Global air pollution modelling

University of Utrecht, The Netherlands
Global and geo-health data center

Meng Lu

Overview

- Introduction
  - Spatio-temporal epidemiology
  - Air pollution modelling and exposure assessment for health research

- Global air pollution modelling

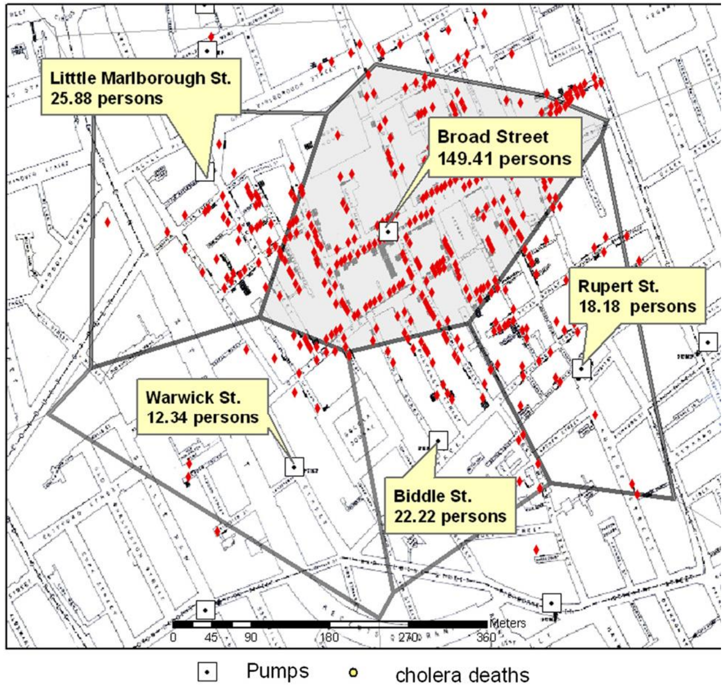
Spatio-temporal Epidemiology

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

Environmental epidemiology: Spatiotemporal epidemiology that focuses on how environmental exposures impact human health.

Origin
1854 John Snow,
Identify possible causes of cholera outbreaks.

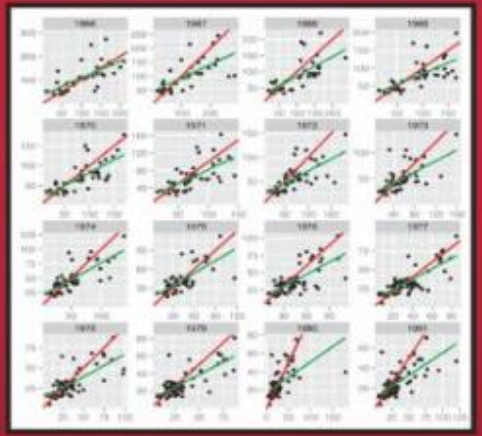
Cholera Mortality per 1,000 persons



Cholera Mortaliy per 1,00 persons for central pump catchments.

Texts in Statistical Science

Spatio-Temporal Methods in Environmental Epidemiology

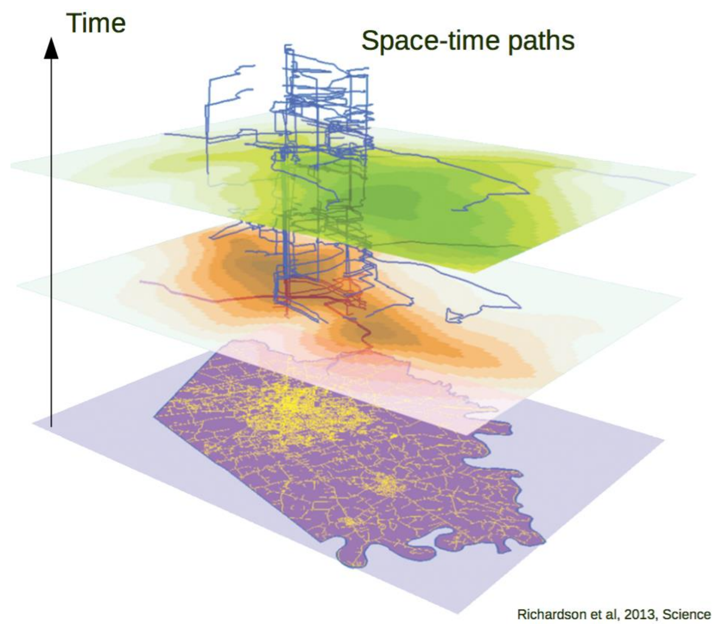


Gavin Shaddick
James V. Zidek

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

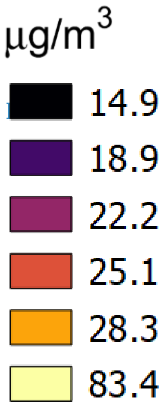
Exposome

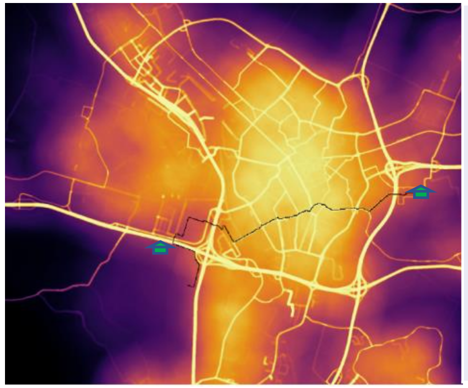
Link environment to health:



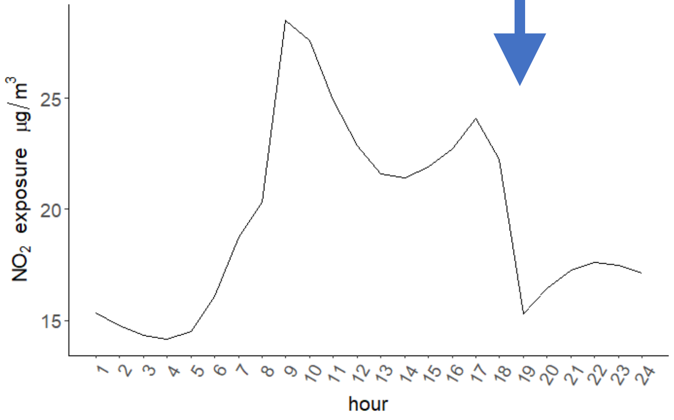
Richardson et al, 2013, Science

Agent-based modelling approach





week days	00:00 - 08:00	home
	08:00 - 09:00	commuting
	08:10 - 17:00	work
	17:00 - 18:00	commuting
	17:09 - 23:59	home



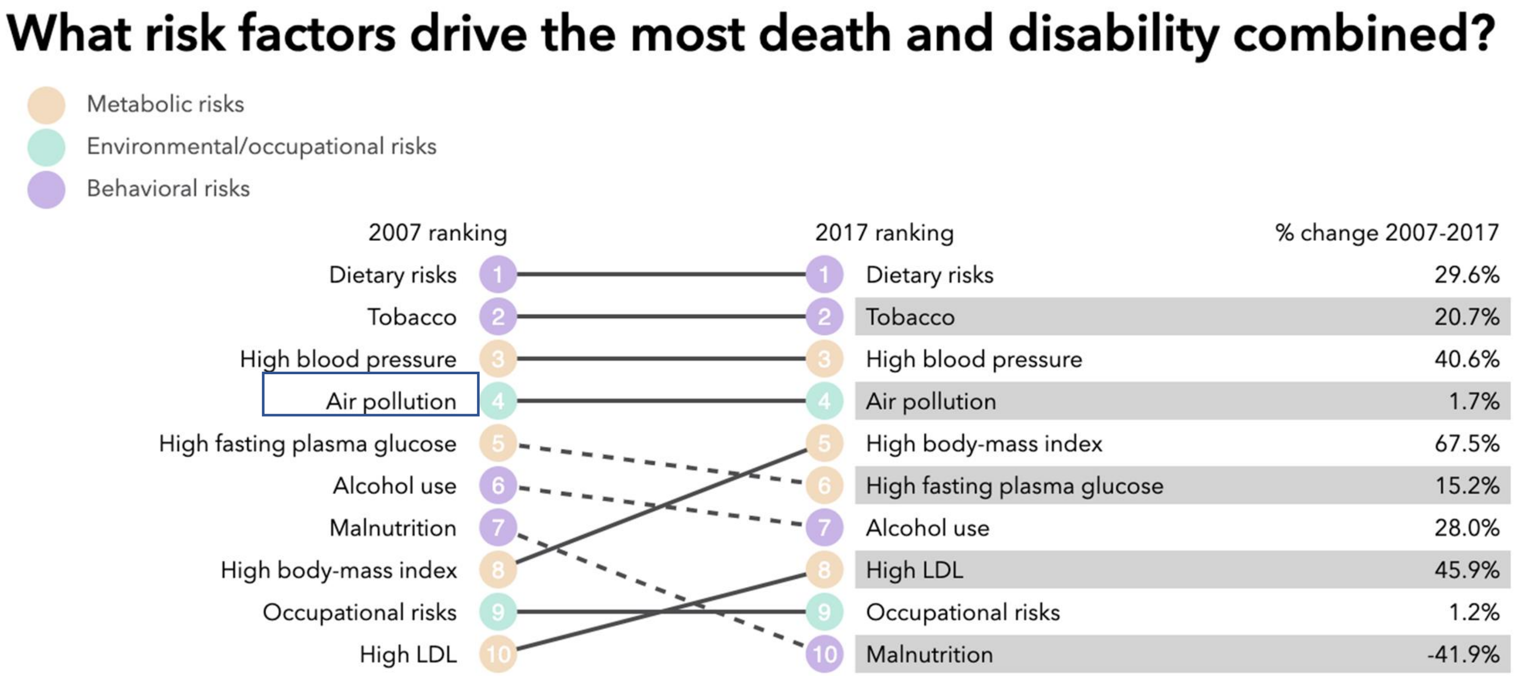
Challenge: detailed space-time paths over a large population and long time period may not be available.

work


10L PER
MINUTE

Air pollution

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



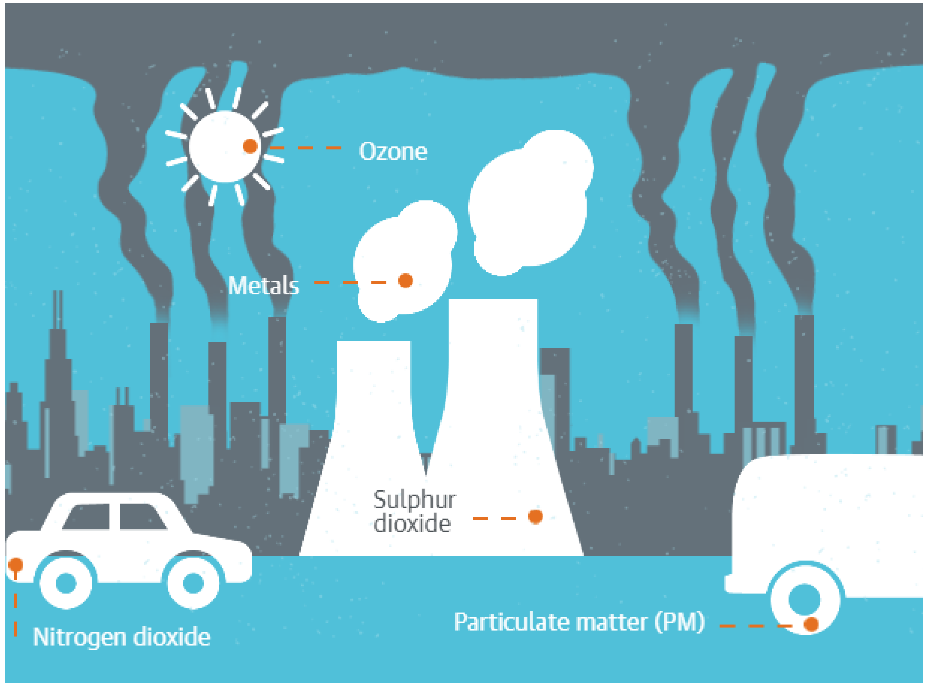
Mortality:

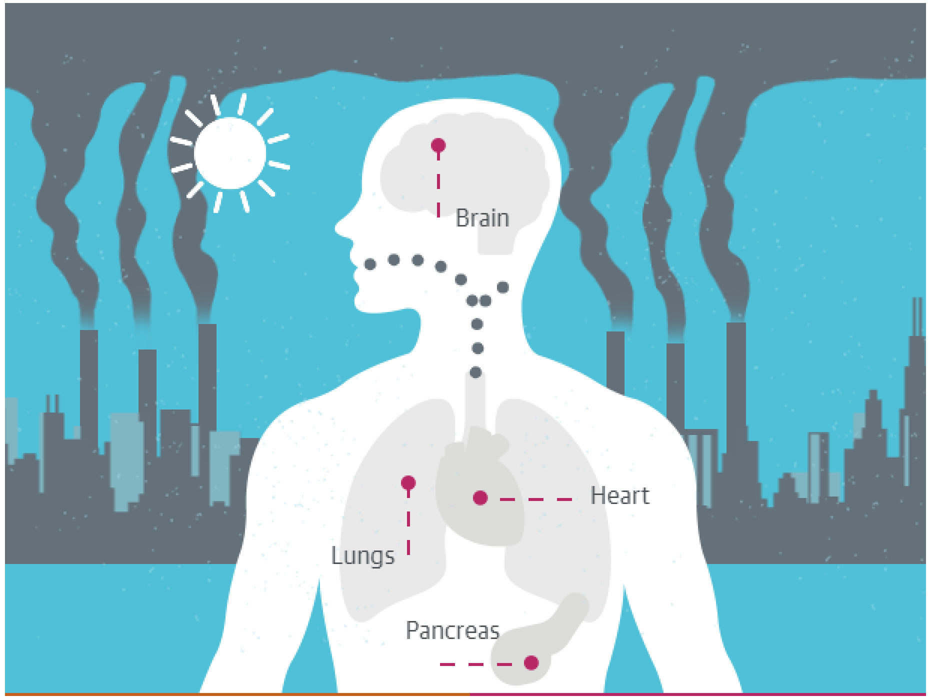
World: more than ~~3.2~~ 8.8 millions death a year

Most measured air pollutants and their health impacts

O_3
NO_2
SO_2
CO
PMx

Little is known about how air pollutant affect health over a population



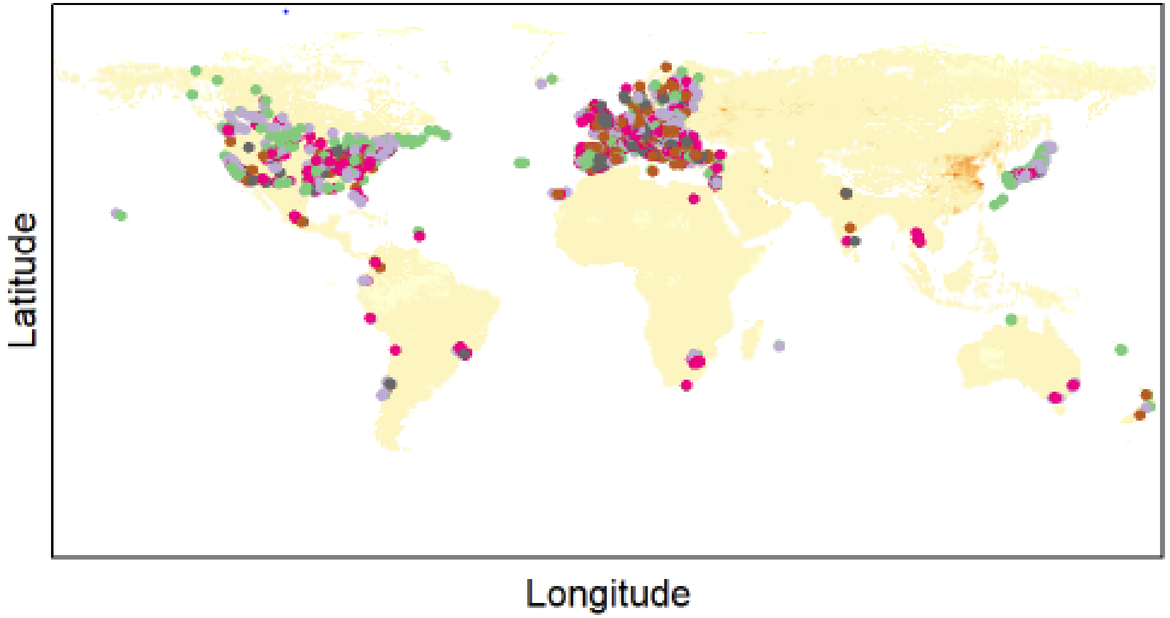


Source [2]

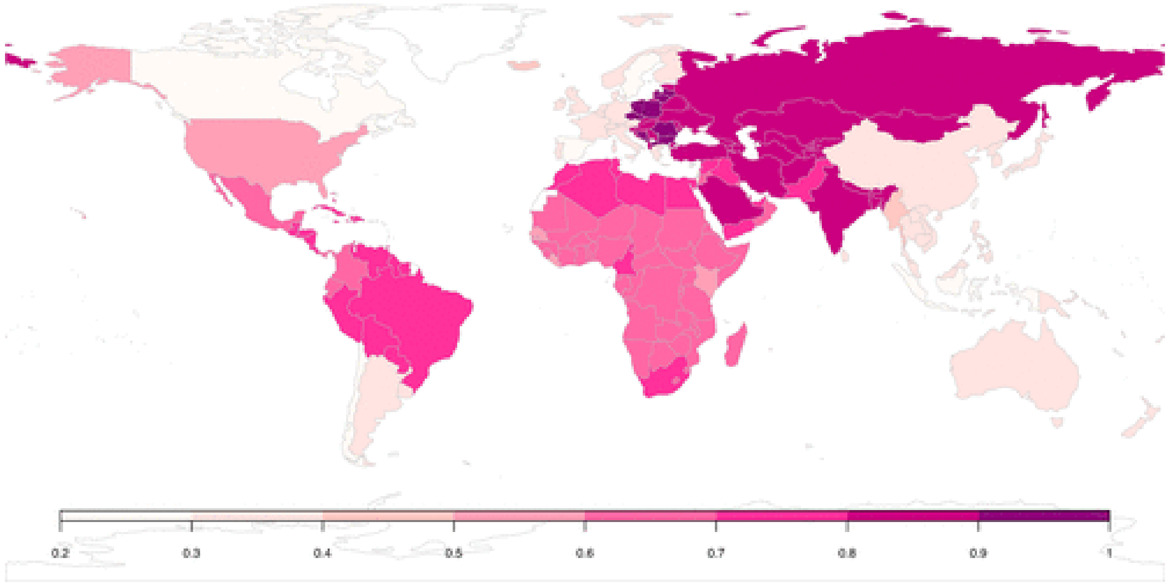
Why is global air pollutant mapping and exposure assessment important?

- Unequally distributed ground monitors
- Consistent comparison

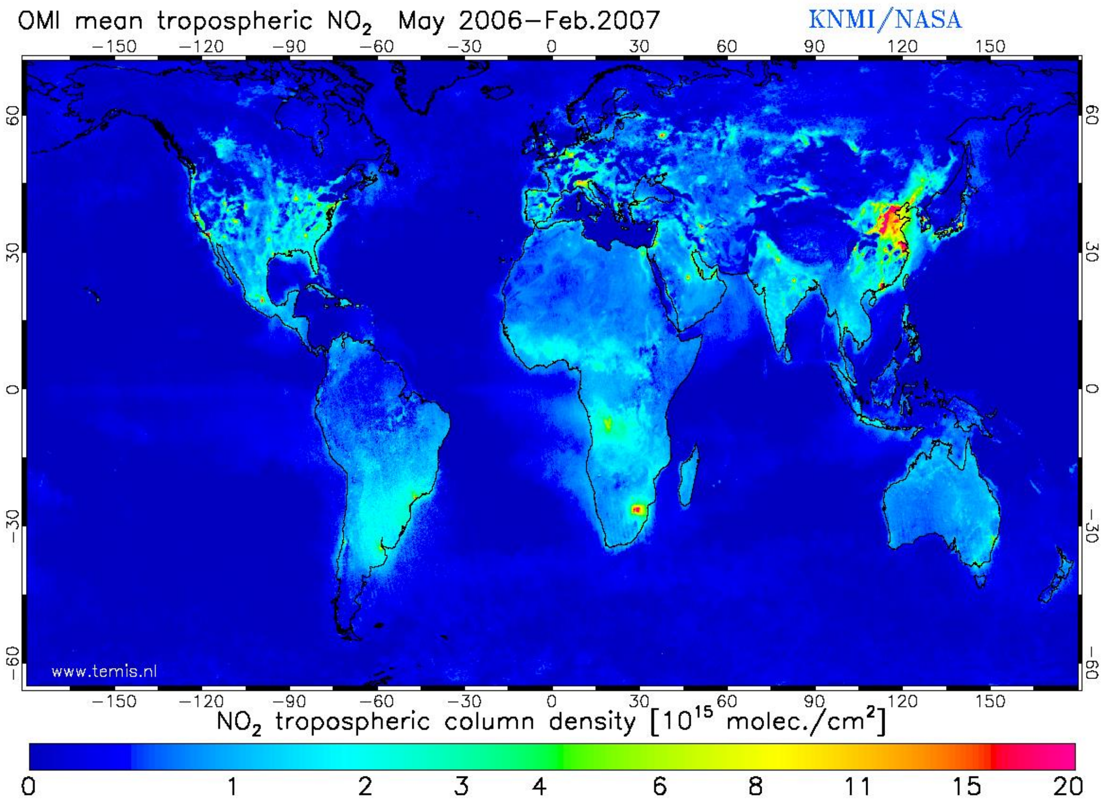
Station measurements



Shaddock et al., 2018



Remote sensing measurements:
OMI (Ozone Monitoring Instrument)



Date of Launch 15 July 2004

At nadir 13 km × 24 km

NO_2, SO_2, BrO, OClO, O_3 (36 km × 48 km)

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

Zoom in mode 13 km× 12 km

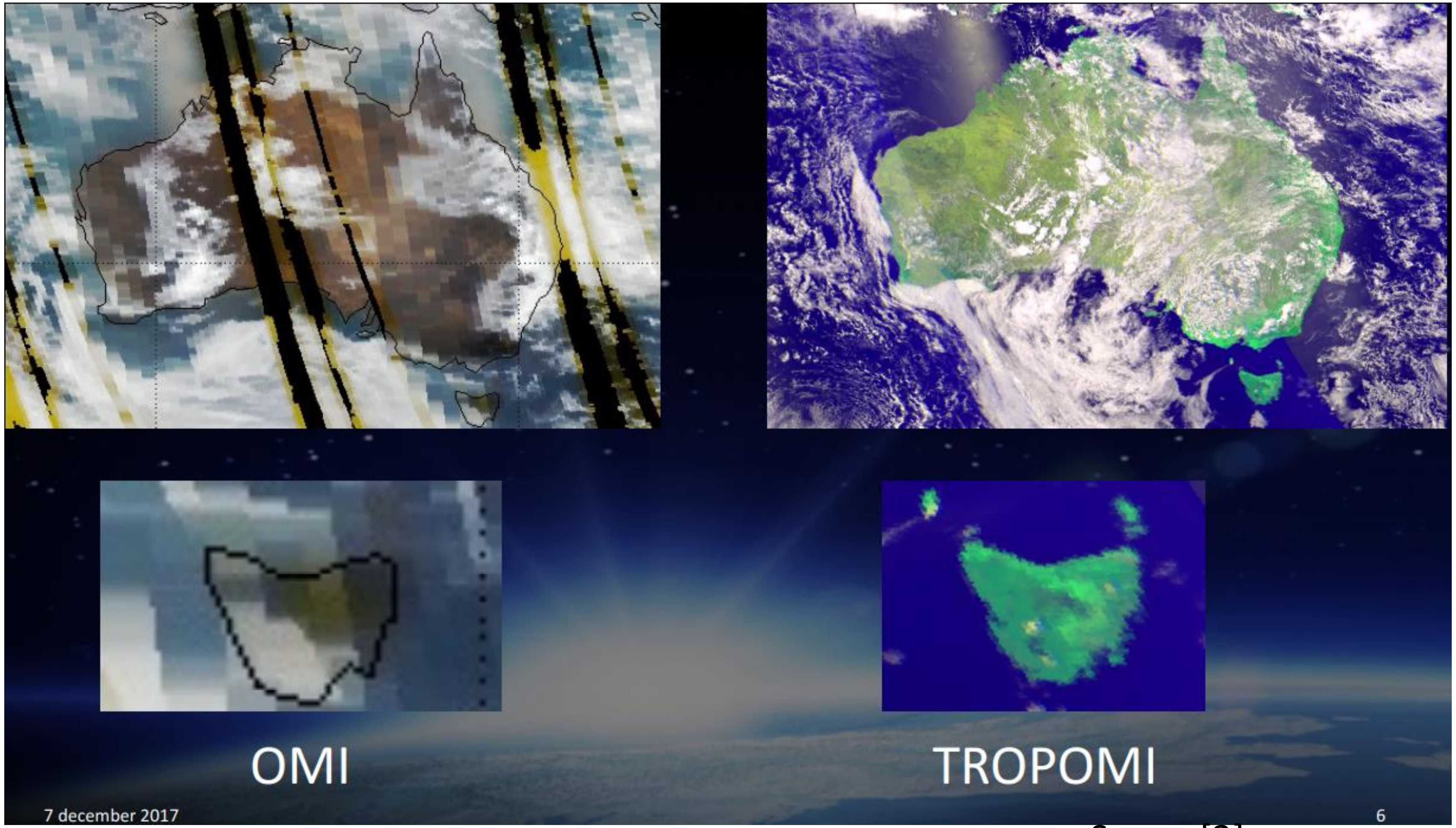
Daily global coverage

Tropomi

(TROPOspheric Monitoring Instrument)

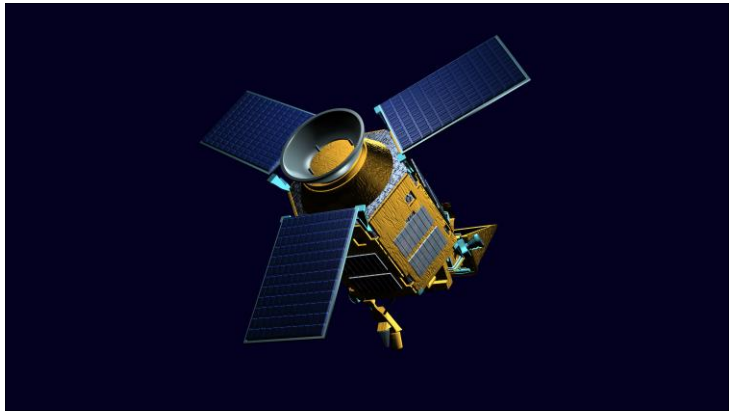
launched 2017, available from Feb 2018

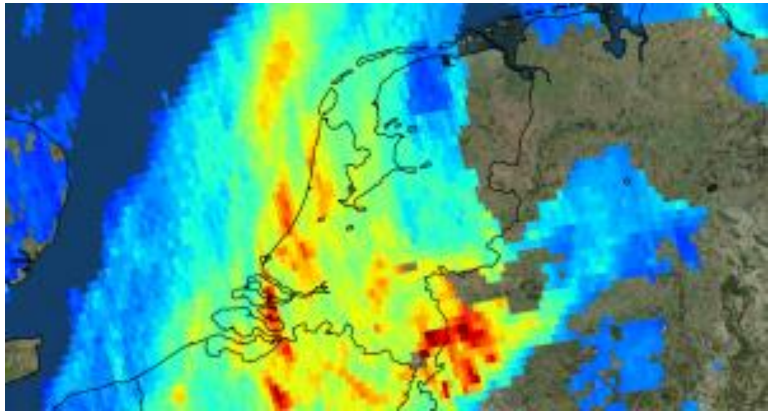
7 km x 7 km



Souce [3]

Tropomi





NO2, O3 (7km × 28km), SO2, methane and CO

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

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	

Air pollution modelling methods

- Statistical methods: regression, Kriging
- Chemical transportation models: GEOS-CHEM
- Hybrid: Kalman filter

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

Remote sensing measurements:

OMI (250 km)
Tropomi (8 km)
…

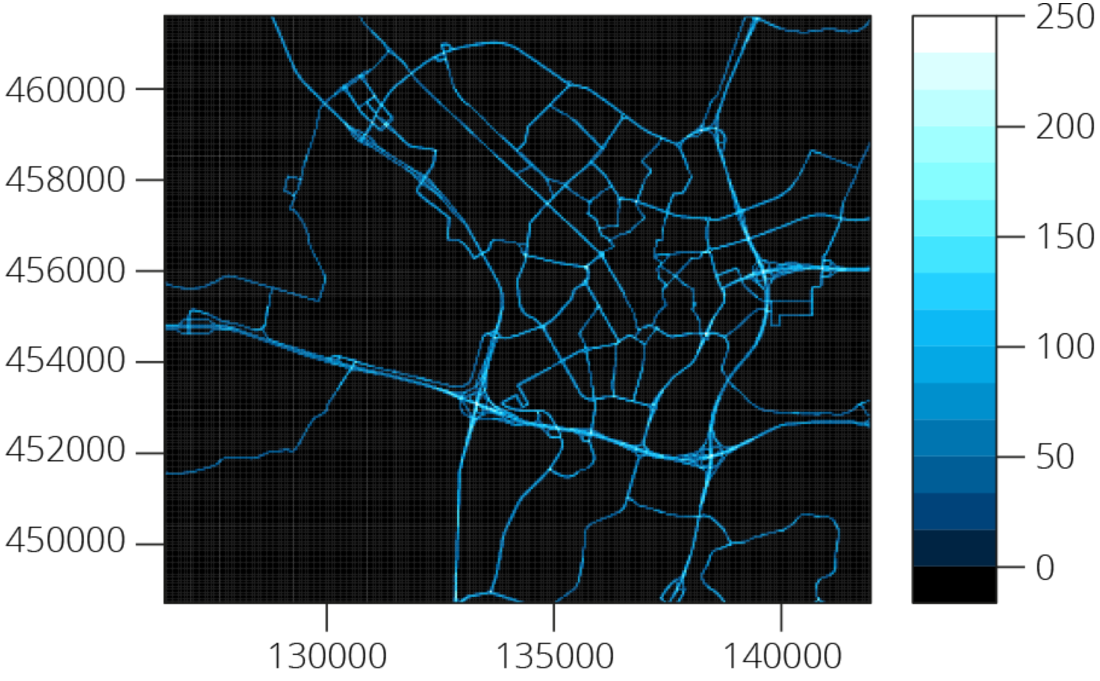
GIS predictors:

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

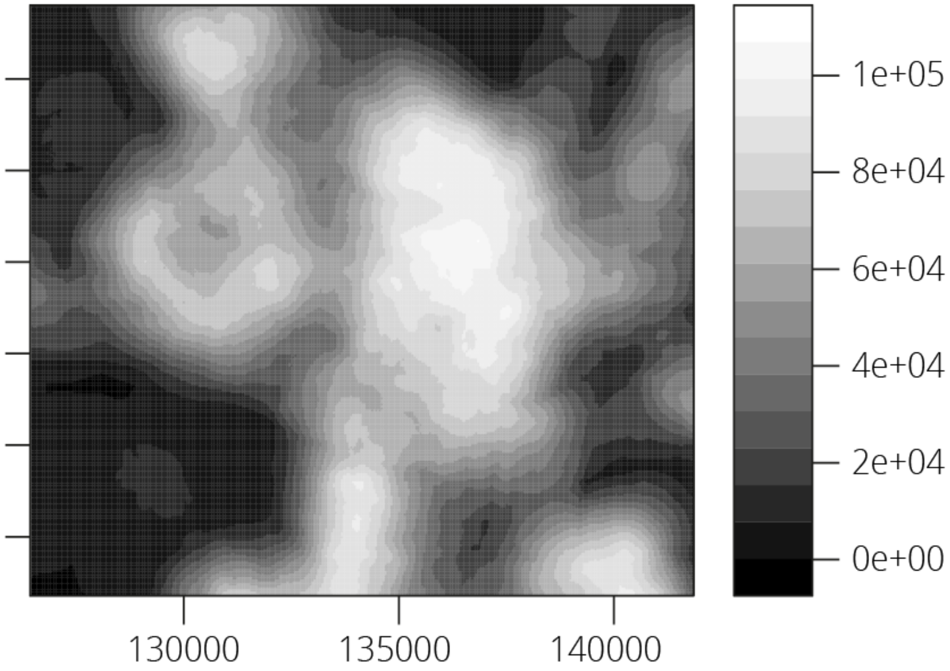
Predictor variables in buffers

Major road length

25m buffer

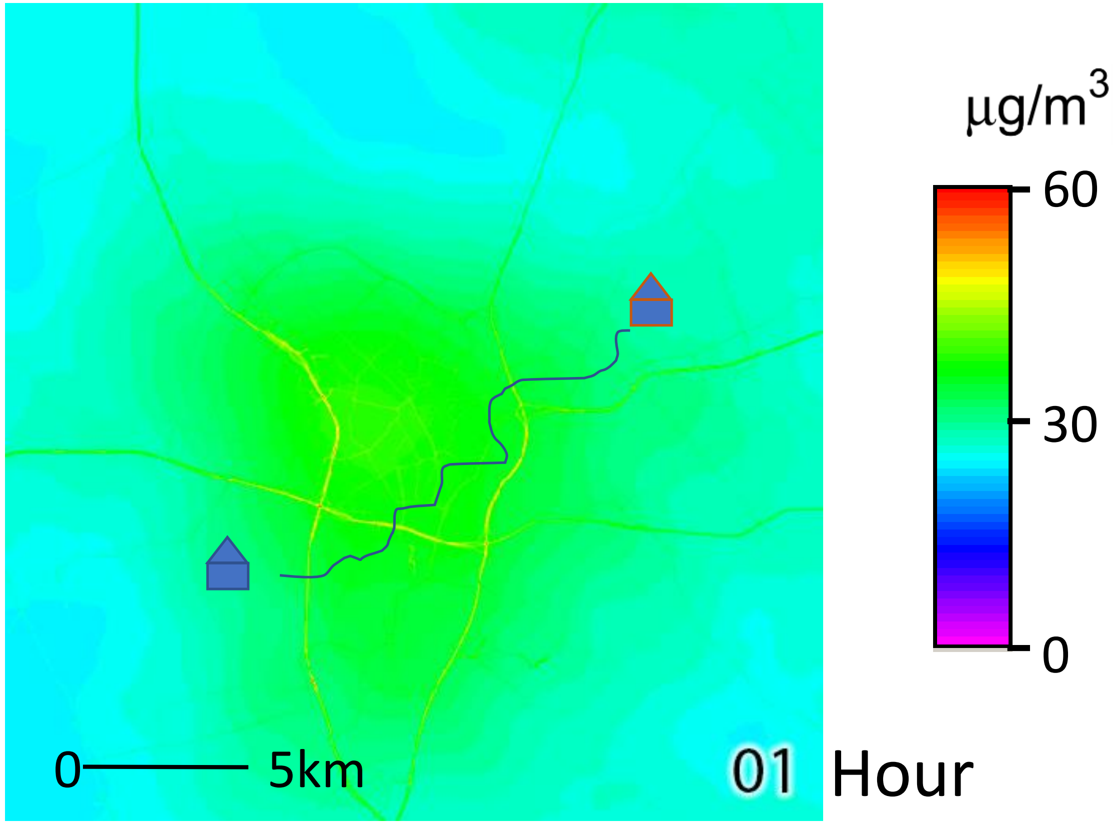


1000m buffer



LUR Prediction

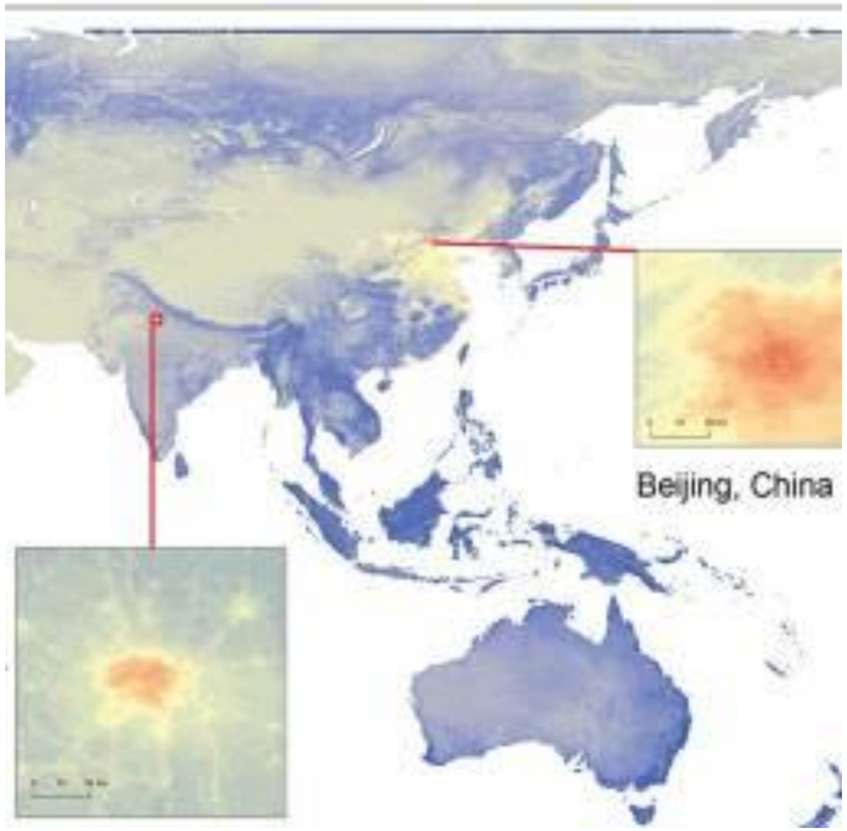
Spatiotemporal dynamic of air pollution showing road effects



5 m resolution

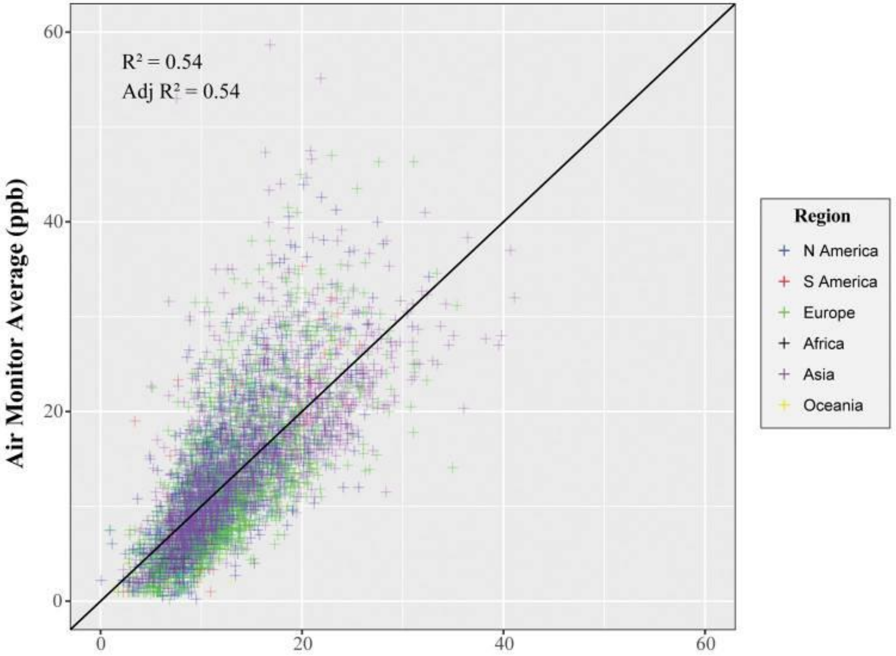
Global NO_2 mapping: Larken et al. 2017 (100m):

LUR model
Lasso, continental variable as prediction

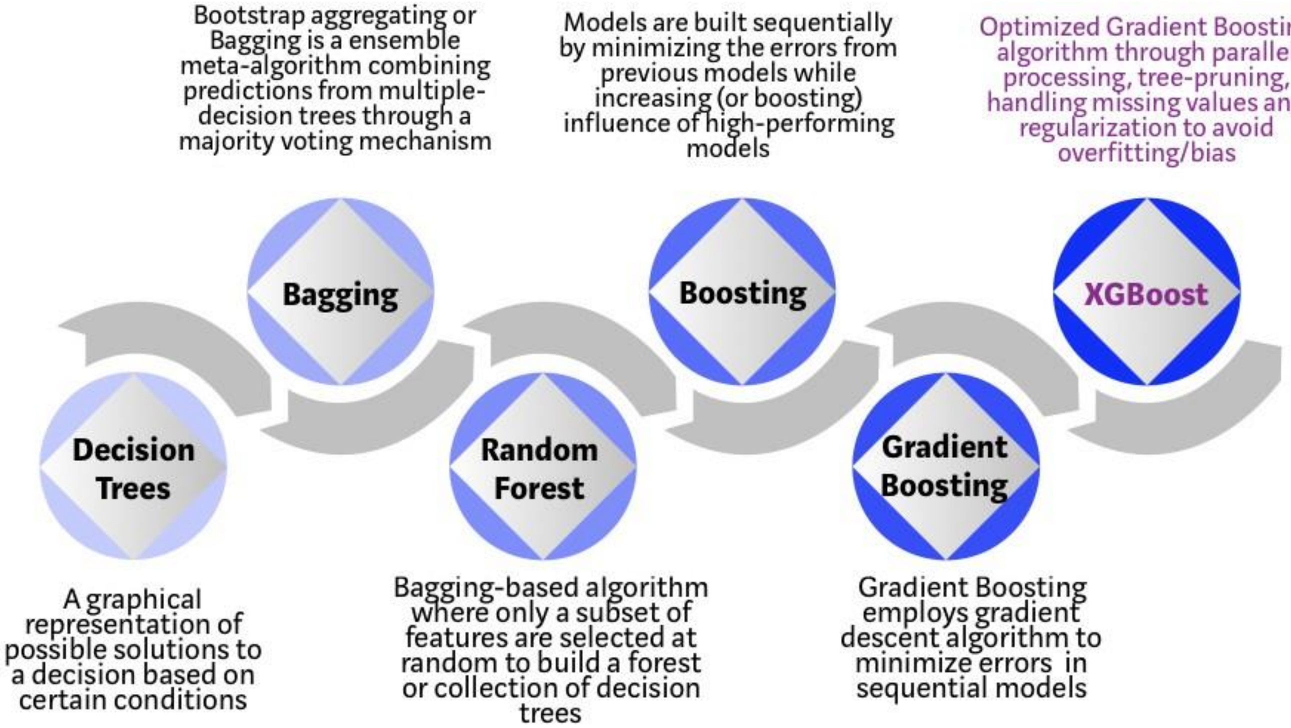


Limitations:

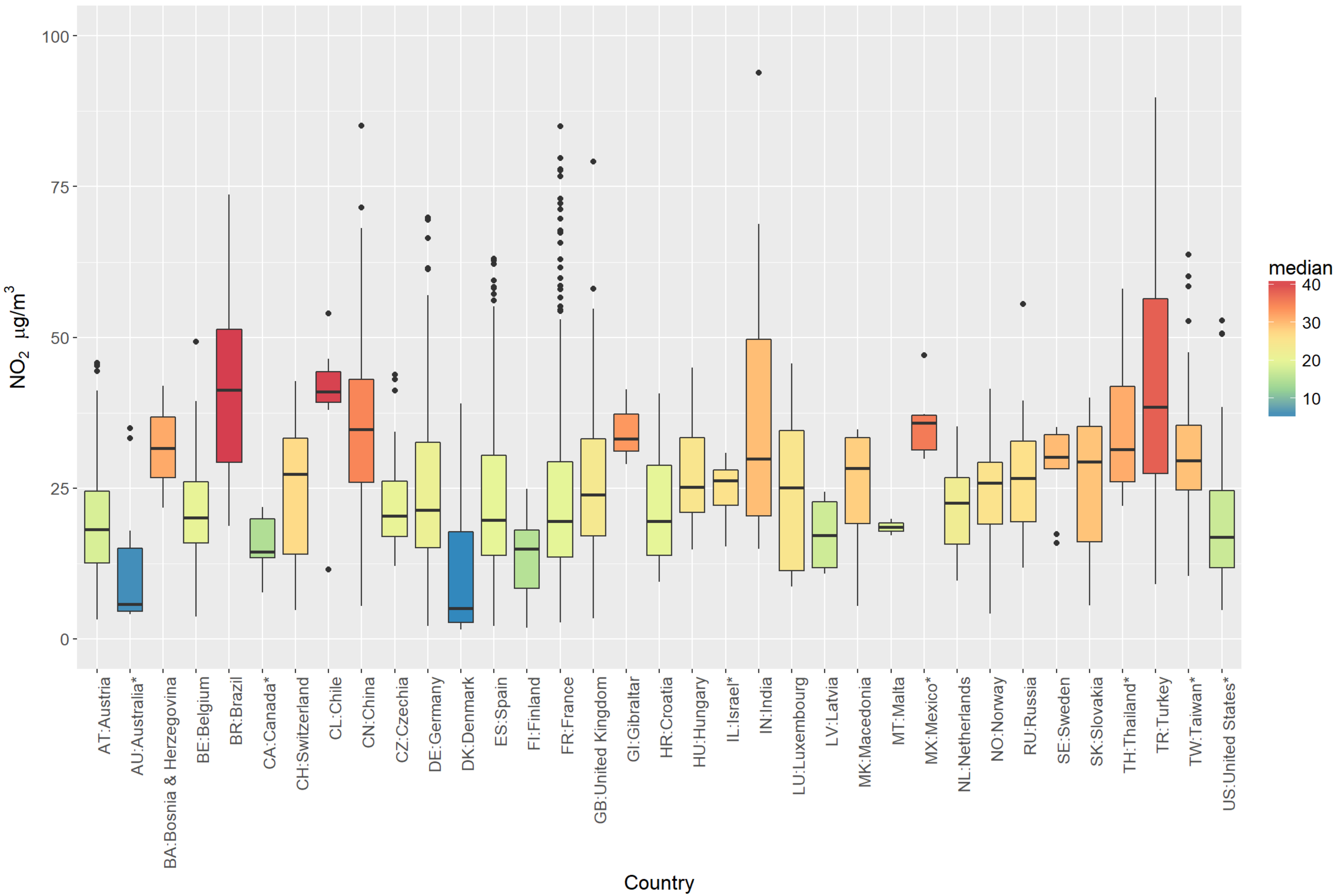
- Linear relationship
- Road effects not modelled
- Only evaluated by Rsquared and RMSE
- Does not include RS measurement



Can tree-based machine learning methods and Tropomi improve global NO_2 mapping ?



Data: OpenAQ



100
75
50
25
0
NO_2 µg/m^3
AT:Austria
AU:Australia*
BA:Bosnia & Herzegovina
BE:Belgium
BR:Brazil
CA:Canada*
CH:Switzerland
CL:Chile
CN:China
CZ:Czechia
DE:Germany
DK:Denmark
ES:Spain
FI:Finland
FR:France
GB:United Kingdom
GI:Gibraltar
HR:Croatia
HU:Hungary
IL:Israel*
IN:India
LU:Luxembourg
LV:Latvia
MK:Macedonia
MT:Malta
MX:Mexico*
NL:Netherlands
NO:Norway
RU:Russia
SE:Sweden
SK:Slovakia
TH:Thailand*
TR:Turkey
TW:Taiwan*
US:United States*
Country
median
40
30
20
10

Predictors

Emission-related
Road length within 25 m – 300 m radius ring
- Highway, primary roads, secondary roads, tertiary roads, unpaved roads

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 from Satellite products and the GEOS-CHEM
Satellite measured NO_2 column density
Distance to coast

Method

Comparing different statistical learning methods

Trees-based

- Random forest
- Stochastic gradient boosting
- Extreme gradient bossting

Regularized regression

- Ridge
- Lasso
- ElasticNet

Mechamical model

- Nonlinear regression integrating air distribution mechanisms

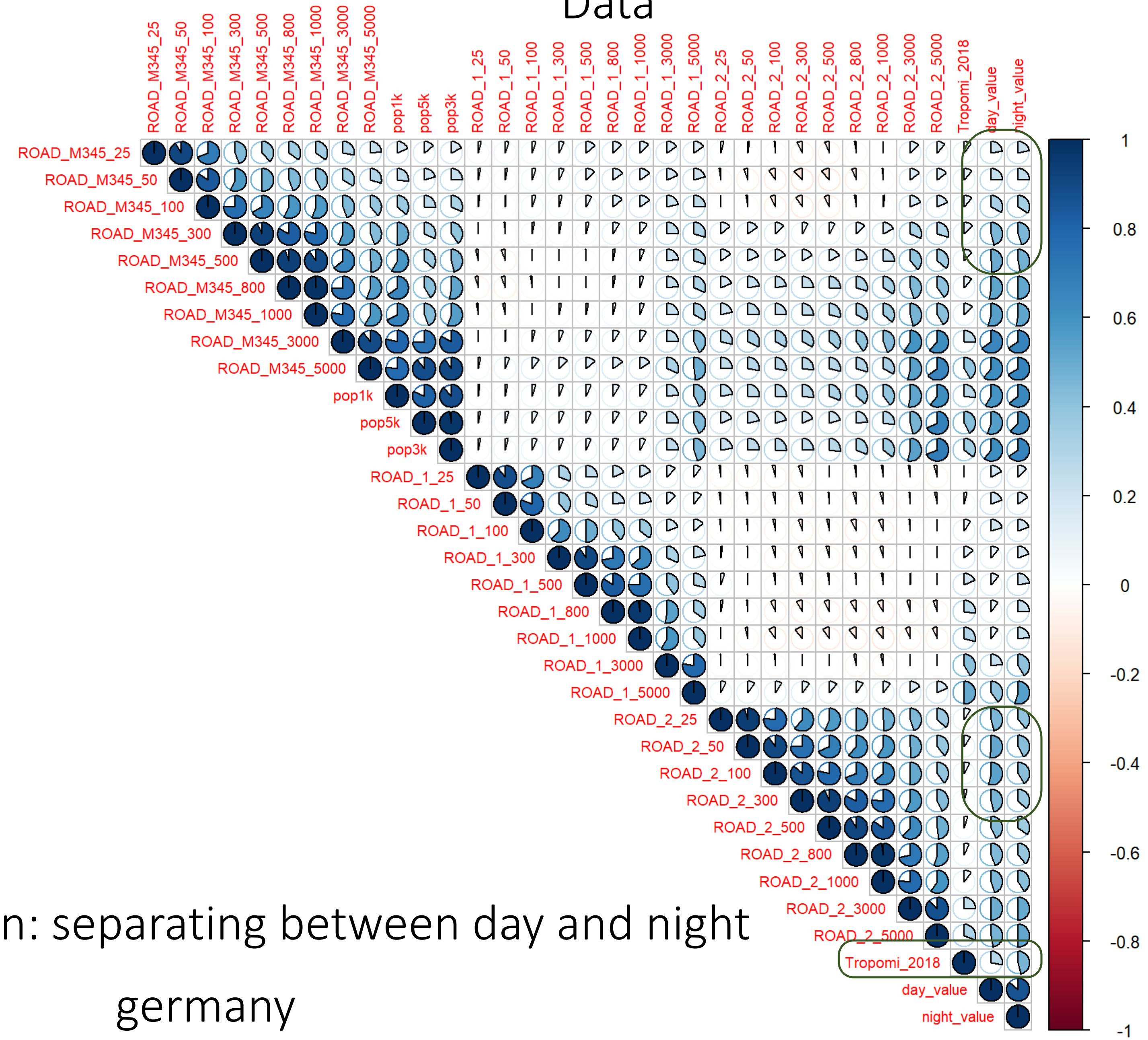
Compare global and national models

Four national models:

US (100),
China (1400),
Germany (350),
Spain (350)

A global model

Data



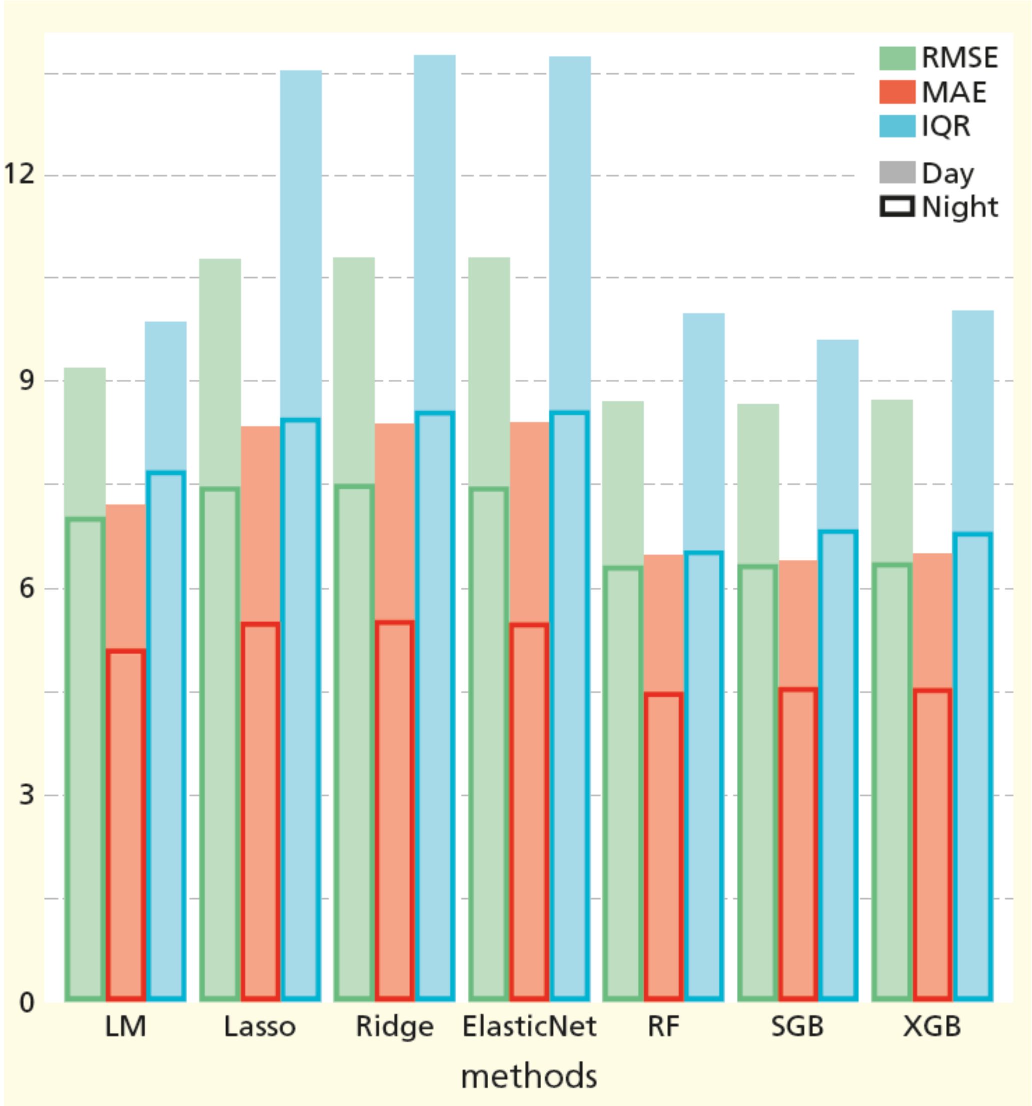
Local roads

Primary roads

Tropomi

Paired correlation: separating between day and night germany

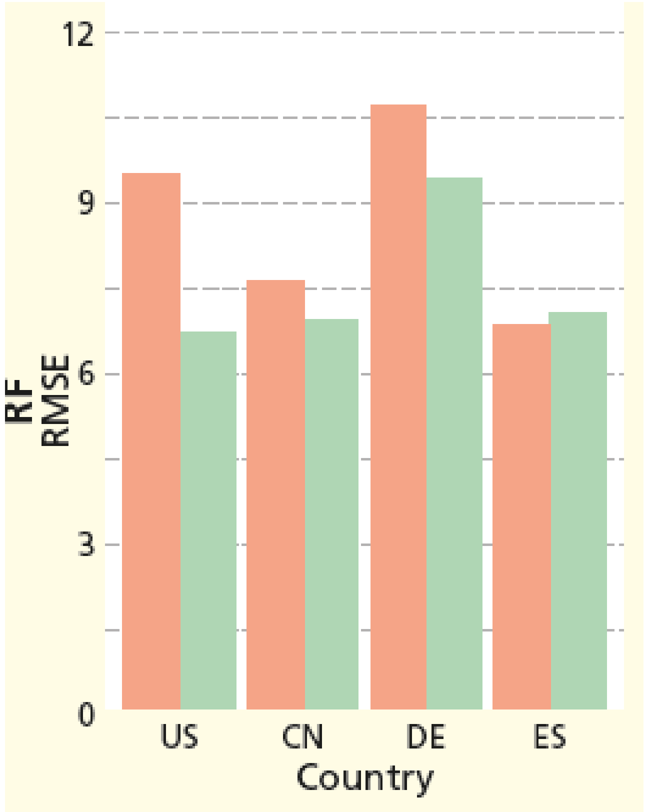
Result: global model accuracy

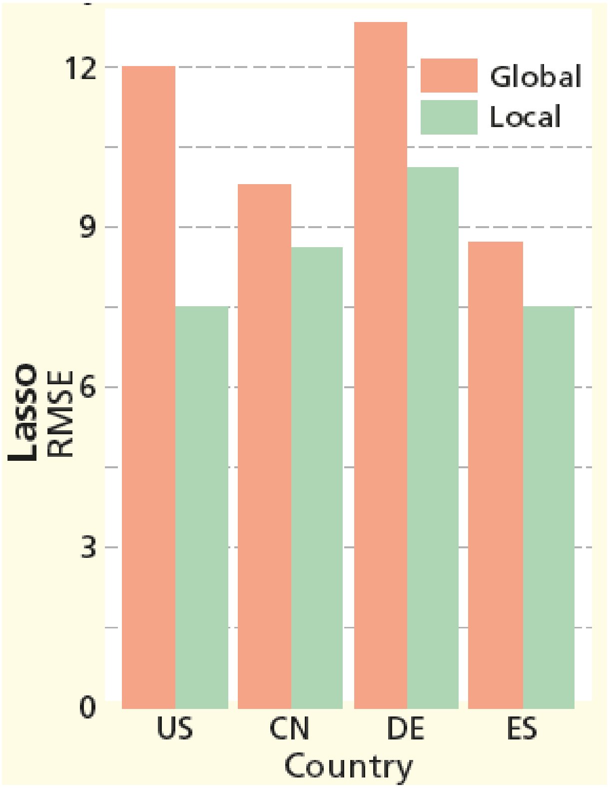


RMSE: root mean squared error
MAE: mean absolute error
IQR: interquartile range

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

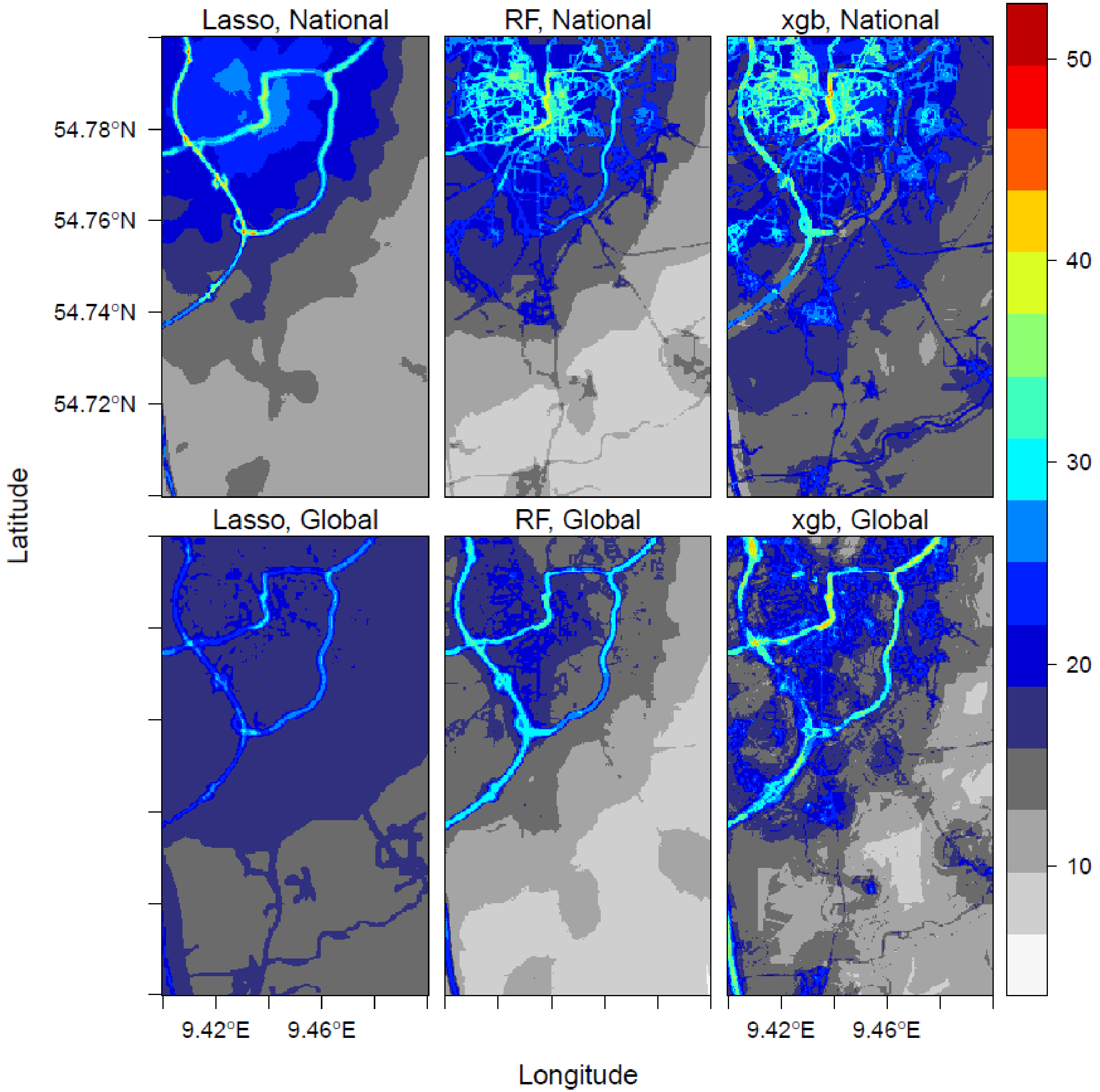
Result: global and national models
RF vs. Lasso





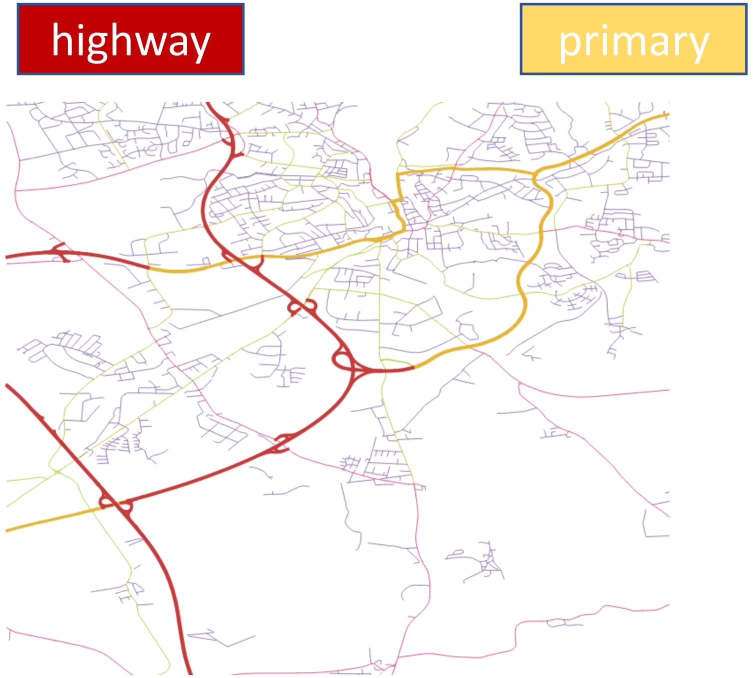
CN: China
DE: Germany
ES: Spain

Conclusion: random forest is more suitable than Lasso for a global NO2 maping and using random forest can achieve an accuracy as good as national models.



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

Germany

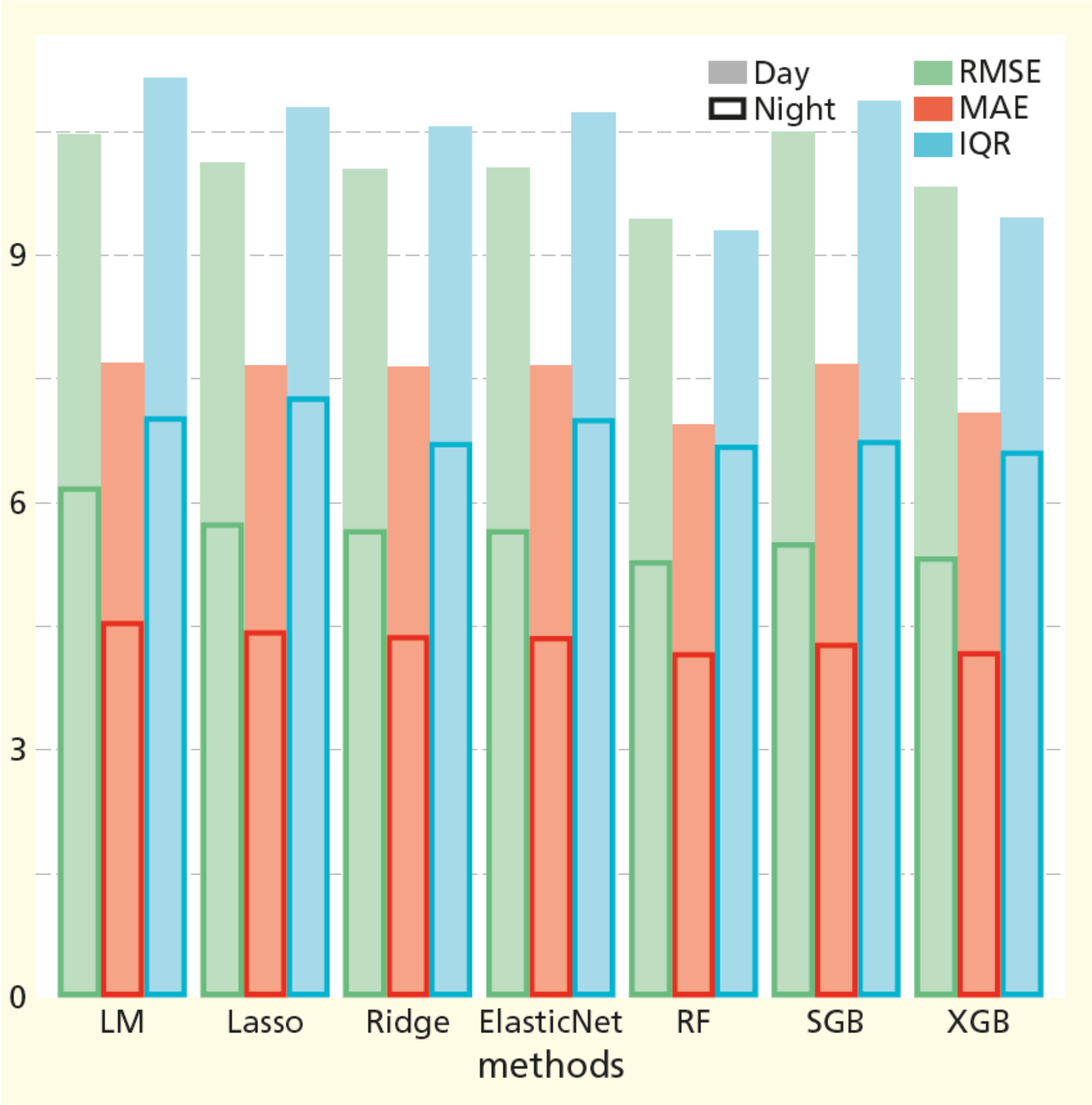


highway
primary

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

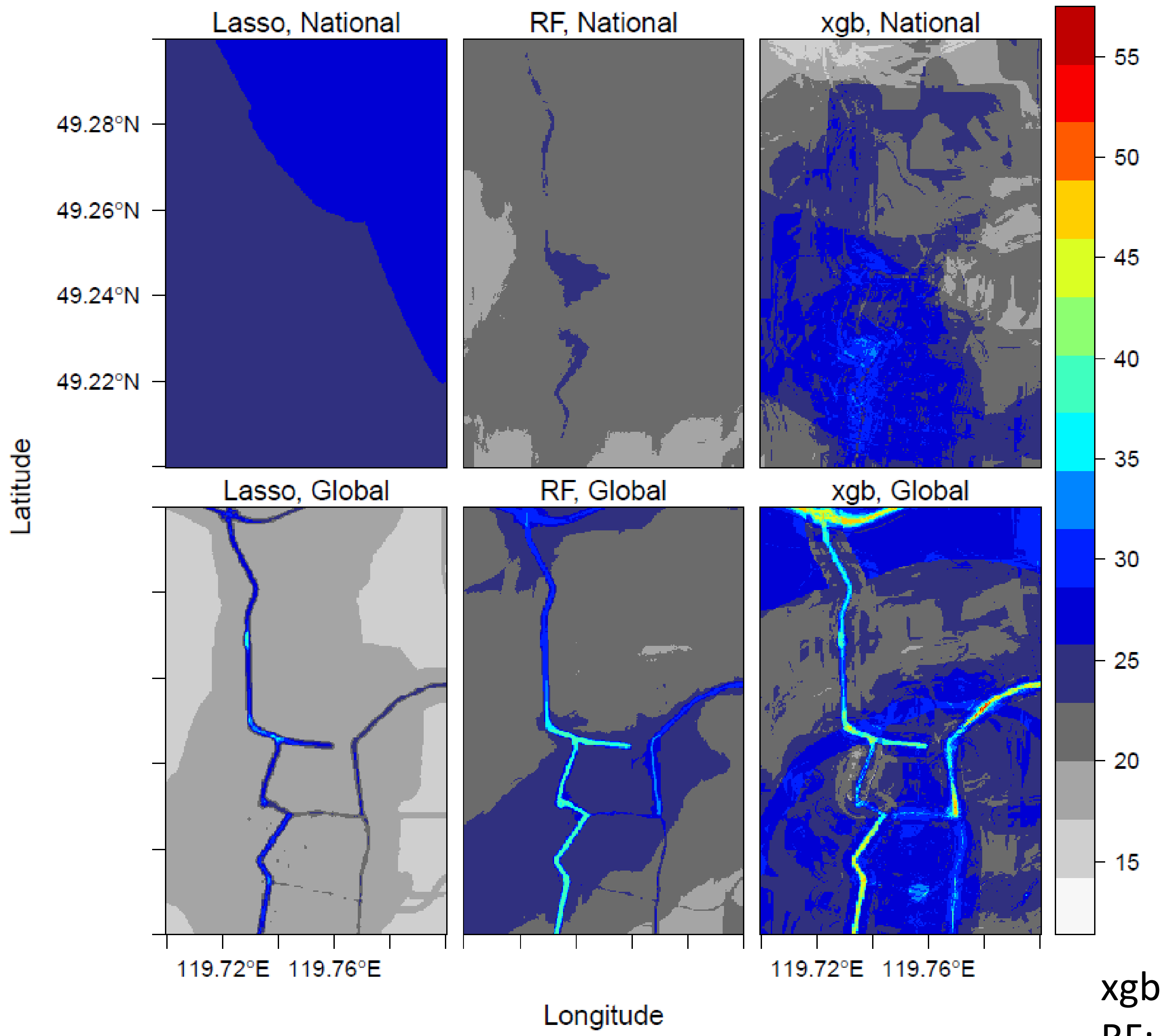
Germany



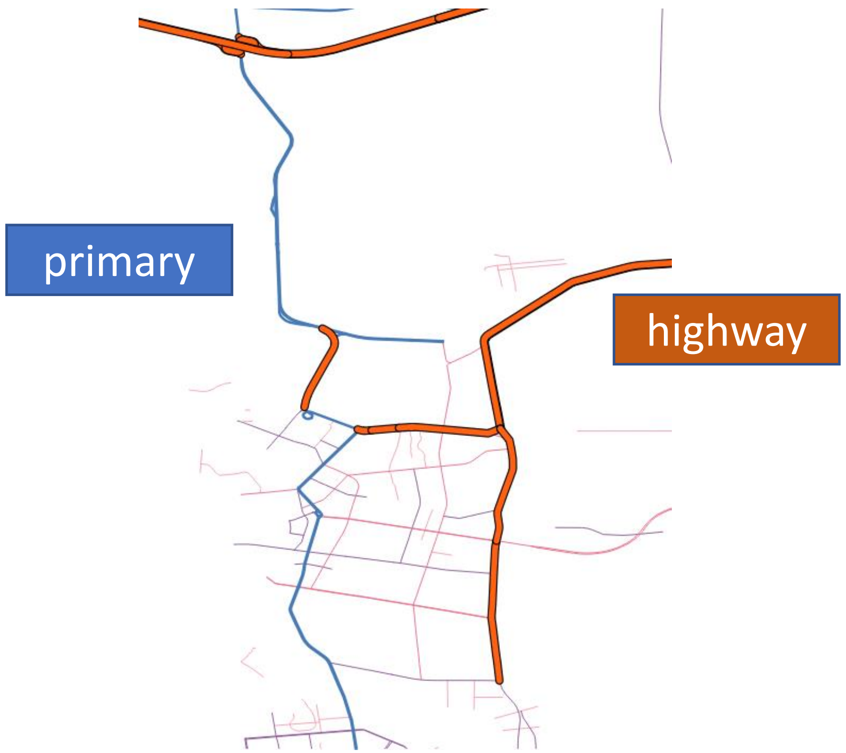
RMSE: root mean squared error
MAE: mean absolute error
IQR: interquartile range

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

Prediction from different methods

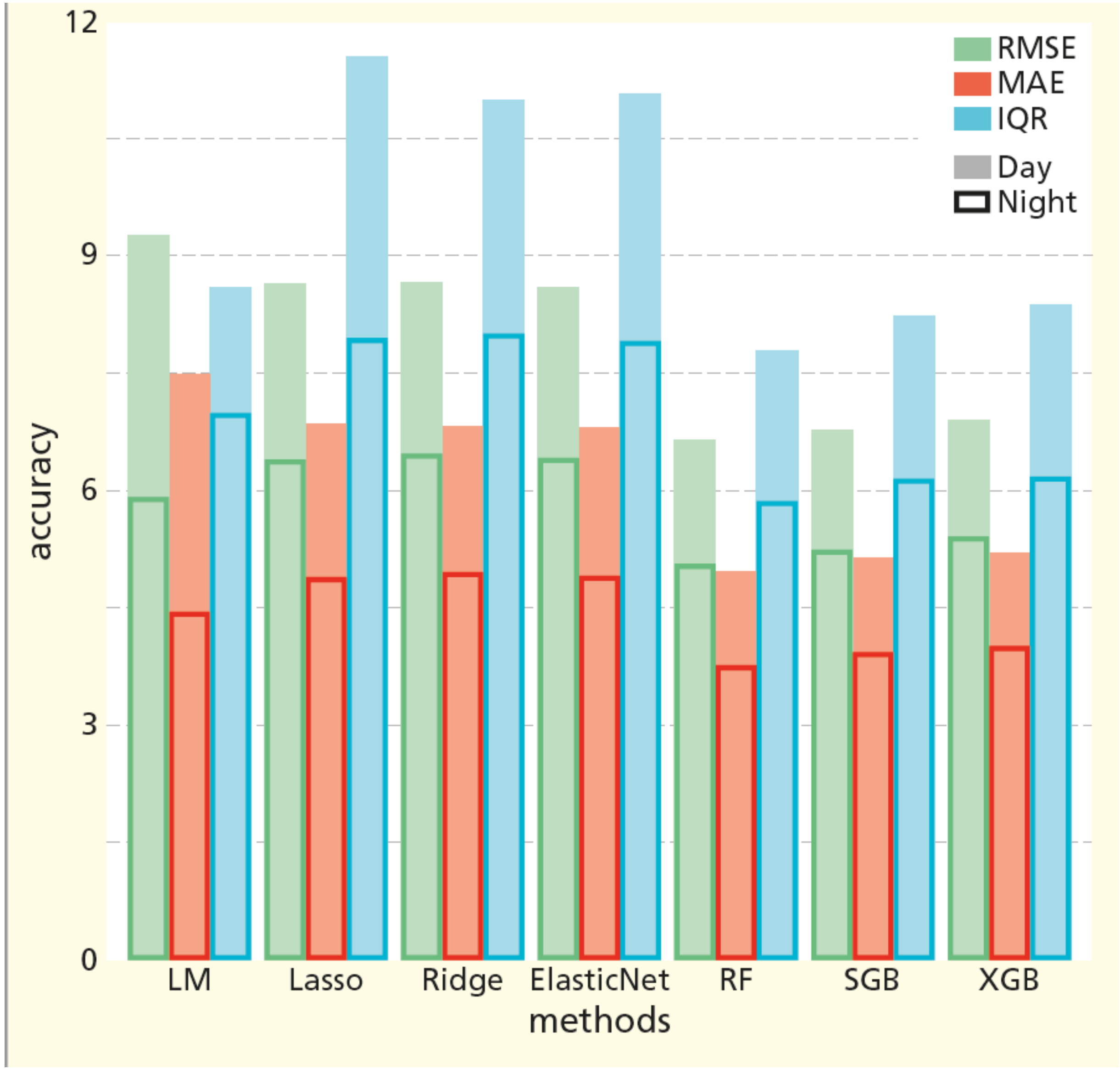


emission-related variables



xgb: xgboost
RF: random forest

China



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

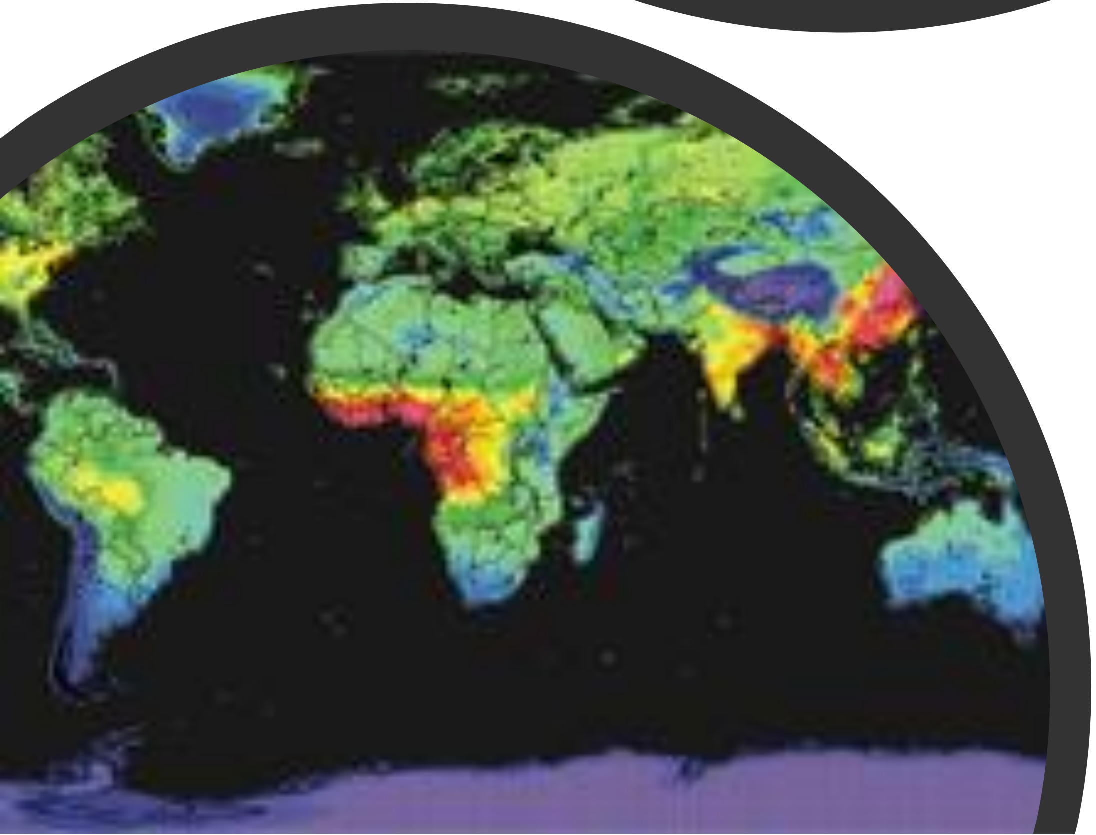
Conclusion

- The validation results indicate that tree-based methods are more suitable than Lasso for a global NO2 maping and their global models can achieve an accuracy as good as national models.

- The differences in validataion accuracy between statistical learning methods are small.

- The patterns of spatial predictions using different methods are notably different.

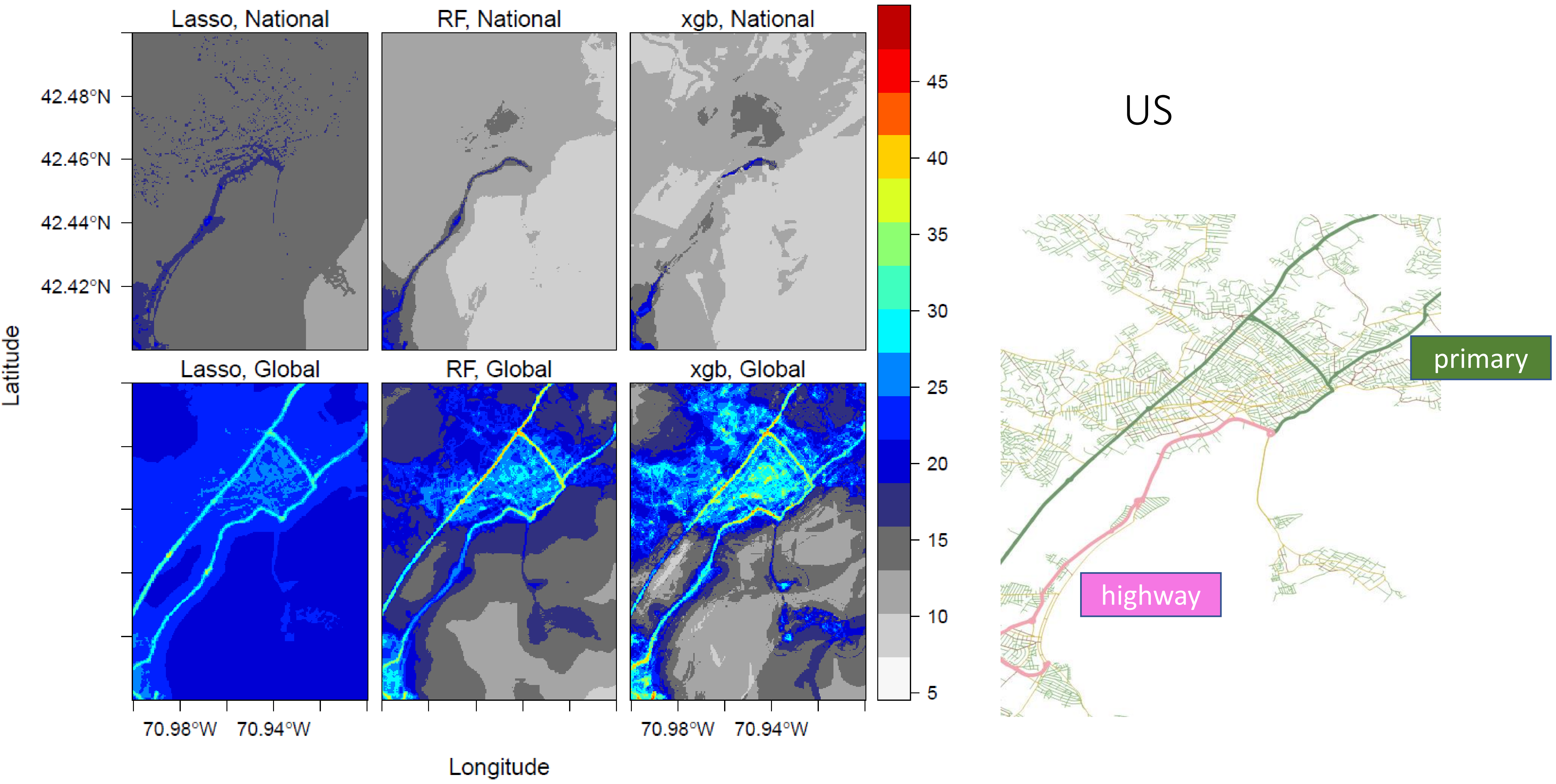
- Field tracking measurements may be needed for validation.



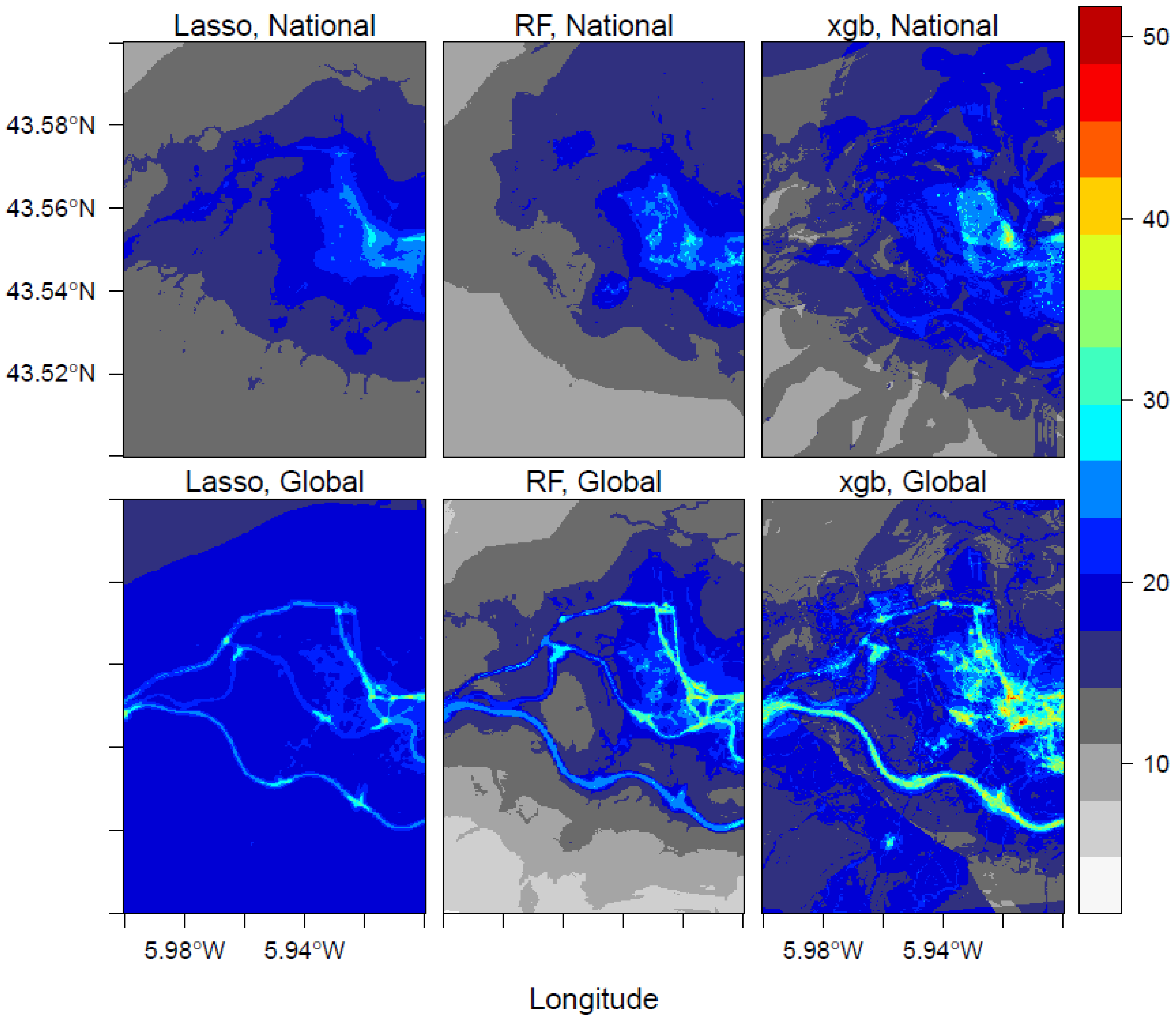
Thank you

[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

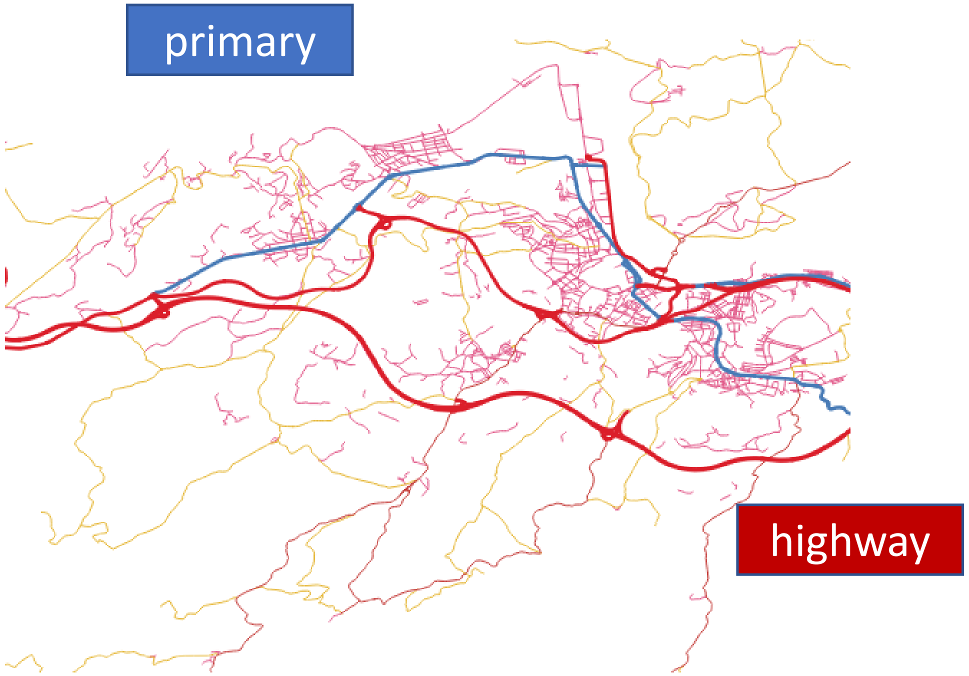
US



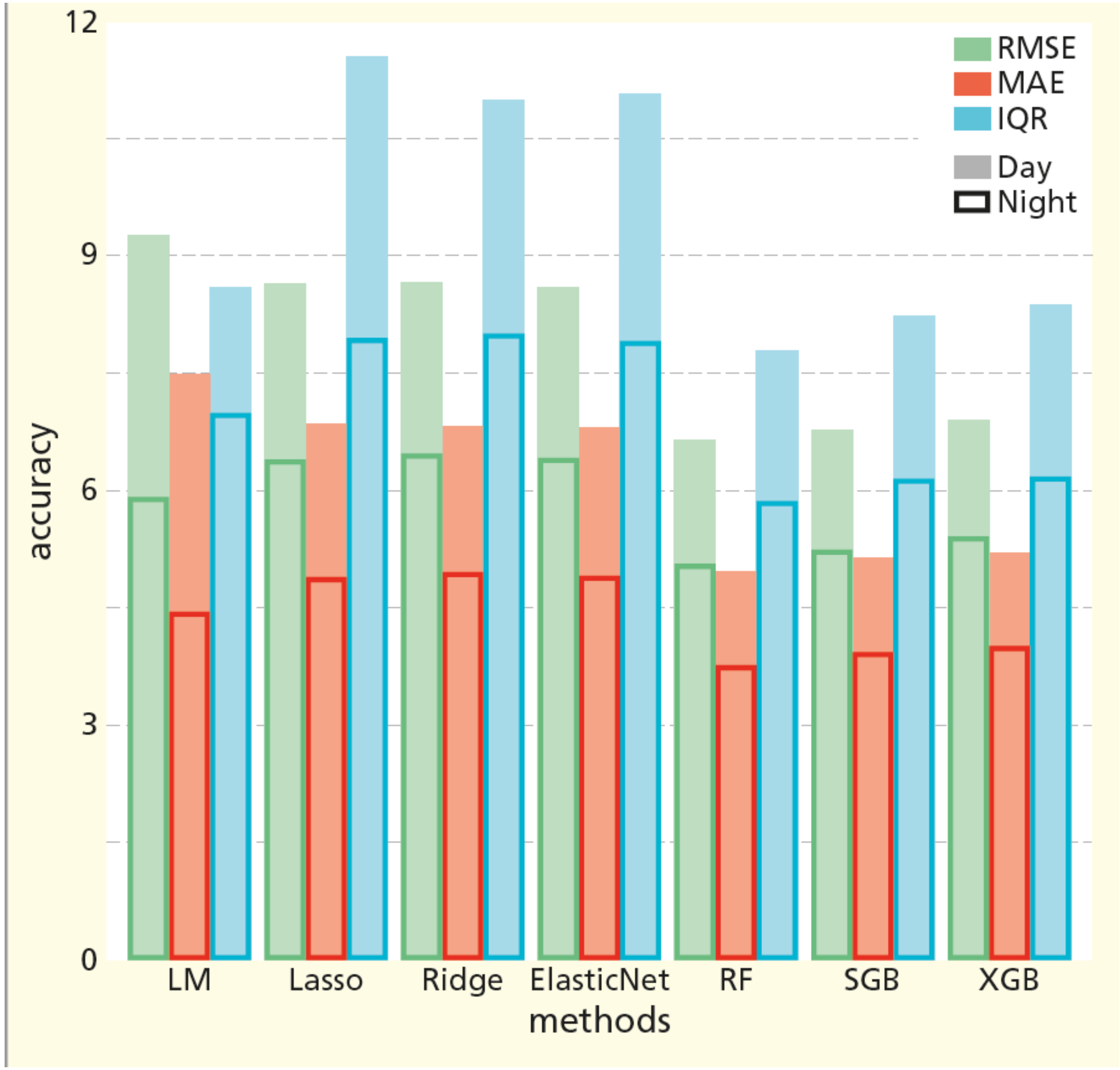
Lasso, National
RF, National
xgb, National
Lasso, Global
RF, Global
xgb, Global
42.48°N
42.46°N
42.44°N
42.42°N
Latitude
70.98°W
70.94°W
Longitude
45
40
35
30
25
20
15
10
5
primary
highway



Prediction: Spain



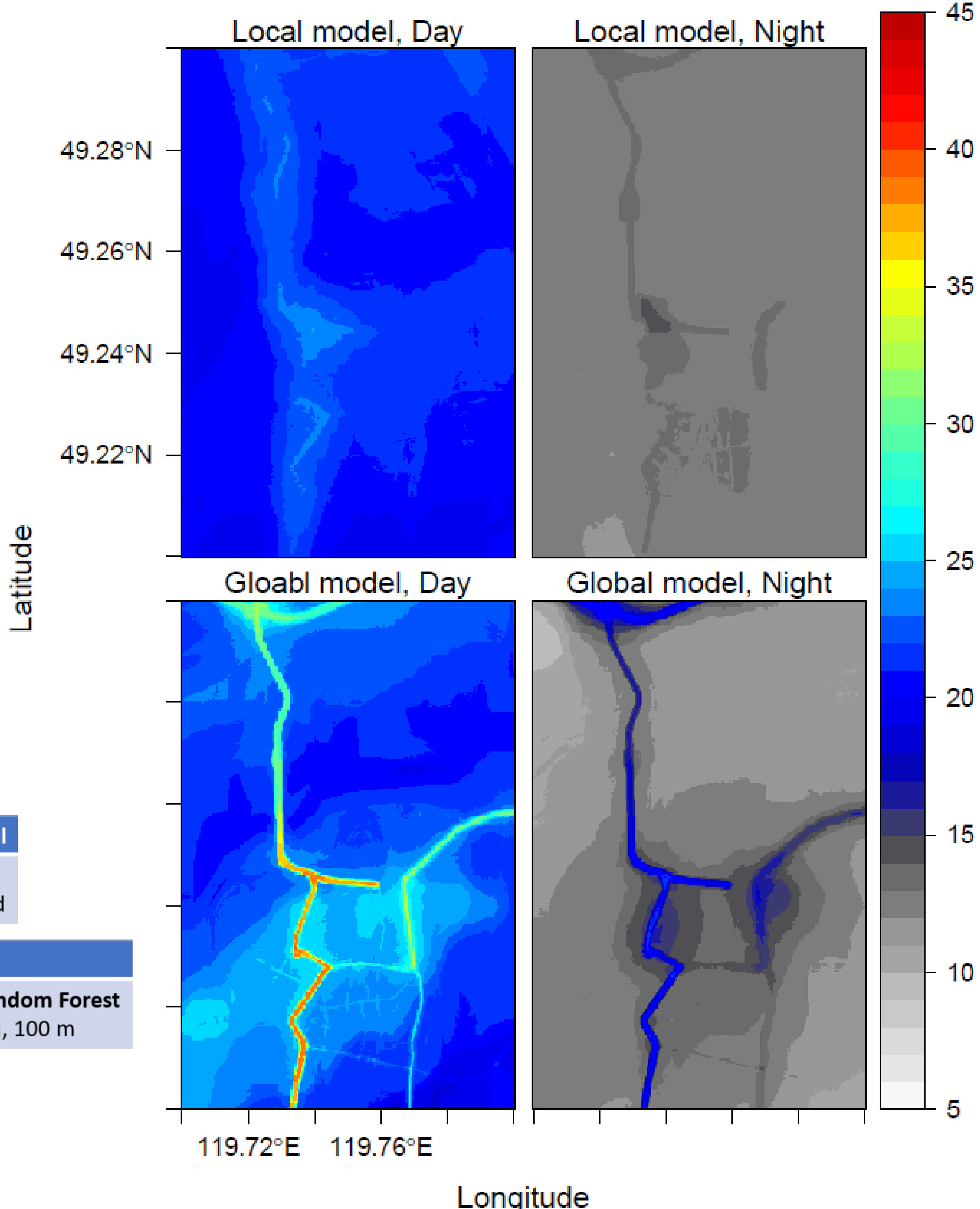
China



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

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