importance



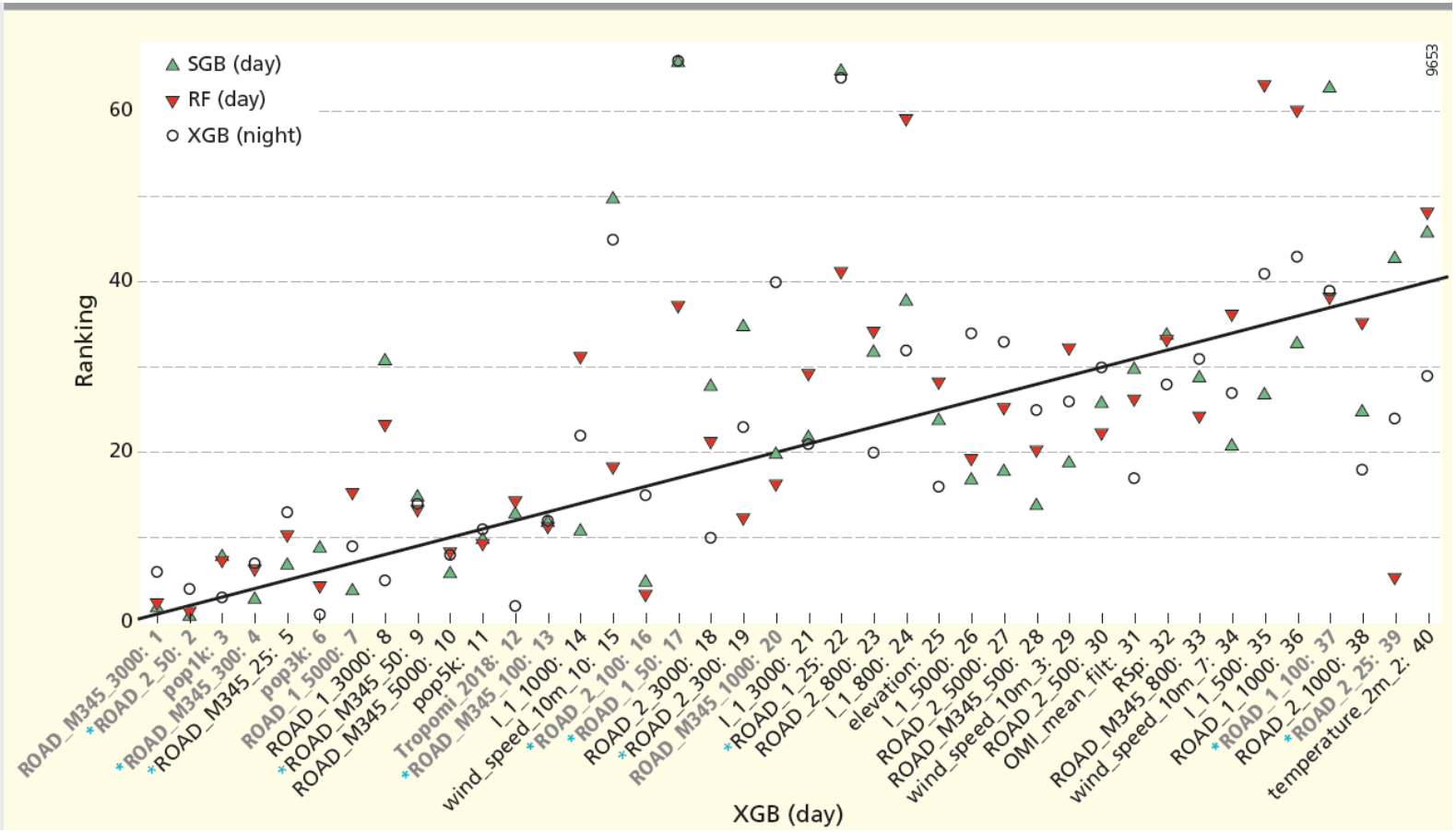
ROAD_M345: secondry and local roads

Pop_: population

ROAD_2: primary roads

Global model

Variable importance



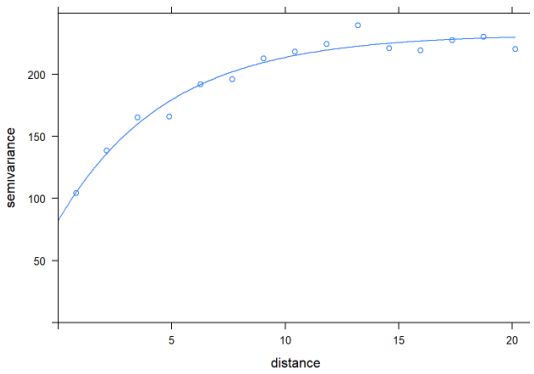
ROAD_M345: secondry and local roads

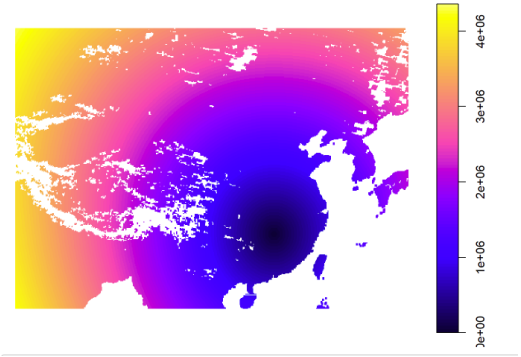
Pop_: population

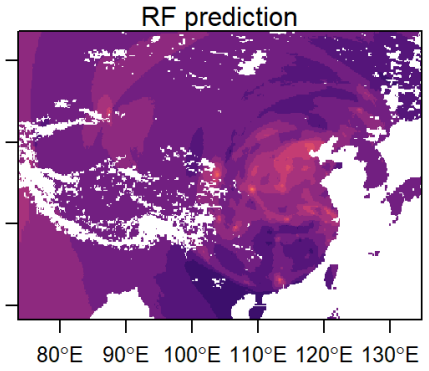
ROAD_2: primary roads

Germany

Using random forest for Geostatistic-like interpolation







http://rpubs.com/menglu/473973