Suitability of low-cost sensor data for land-use-based particulate matter modelling

Janani Venkatraman Jagatha, Christoph Schneider, Tobias Sauter

Humboldt-Universität zu Berlin, Geography Department, Climate Lab

Contact: Janani Venkatraman Jagatha, janani.venkatraman.jagatha@geo.hu-berlin.de, +49 30 2093 6898

Idea and concept

- Air pollution is considered to be a major global health concern that causes one in nine deaths worldwide. Although a number of sources and factors have been identified as a cause of air pollution it is difficult to pinpoint a particular source and manage it due to the variability of the pollutants in space, time, and the socio-economic factors involved.
- advances in micro-sensor technologies and low-cost due to production facilities enable sensors that are simple in design, lightweight and easier to deploy in larger scales.

Research Question

- Is it possible to use low costs sensor based mobile measurement systems to identify major sources that contribute to higher aerosol concentration in time and space using land use based regression methods?
- How do conventional linear regression models compare to Random forest methods?

Methodology

- Particulate matter sensor, OPC-N2 and temperature and humidity sensors (SHT35) were used in a mobile measurement platform (bicycle) to collect data in three suburbs of Berlin, Germany (Fig. 1).
- A random-forest model is developed using the collected data as the target and spatial variables such as local climate zones, land use types, building height, building type, leaf area index, etc., as predictors.

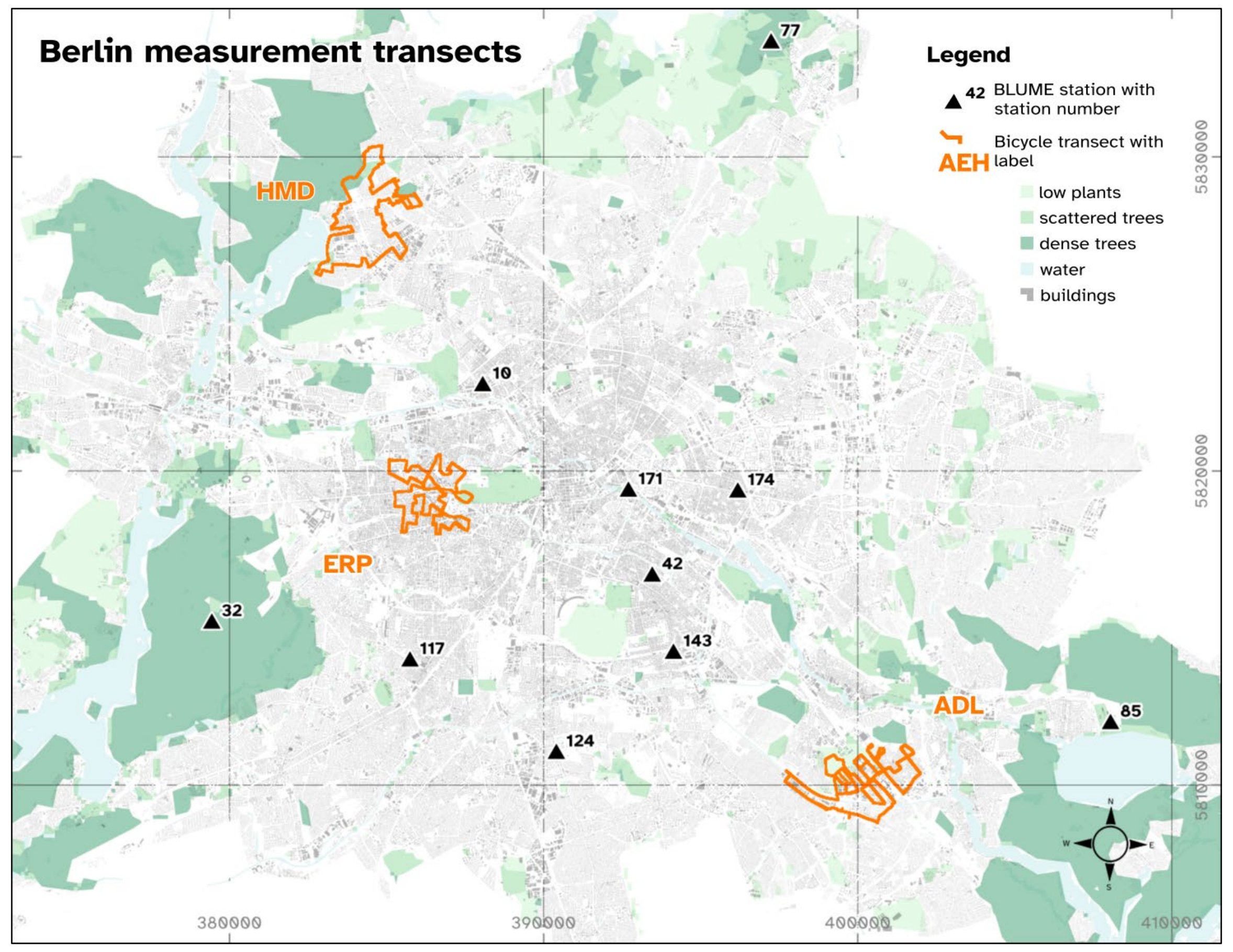

Fig. 1: Mobile measurement transects in Berlin-Germany at Hermsdorf (Hmd), Charlottenburg-Ernst-Reuter-Platz (ERP) and Adlershof (Adl).

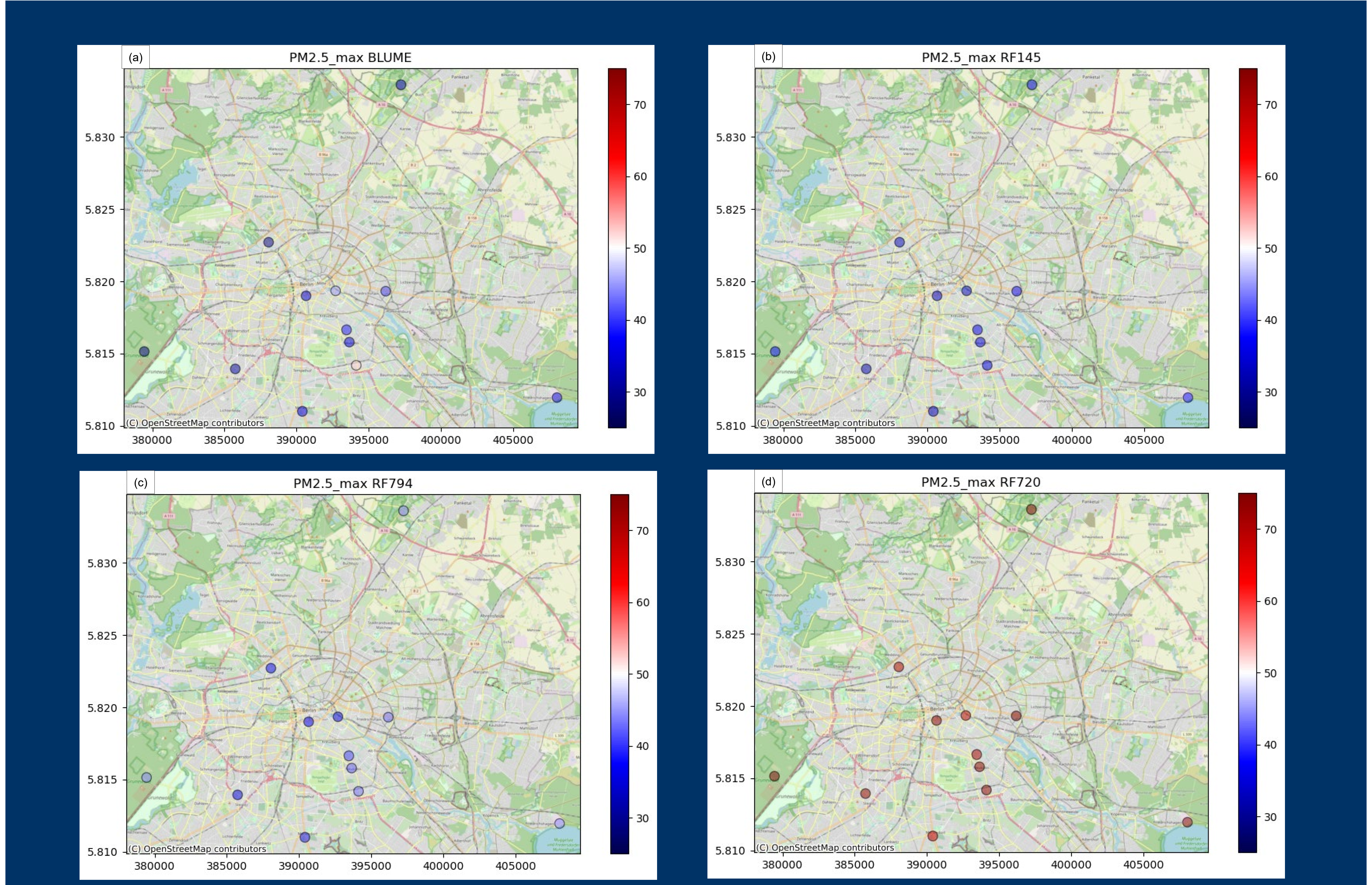

Fig. 4: (a) Maximum PM2.5 concentrations recorded in various BLUME-Air-quality measurement network stations in Berlin-Germany (https://luftdaten.berlin.de/lqi). (b, c, d) Maximum PM2.5 concentrations predicted using Random forest model RF145 (b), RF794 (c) and RF720 (d). The concentration of PM2.5 is measured in µg/m³ . The PM2.5 concentration corresponding to the points are shown on the colour bar on the right hand side of each map. UTM Easting and UTM Northing are represented on the X and Y axes respectively. CRS = EPSG 25833.

	Best Performance	Worst Performance
Including stations with larger green spaces	RF145 : RMSE = 6.64	RF720 : RMSE = 36.68
Excluding stations with larger green spaces	RF794 : RMSE = 5.67	RF720 : RMSE = 34.36

Key takeaways

- Random forest model performs best when compared to multi variate regression and Lasso regression.
- Population density and traffic volume contribute the most to PM2.5 concentrations
- Citywide statistical modeling is possible. Low computation intensity and can be applied quickly.
- Training is required with more data representing green spaces.

SPONSORED BY THE

Federal Ministry of Education and Research

FONA Forschung für Nachhaltigkeit

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Preliminary results

- Variance decomposition reduces the number of predictors from 180 to 54 for PM2.5max (Fig. 2).

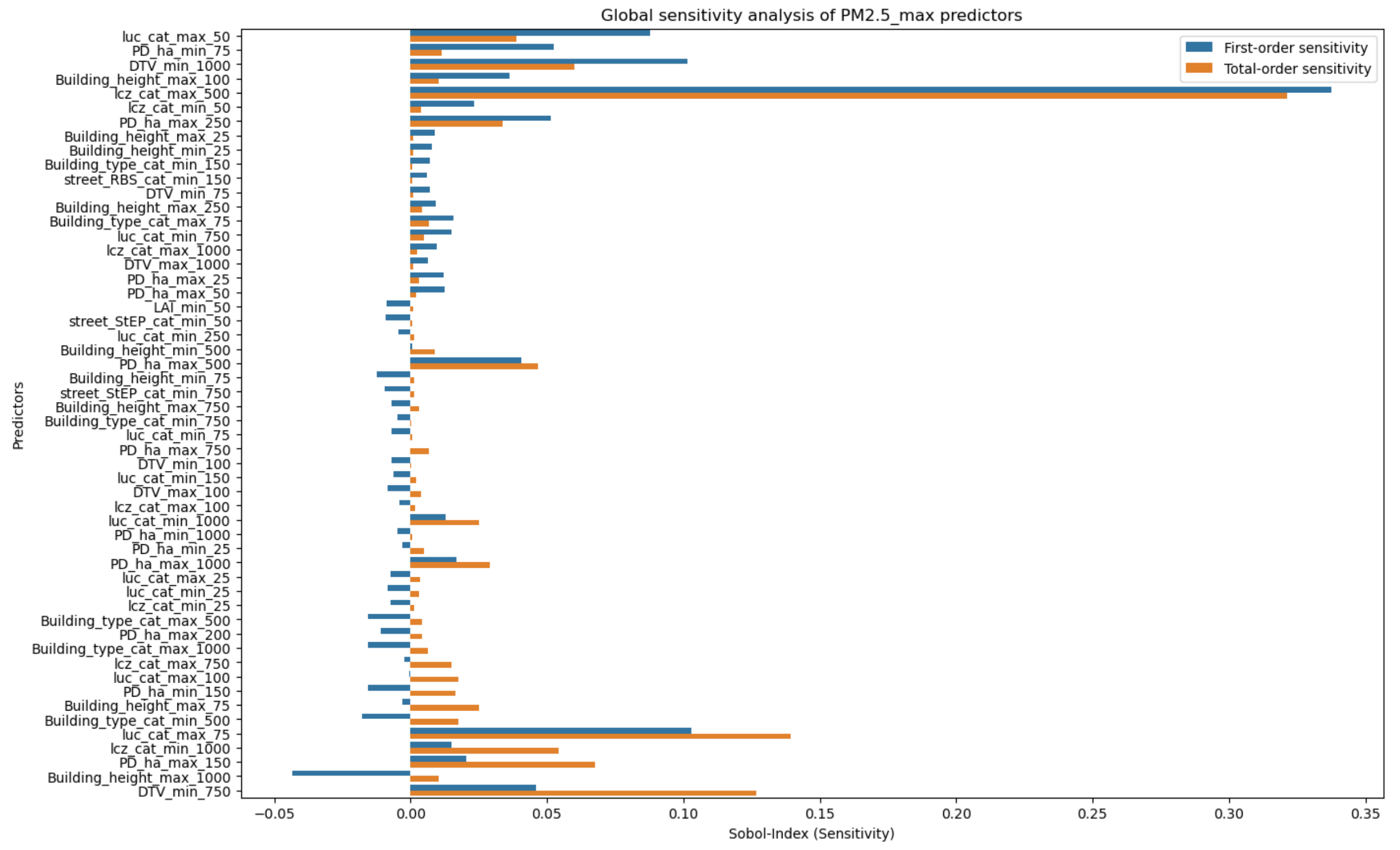

Fig. 2: Variance decomposition of PM2.5_max predictor data. Blue indicate the first order sensitivity of the predictors and orange indicates the total-order sensitivity of the predictors. The predictors used are shown on the y-axis, where the number indicates the buffer size in meter.

- Random forest model performs better with an R² of 0.83 compared to MVR (0.43) and Lasso regression (0.66) (Fig. 3).

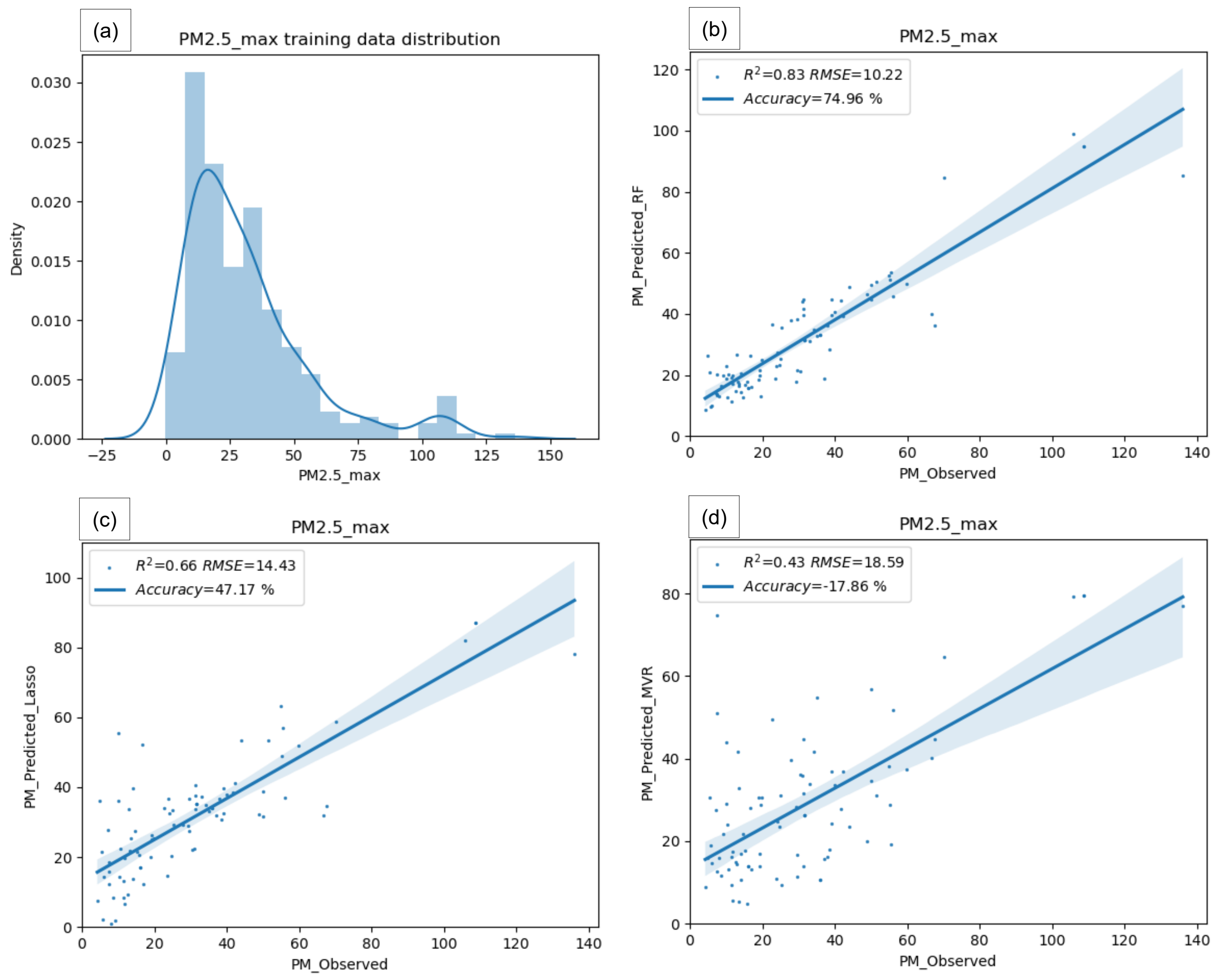

Fig. 3: (a) Distribution of PM2.5_max training data. (b,c,d) scatter-plot between observed PM2.5_max against the predicted PM2.5_max using random-forest model (b), Lasso regression (c) and multi variate regression (d).

Next steps

- Apply model to whole of Berlin city.
- Check if similar methodology can be used for predicting PM1 and PM10 concentrations.