Air pollution in Berlin

Using machine learning algorithms to model atmospheric aerosol particles in Berlin

Source: IMAGO / phototek

Air pollution

#3 death cause globally

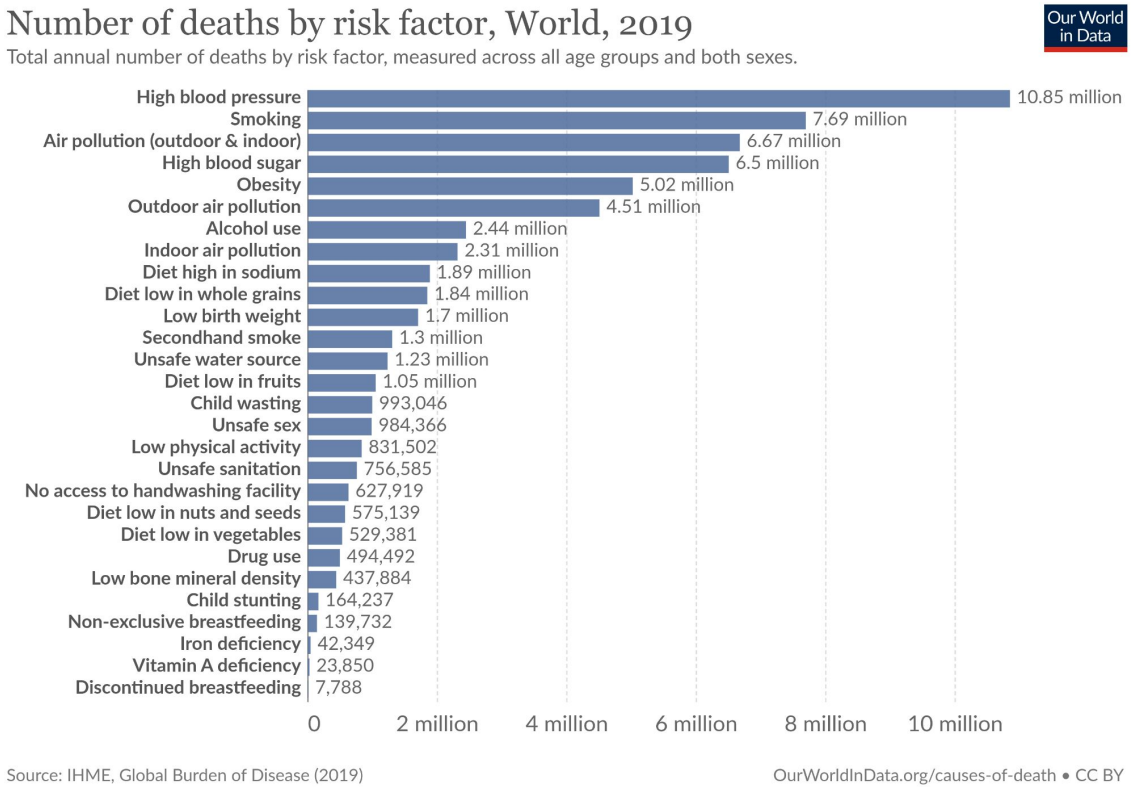
01 | stroke

02 | heart disease

03 | lung cancer

04 | respiratory diseases (asthma)

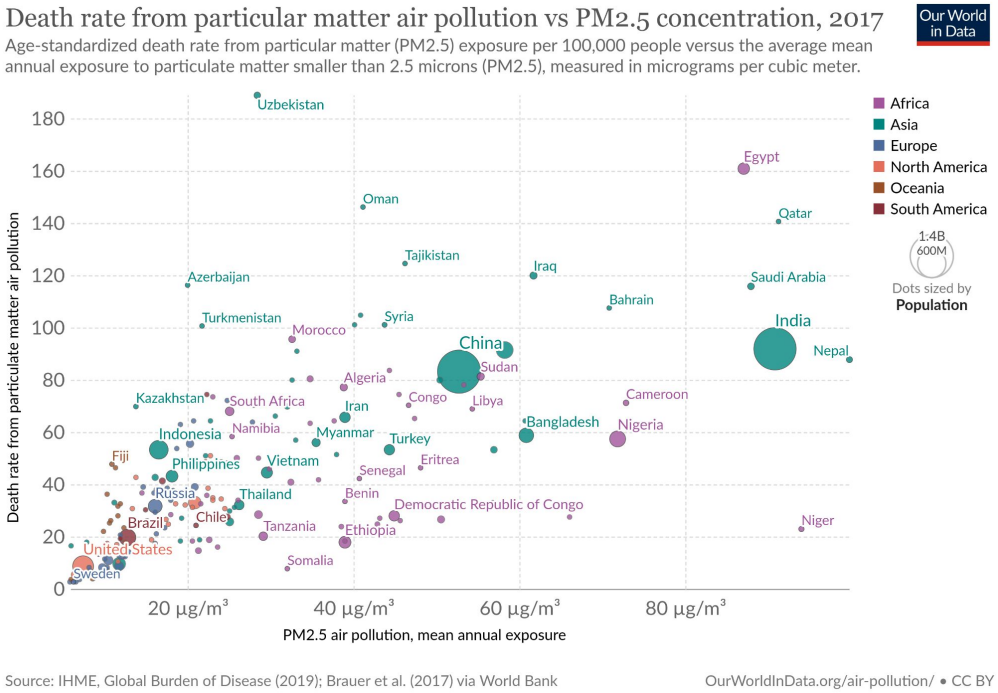
05 | infertility



PM 2.5 concentration

Particulate matter

- Atmospheric aerosol particles with diameter <2.5 μm
- Microscopic particles suspended in the air
- Respirable (penetrate into lungs)



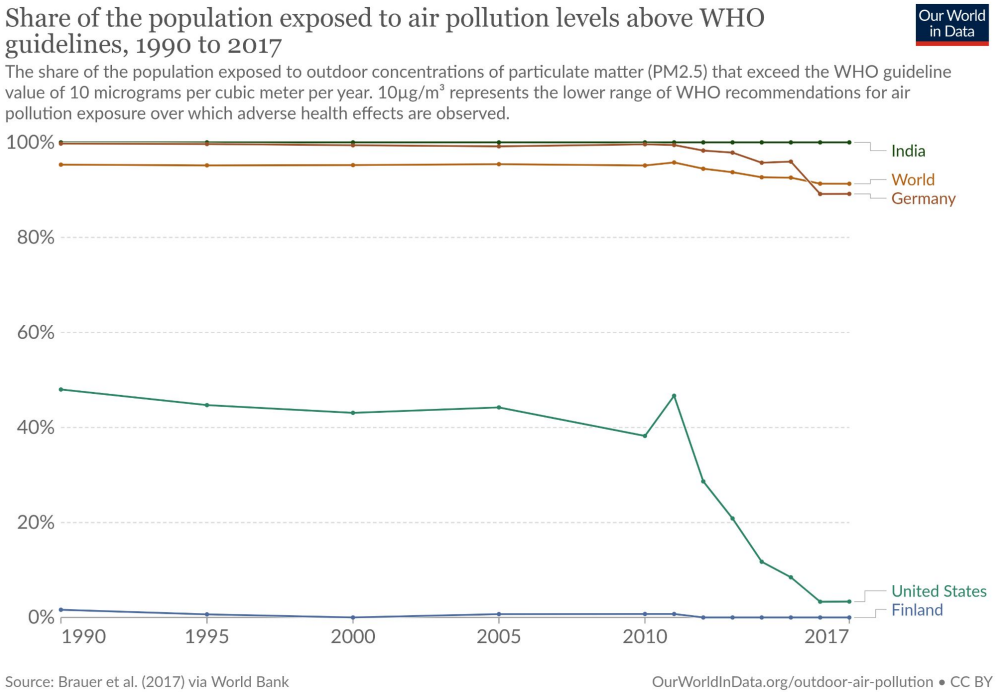
Source: IHME, Global Burden of Disease (2019); Brauer et al. (2017) via World Bank

OurWorldInData.org/air-pollution/ • CC BY

WHO guidelines

Berlin: worst air quality in Germany

Source: Federico Gambarini/dpa



Influencing factors

Weather

Temperature: summer vs. winter
Wind speed: dilution / accumulation
Wind direction: transport from source
Precipitation

Topography

Mountains, buildings, narrow streets, vegetation

Emission source

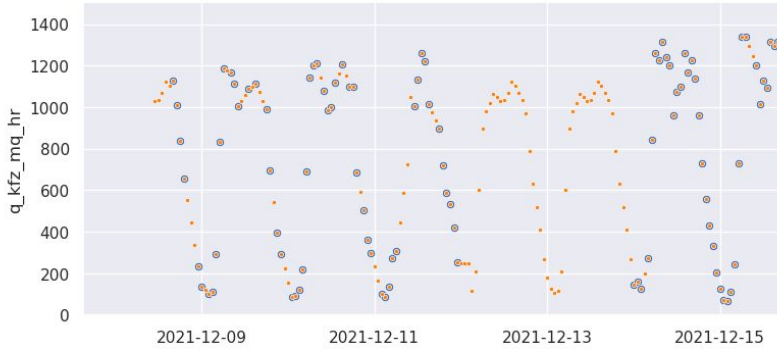
Natural vs. anthropogenic
Distance from source:
Traffic vs. volcanic eruption, Sahara dust

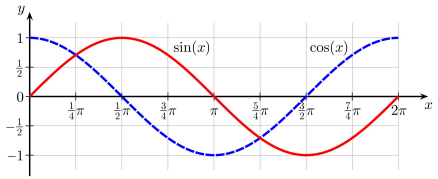
Data: hourly time series

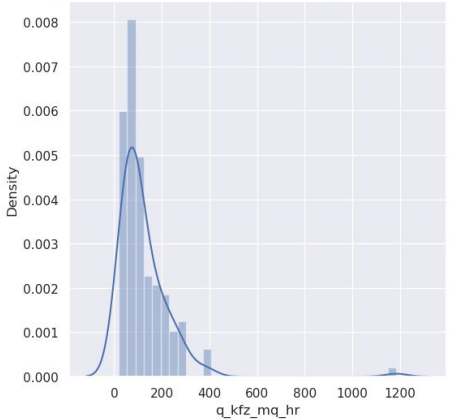
- **Air quality**
  - PM2.5 concentrations at measurement station in traffic (Mariendorfer Damm)
    - https://luftdaten.berlin.de
- **Features**
  - **Weather** data: air temperature, humidity, wind speed, wind direction, precipitation (Alexanderplatz)
    - https://opendata.dwd.de
  - **Traffic** data: quantity and velocity of cars per street (Mariendorfer Damm & distribution over other streets)
    - https://api.viz.berlin.de/daten/verkehrsdetektion
  - **Time** features: hour, weekday, month

Preparation steps

- **Clean data**
  - Measurement errors
- **Fill gaps**
  - Not straightforward for e.g. cyclical data







- **Feature engineering**
  - Transform cyclical features: wind, time
    - sine & cosine
  - Lagged features of traffic data
  - Add traffic distribution stats as features:
    - mean, sd, min, max, skewness, kurtosis
- **Feature Selection**
  - Feature importance

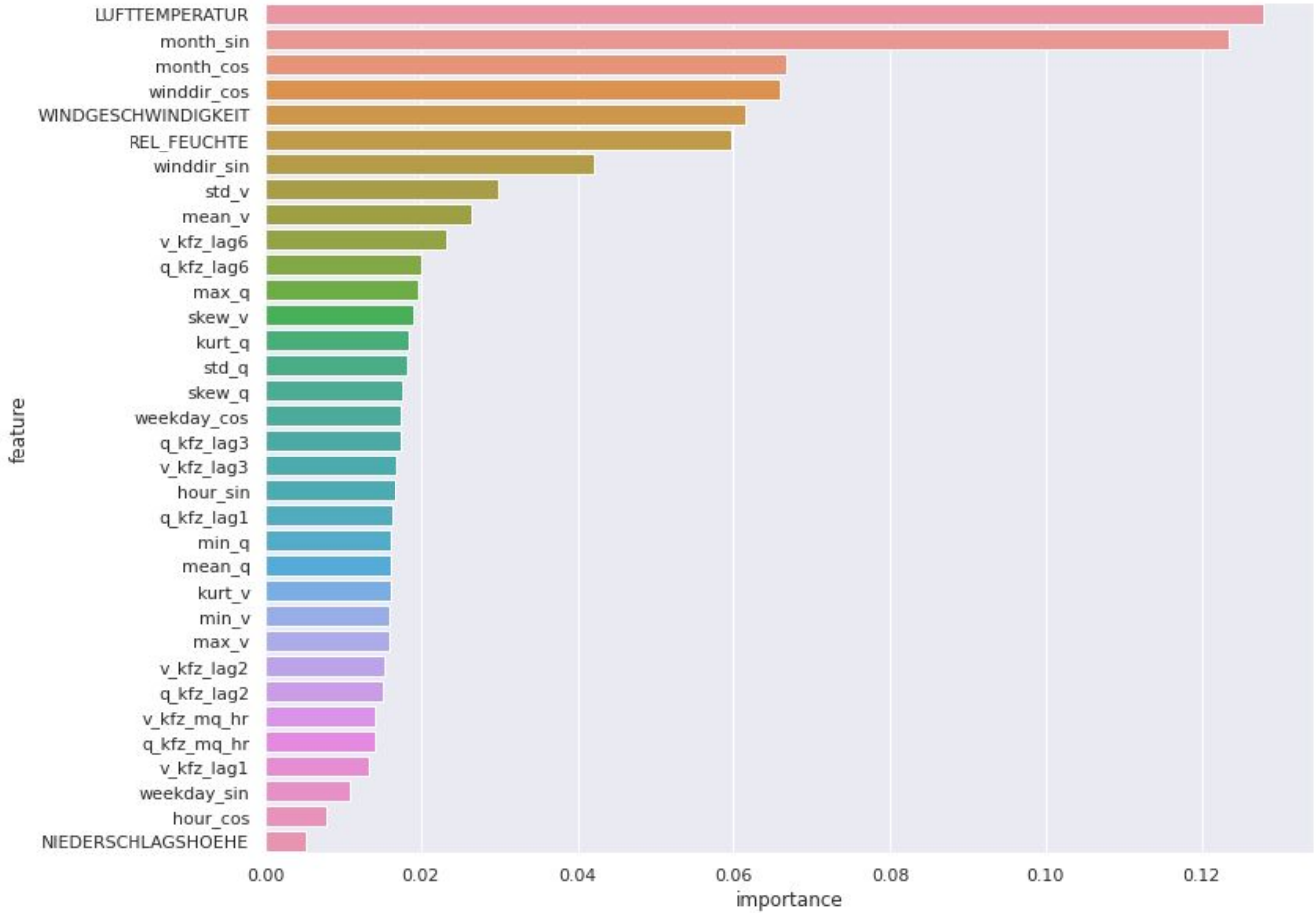
Models and results

- Supervised learning – regression – time series
- GridSearchCV for all models
  - K-fold cross-validation
  - Hyperparameter tuning
- Linear Regression
  - Accuracy: train 0.29, test 0.28
- Linear regression with polynomial features
  - Accuracy: train 0.64, test 0.49
- Random Forest Regression
  - Accuracy: train 0.69, test 0.61

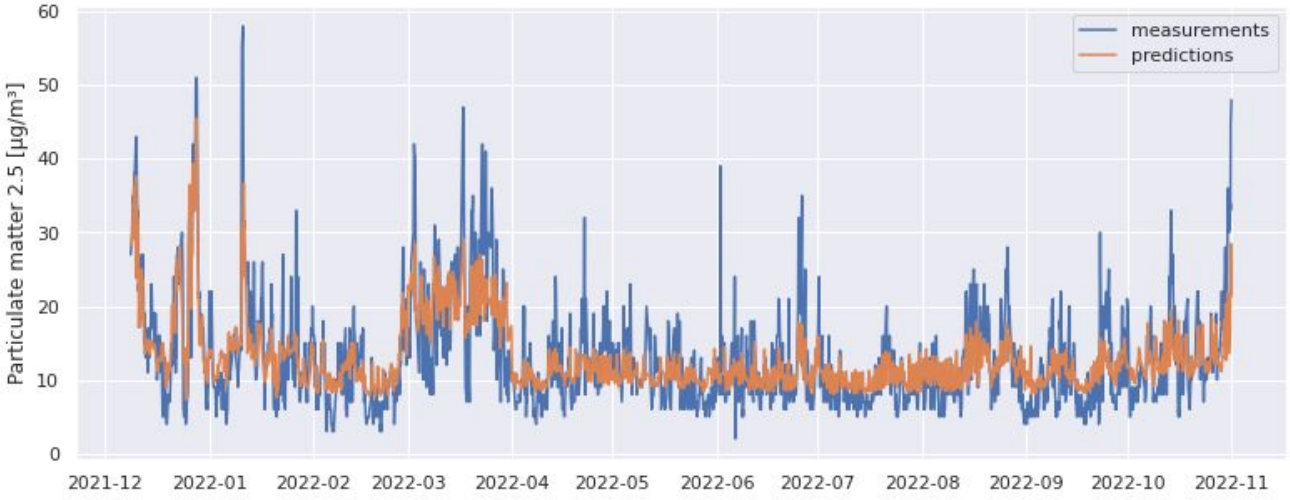
Source: www.askideas.com

Feature importance

drop features:
importance <0.02



Results


measurements
predictions
Particulate matter 2.5 [µg/m³]
60
50
40
30
20
10
0
2021-12
2022-01
2022-02
2022-03
2022-04
2022-05
2022-06
2022-07
2022-08
2022-09
2022-10
2022-11

Implications

- Weather features most important
  - nonlinearity
- Traffic: speed and speed variability seems more important than number of cars
  - speed limits would improve air quality
- Limits of machine learning for environmental observational data
  - often have gaps, discontinuities, not long enough
  - need for relevant features (often not available)
- Given the limits of the data: RandomForest proved quite powerful

Sources: REUTERS/Fabrizio Bensch | Grzegorz Polak/Fotolia.com
MagMac83/depositphotos.com

Many thanks!