Statistical methods of global air pollution modeling

University of Utrecht, The Netherlands

Meng Lu

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.

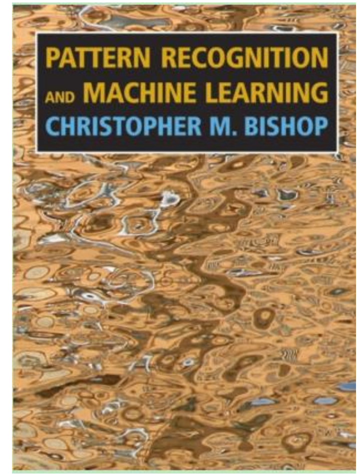
Enlarge This Image

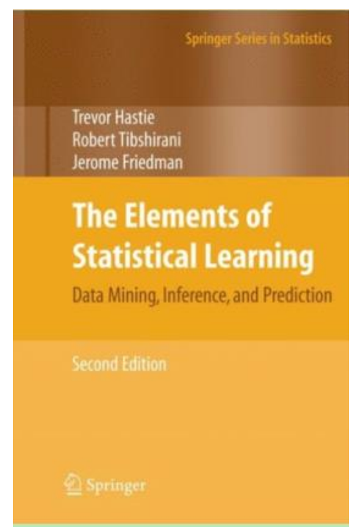
Thor Swift for The New York Times
Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

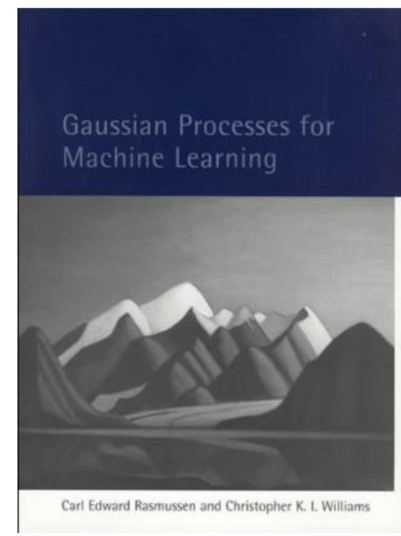
"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

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)

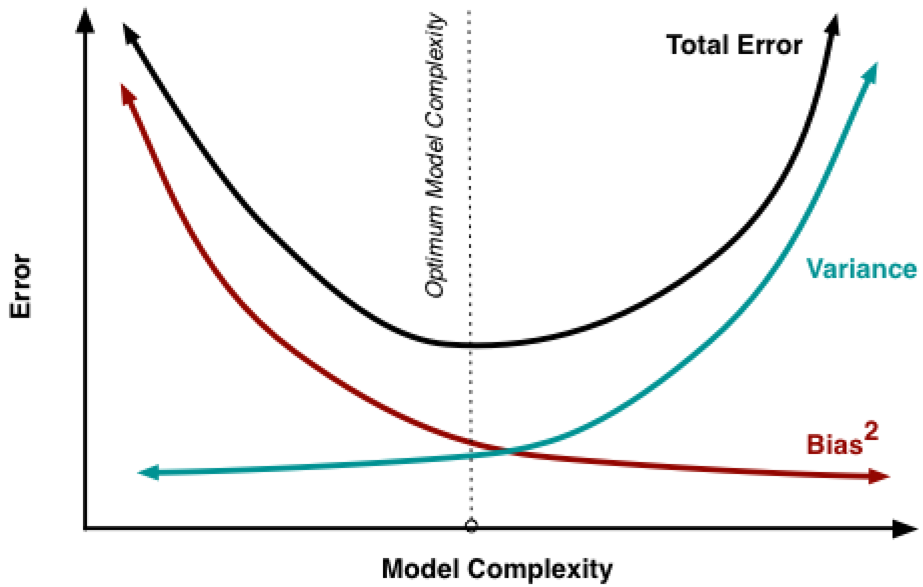
$$h: X \rightarrow Y$$

Y : categories; continuous data, graphic output

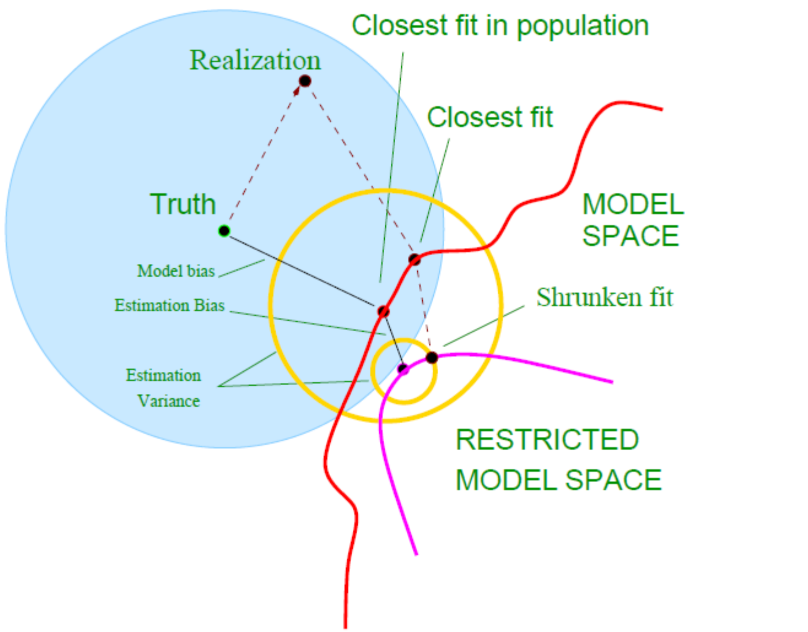
Bias-variance trade-off

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

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



All algorithms are affected by bias-variance trade-off

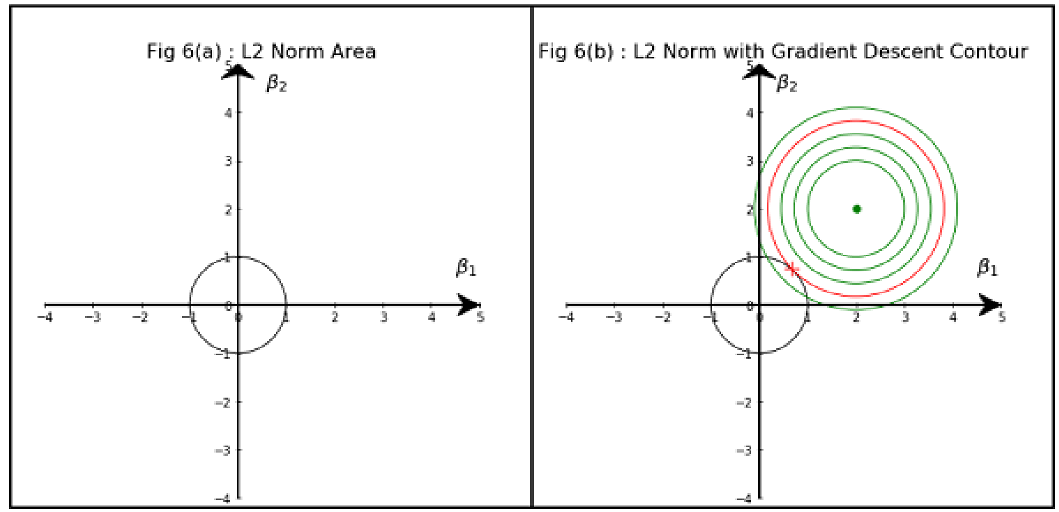


Schematic of the behavior of bias and variance.

Regularization

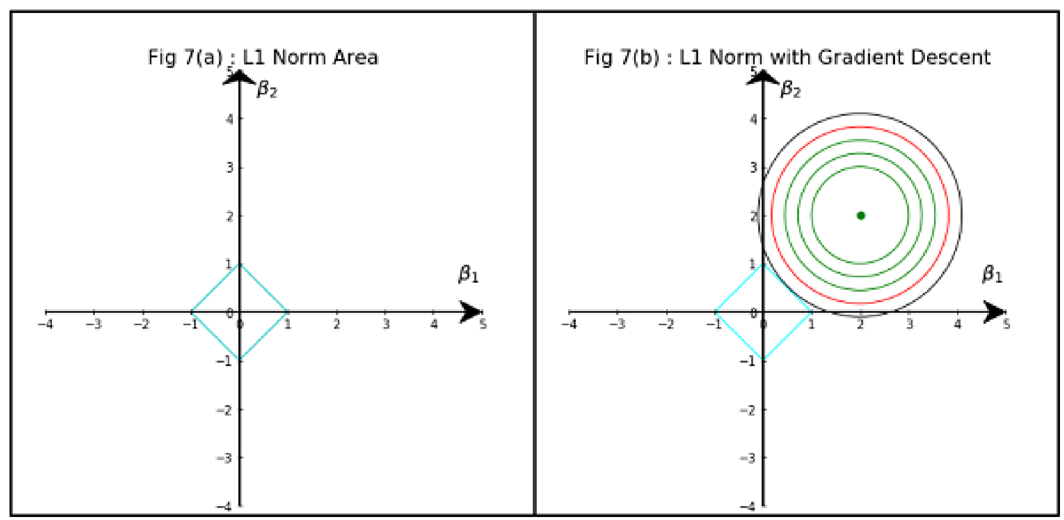
Ridge regression

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

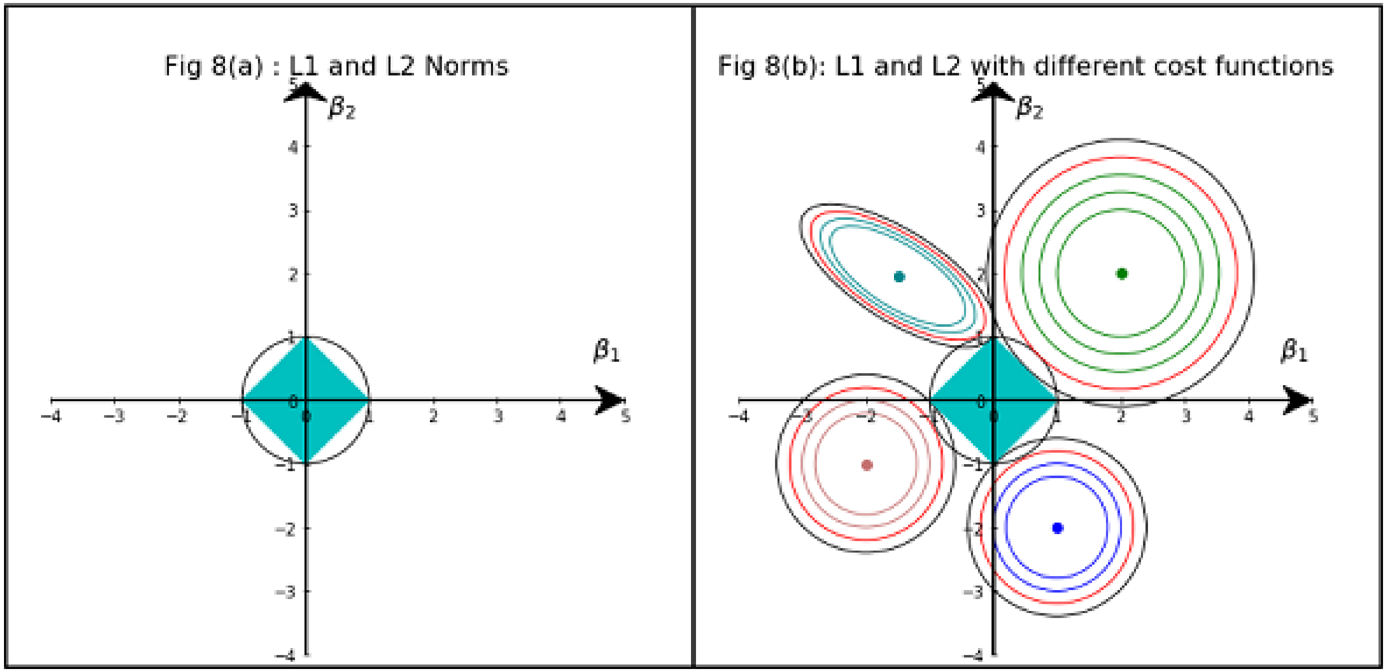


Lasso regression

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



Lasso vs. Ridge
ElasticNet



Regression trees

Features:

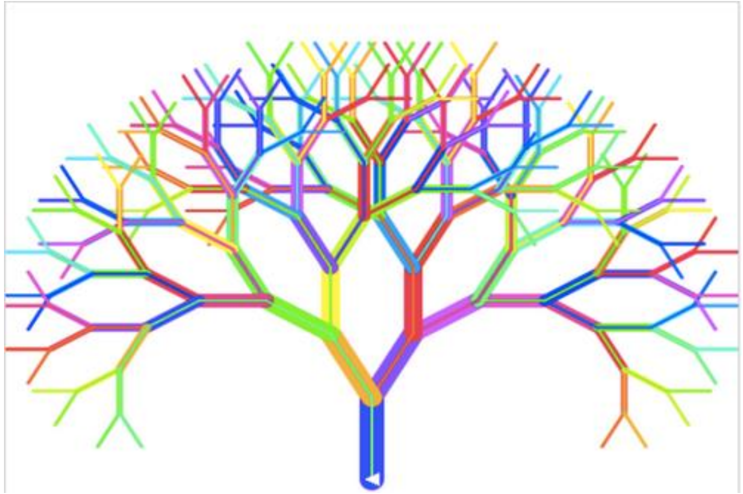
- Non-parametric
- Different kinds of variables
- Redundant variables are ignored
- Handle missing data
- Small trees are easy to interpret

Leveraging trees to improve the performance:

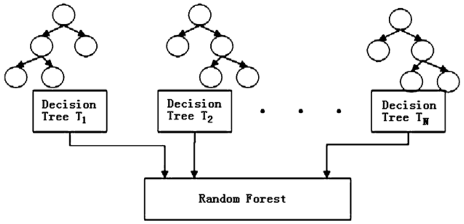
- Bagging
- Boosting
- Random forest

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 $\frac{1}{B}\sigma^2$

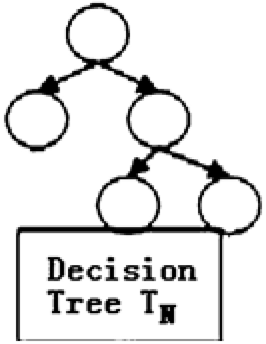
If the variables are not independent (but identically distributed) with positive pairwise correlation p , the variance of the average:

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

“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

General boosting and gradient boosting

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

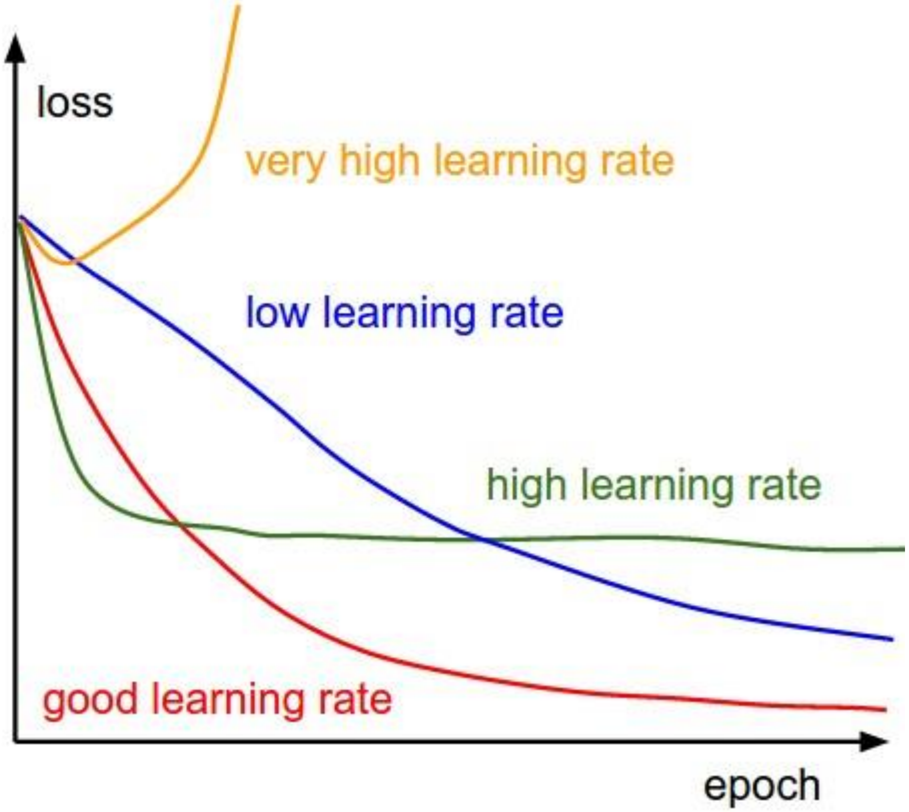
$$\text{Set } F_m(x) = F_{m-1}(x) + \epsilon \beta_m b(x; \gamma_m)$$

(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

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

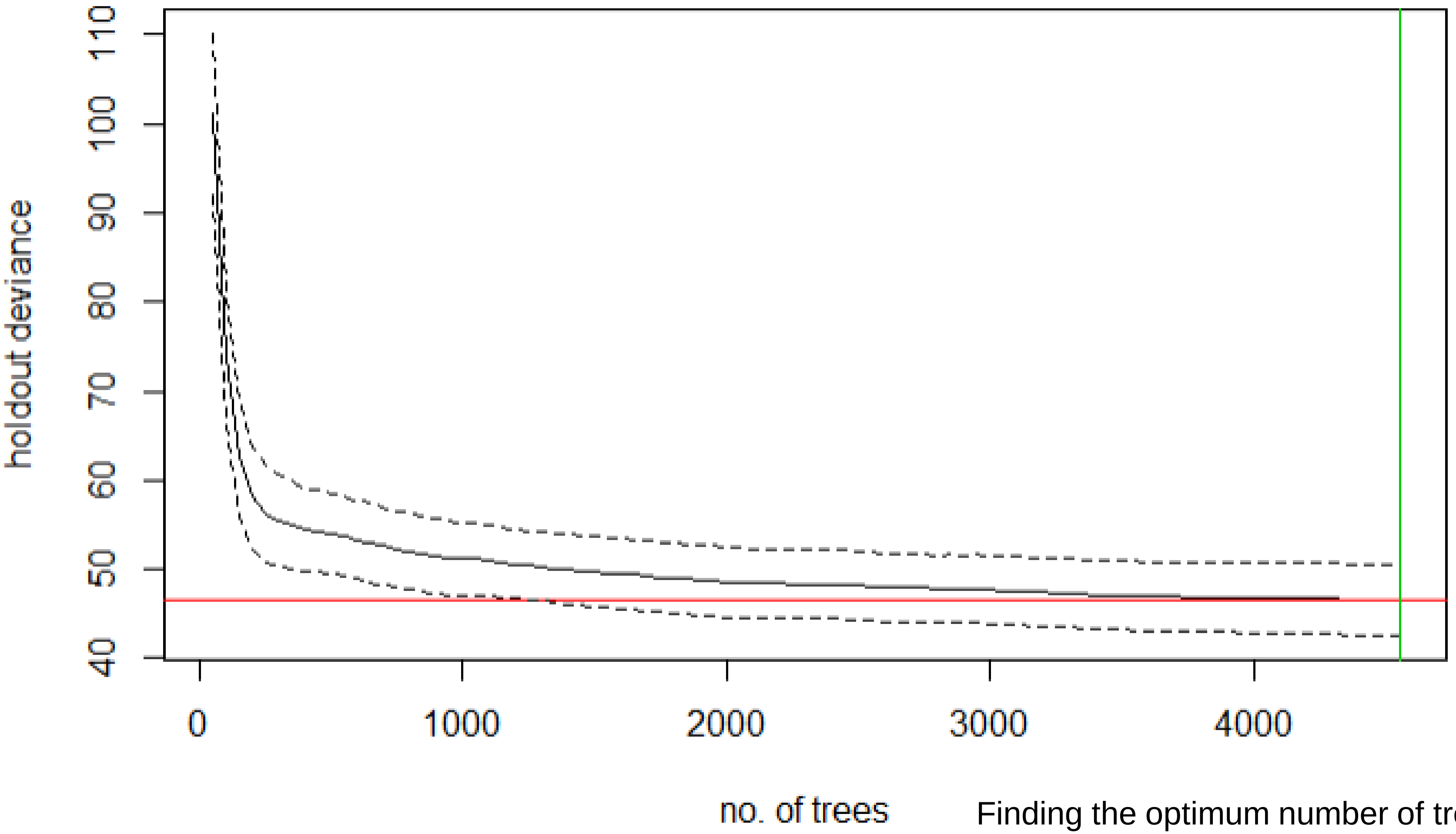
Idea
Not only impurity, but also model complexity

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

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

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

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

Postprocessing

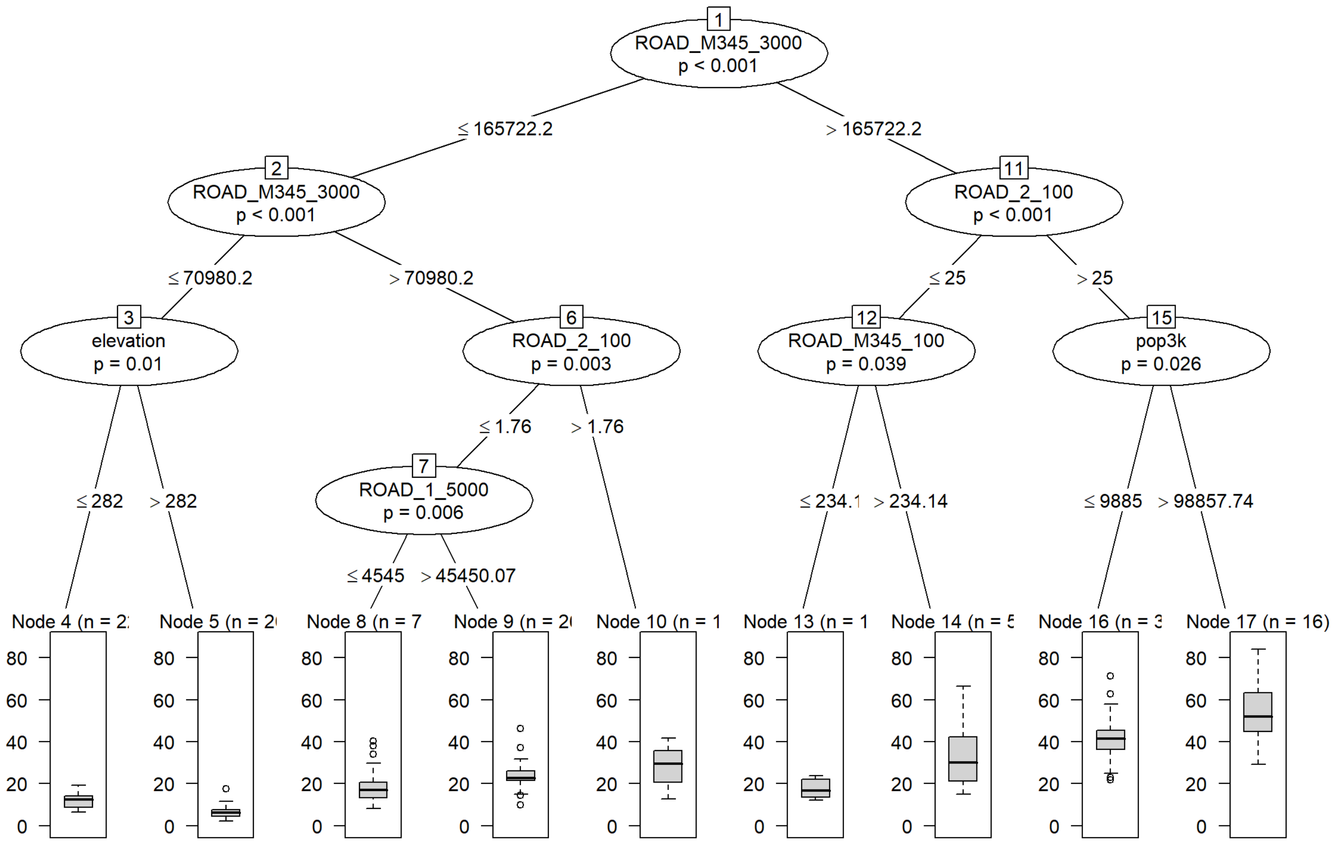
Lasso regularization of regression trees
--- discarding trees that are not useful

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

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.

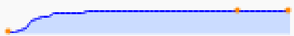
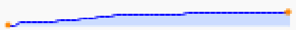
$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C)d\mathbb{P}(x_C)$$

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,

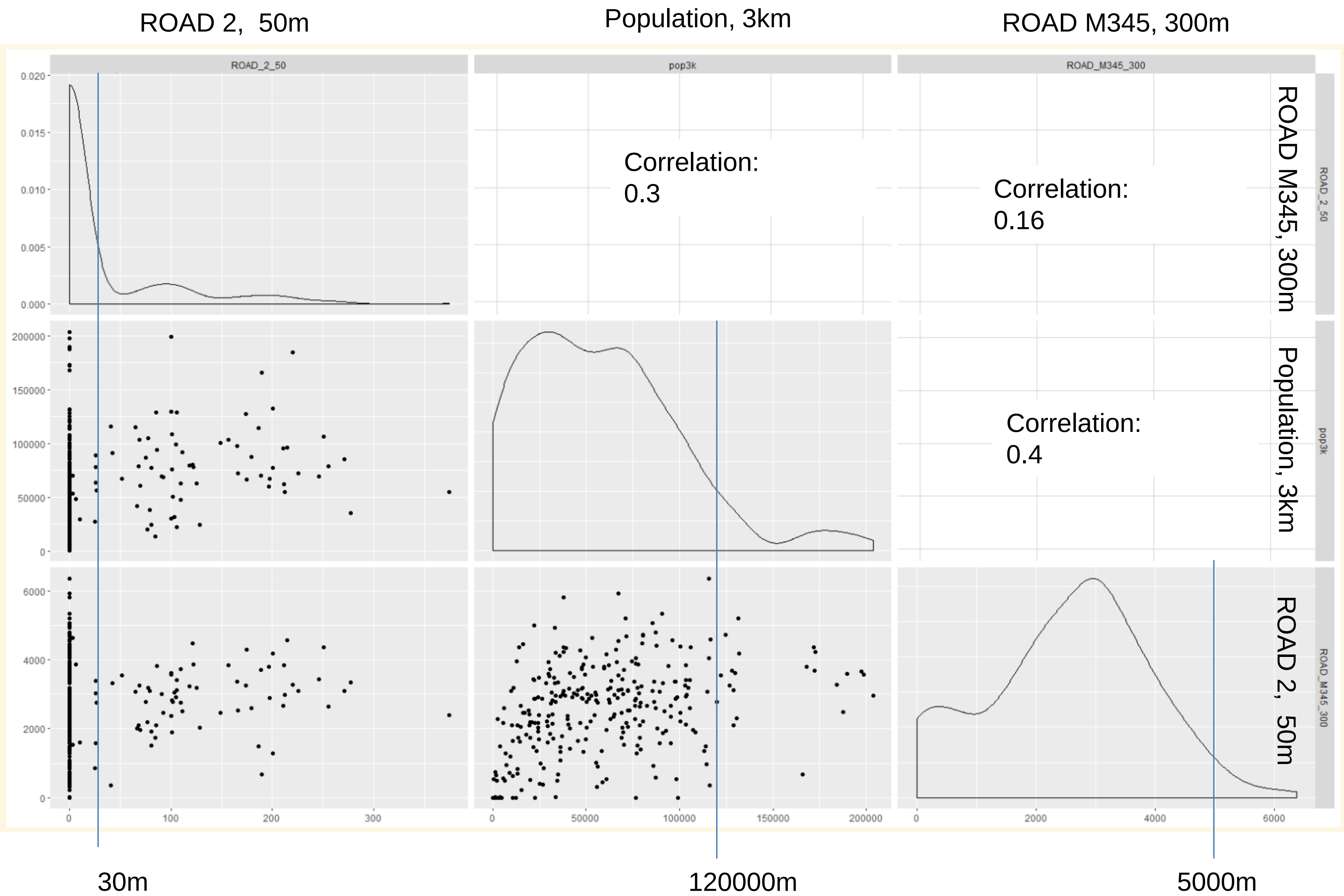
Assumption: the features in C are not correlated with the features in S

Show 10 entries Search:

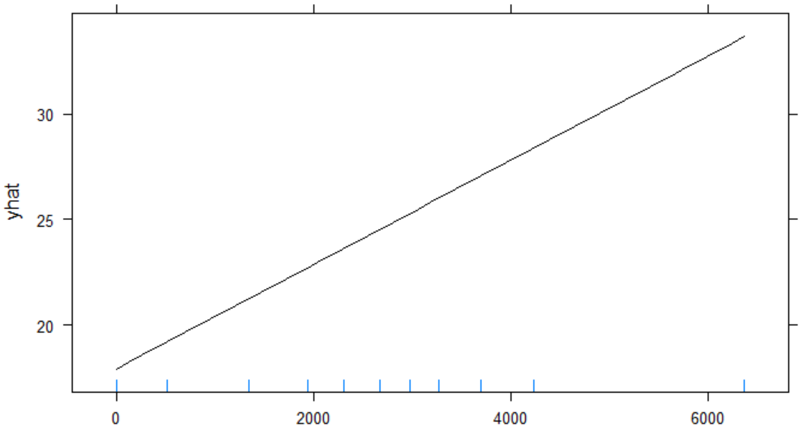
	Variable	Importance	Effect
1	ROAD_2_50	3.032	
2	ROAD_M345_3000	1.542	
3	pop3k	1.379	
4	ROAD_2_100	1.084	
5	ROAD_M345_300	1.058	
6	pop5k	0.840	
7	pop1k	0.756	
8	ROAD_M345_5000	0.674	
9	Tropomi_2018	0.654	
10	ROAD_M345_100	0.578	

Showing 1 to 10 of 65 entries

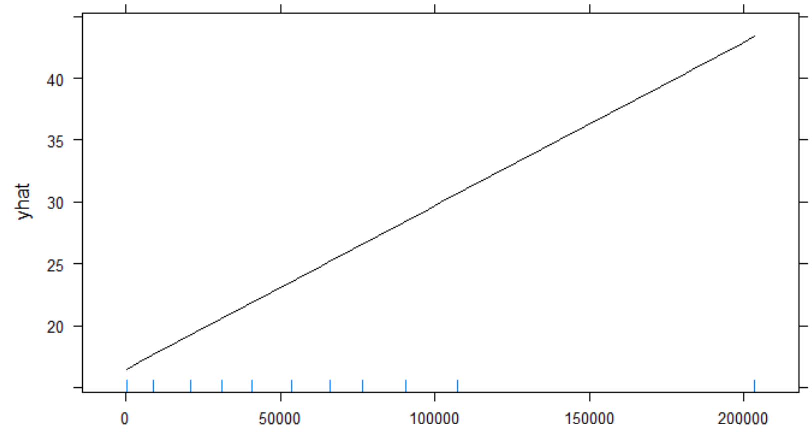
Previous 1 2 3 4 5 6 7 Next


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

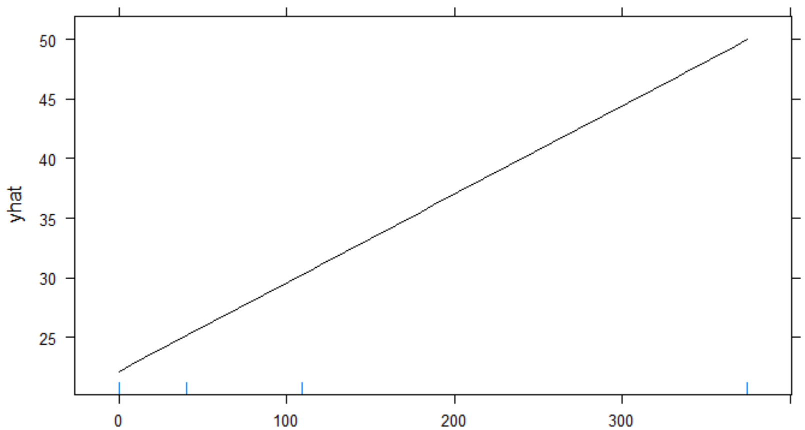
Partial dependent plots: Linear regression



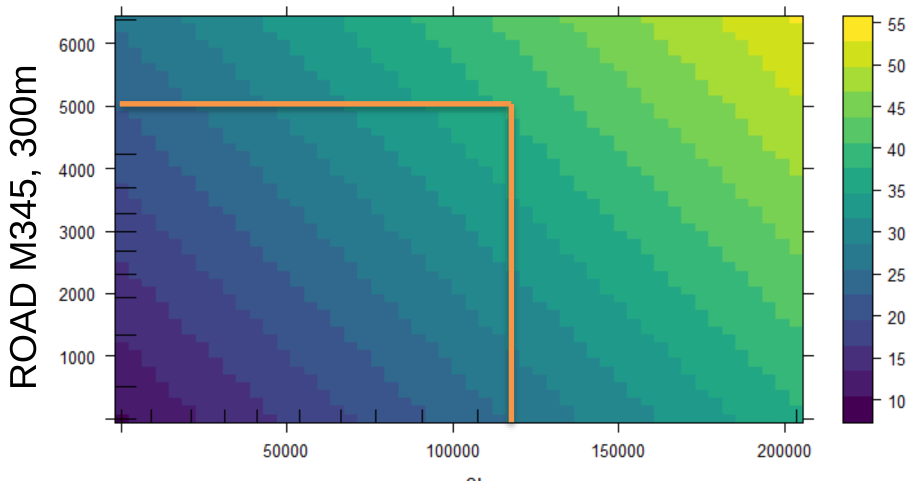
ROAD M345, 300m



Population, 3km

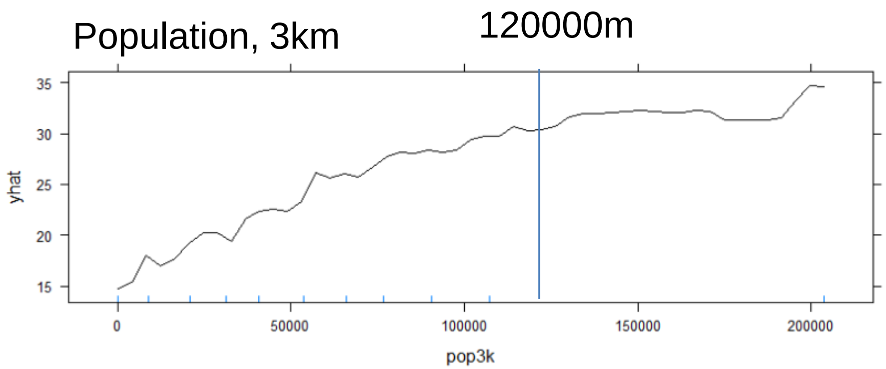


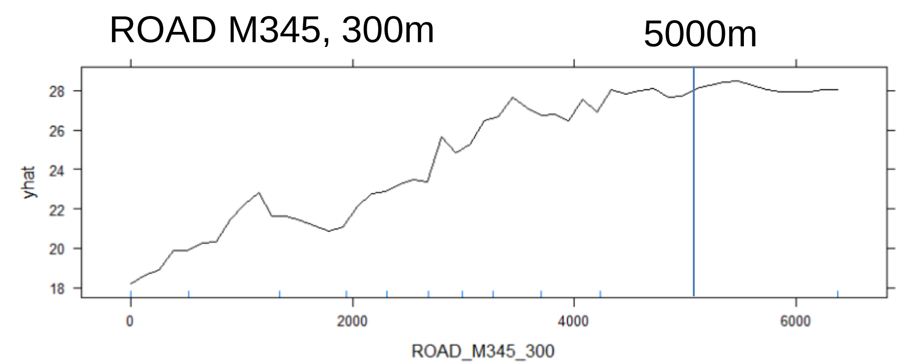
ROAD 2, 50m

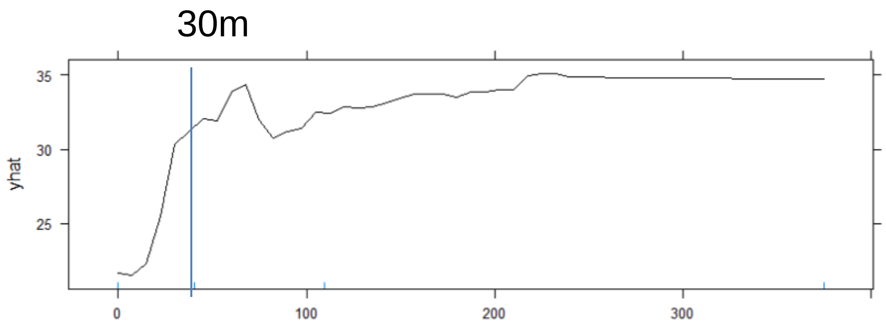


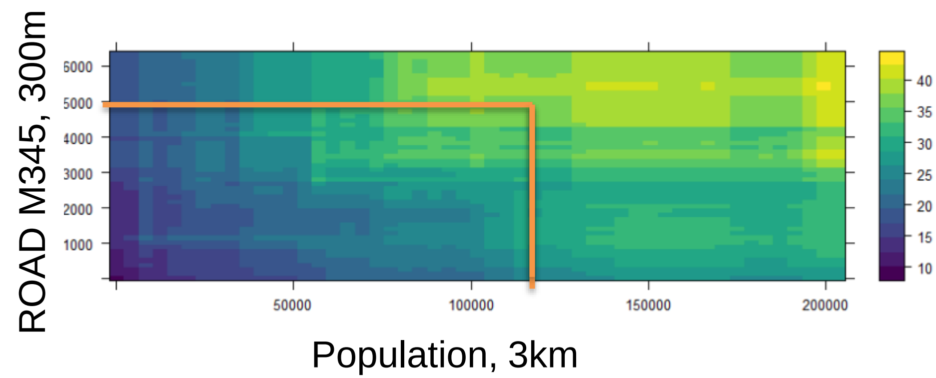
Population, 3km

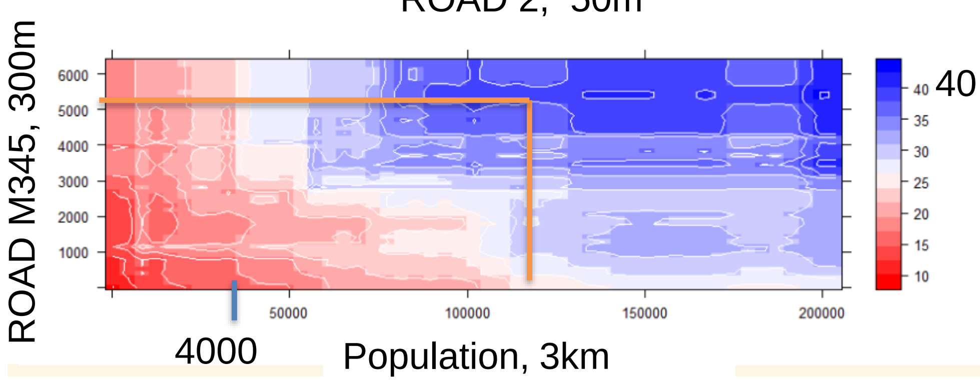
Partial dependent plots: Random forest





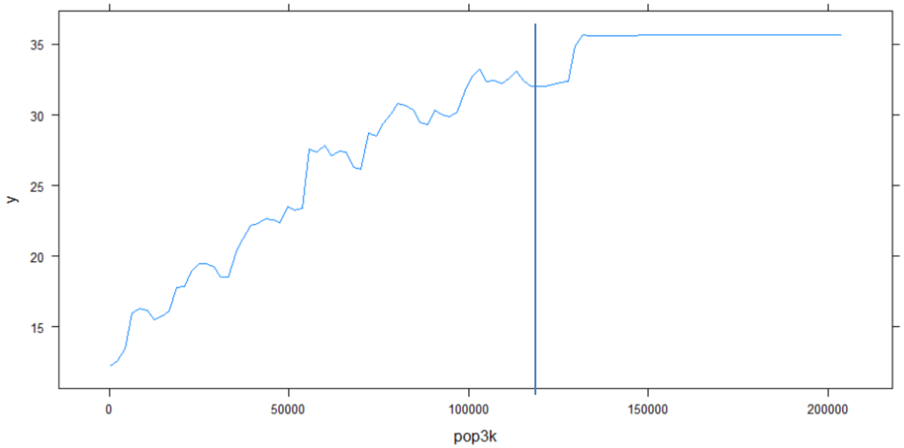




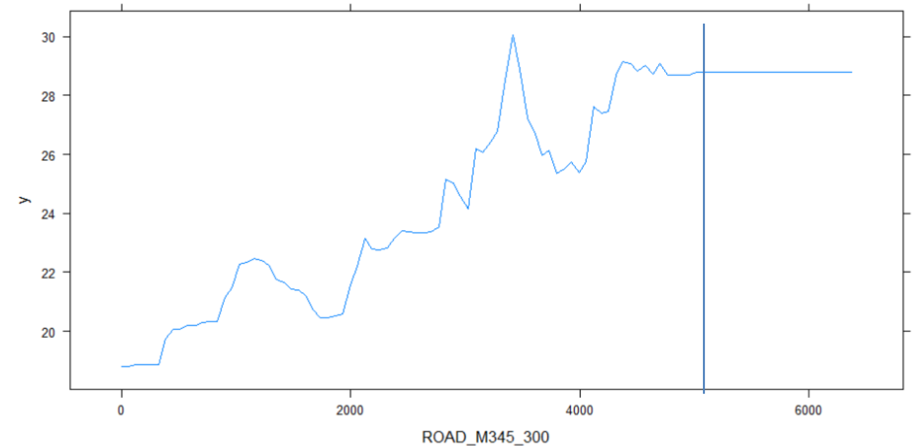


Partial dependent plots: boosted regression trees

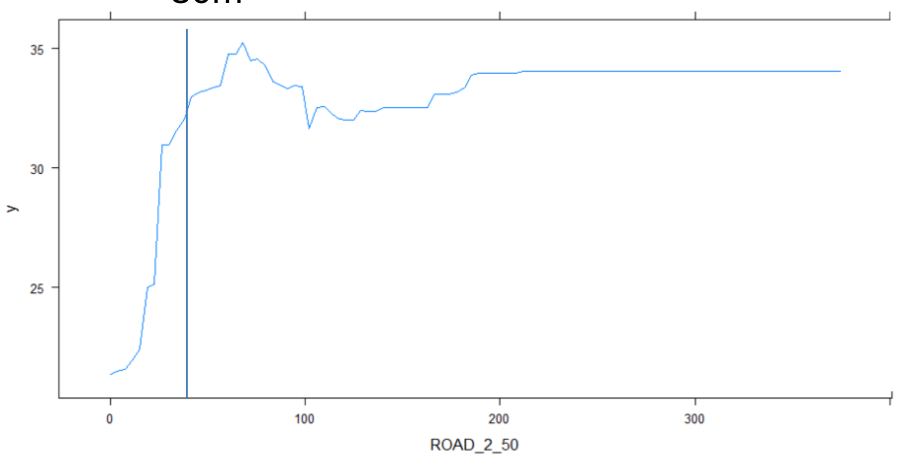
Population, 3km 120000m



ROAD M345, 300m 5000m

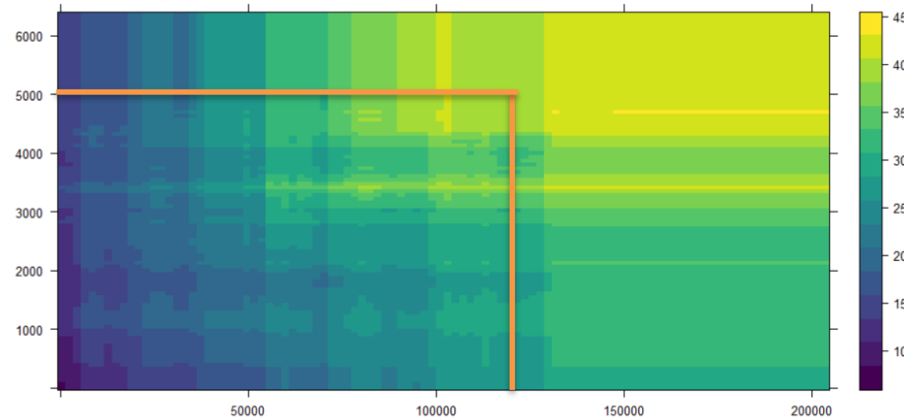


30m



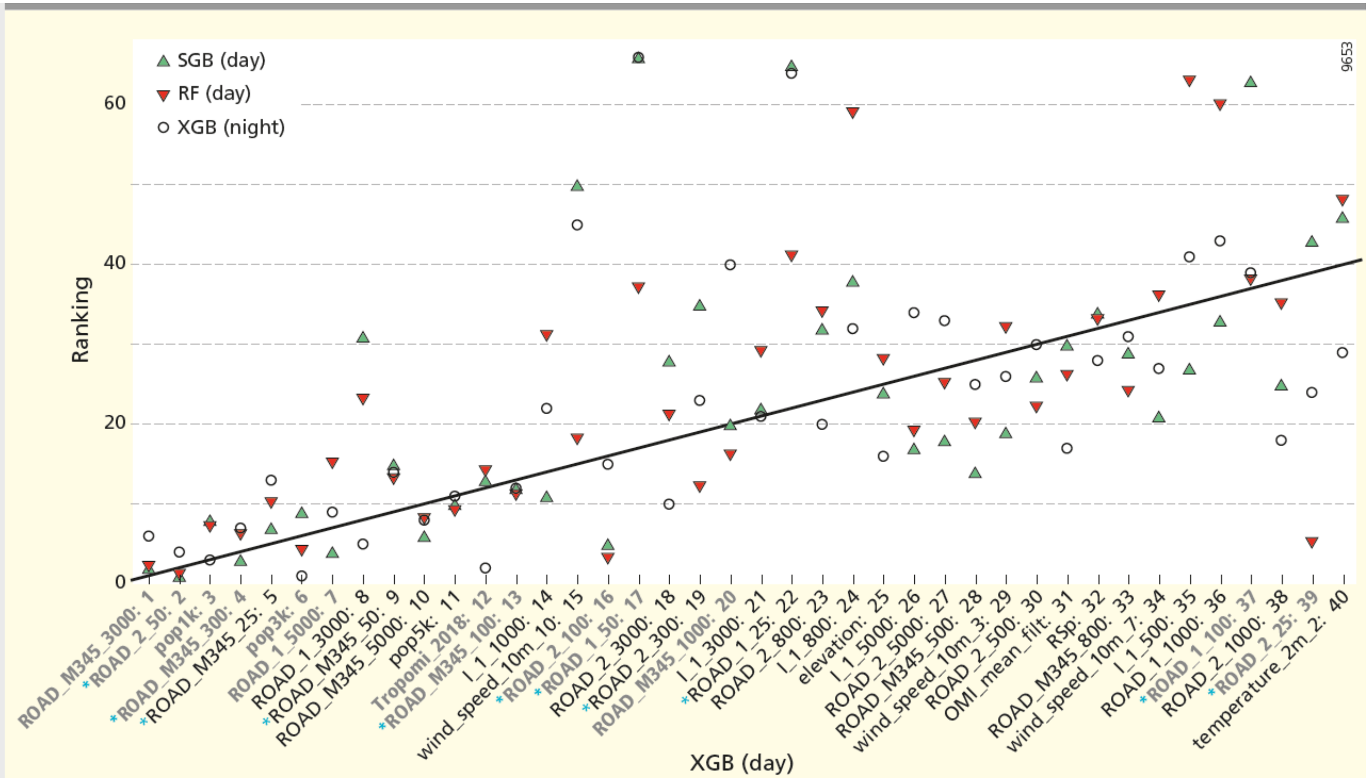
ROAD 2, 50m

ROAD M345, 300m



Population, 3km

Variable importance



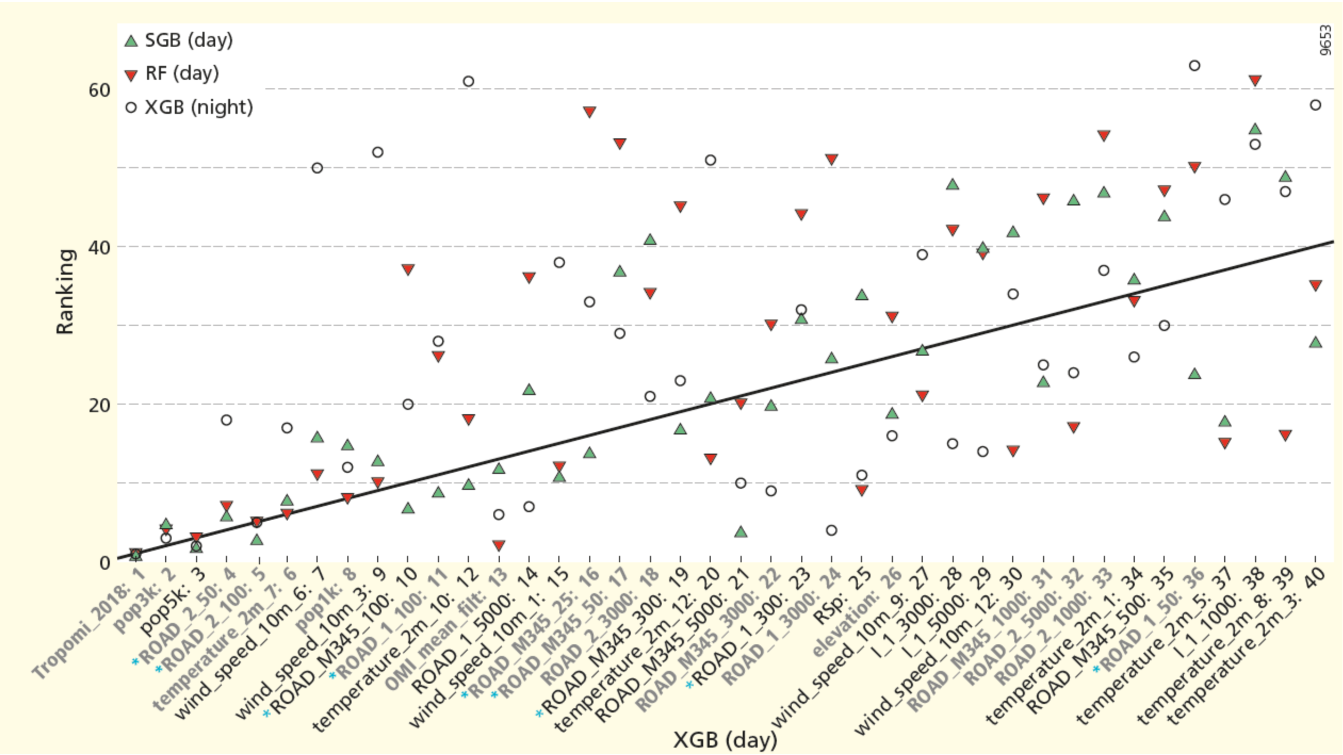
ROAD_M345: secondry and local roads

Pop_: population

ROAD_2: primary roads

Germany

Variable importance

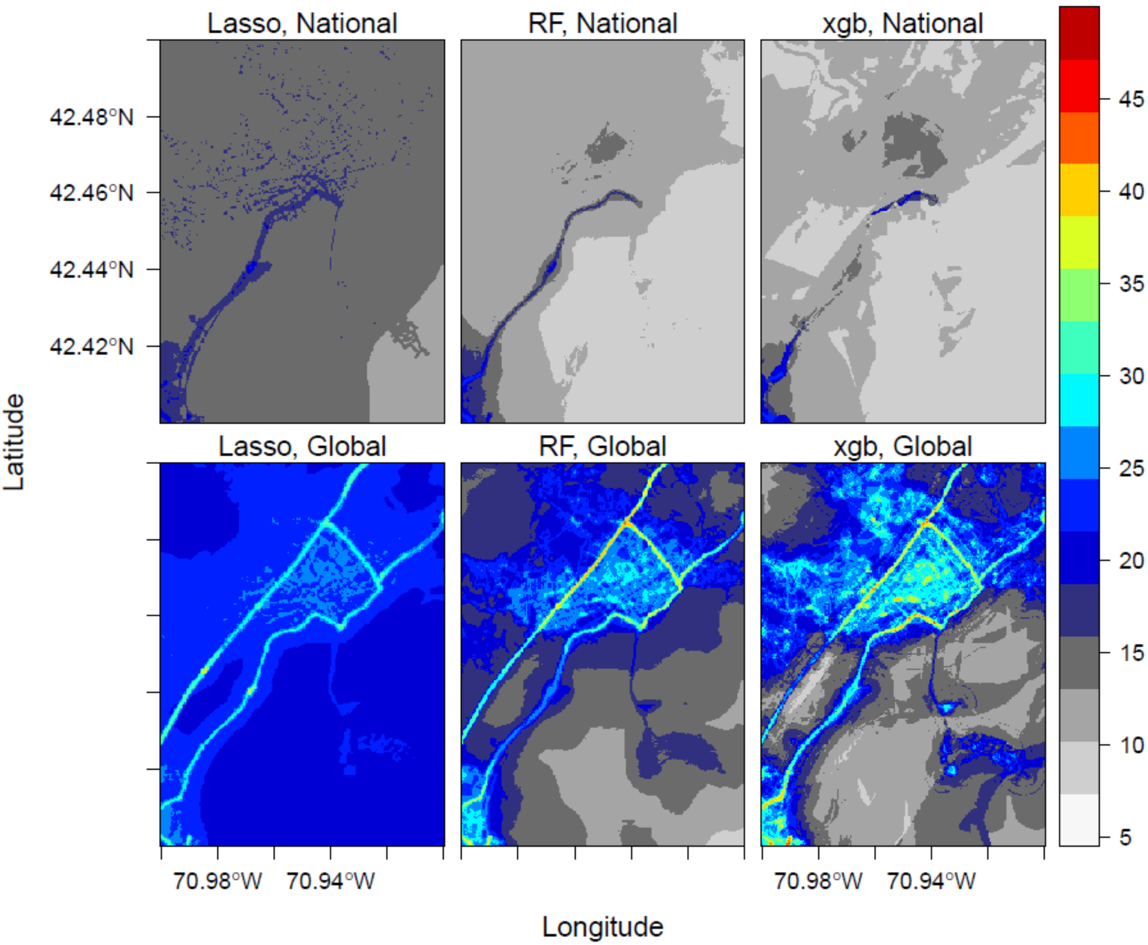


ROAD_M345: secondry and local roads

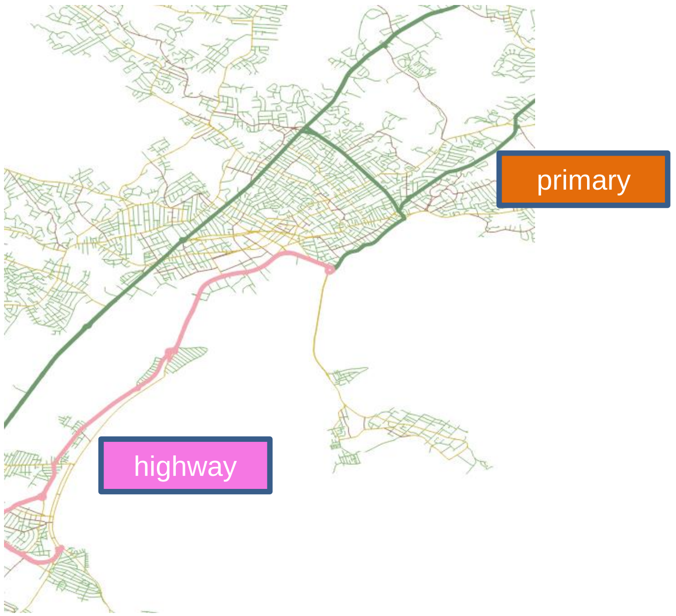
Pop_: population

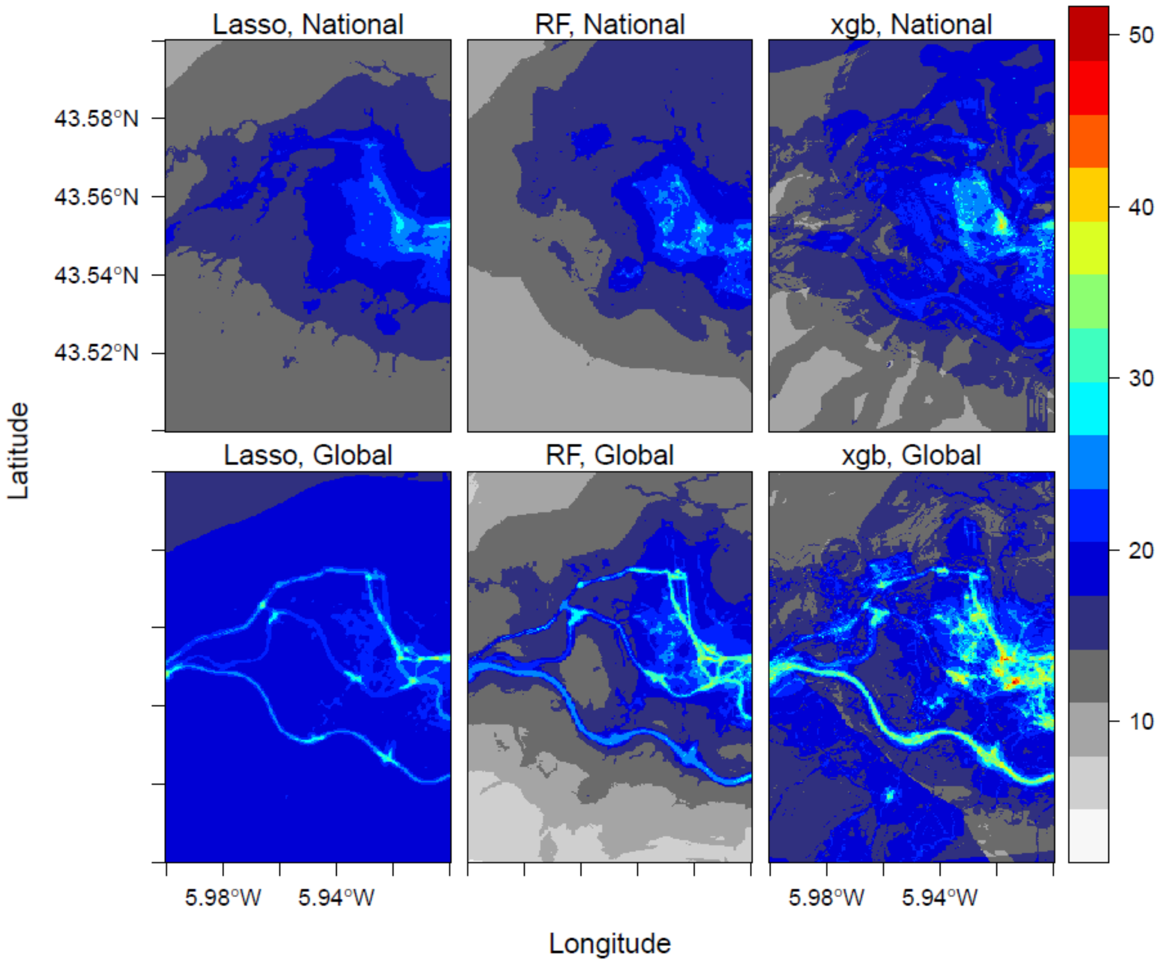
ROAD_2: primary roads

Global model

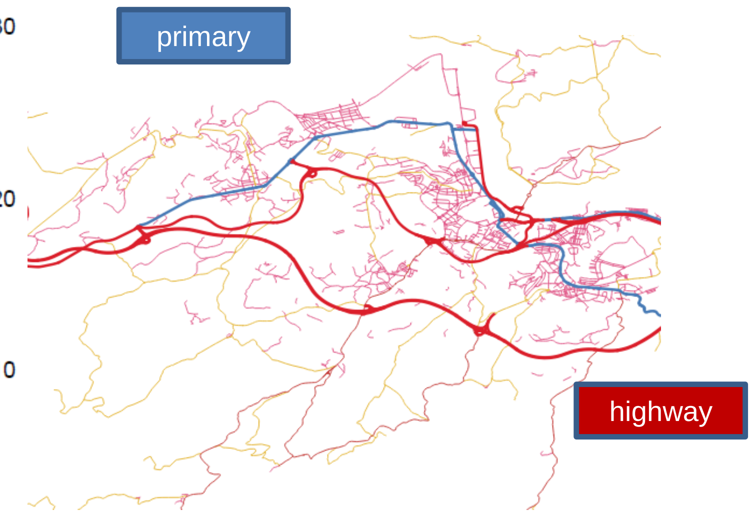

Lasso, National
RF, National
xgb, National
Lasso, Global
RF, Global
xgb, Global
42.48°N
42.46°N
42.44°N
42.42°N
Latitude
70.98°W
70.94°W
Longitude
45
40
35
30
25
20
15
10
5

US

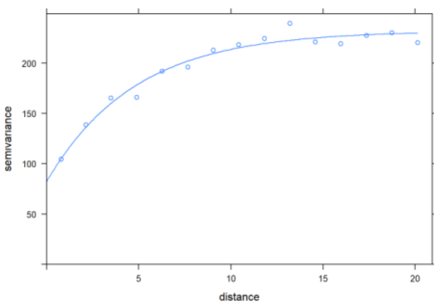

primary
highway

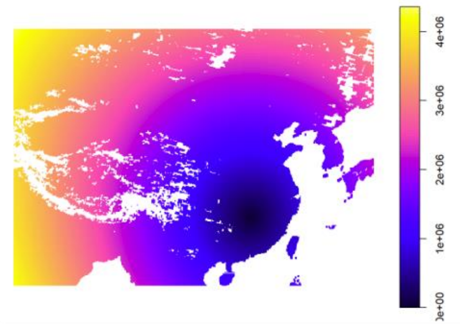

Lasso, National
RF, National
xgb, National
Lasso, Global
RF, Global
xgb, Global
43.58°N
43.56°N
43.54°N
43.52°N
Latitude
5.98°W
5.94°W
5.98°W
5.94°W
Longitude
50
40
30
20
10

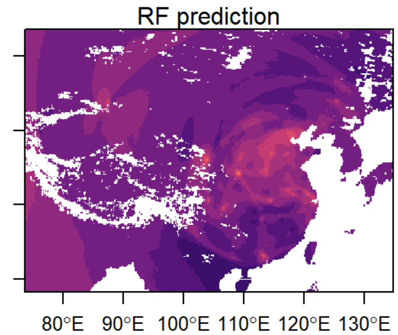
Spain


primary
highway

Using random forest for Geostatistic-like interpolation







http://rpubs.com/menglu/473973