Characterizing Air Quality in Urban Areas with Mobile Measurement and High Resolution Open Spatial Data: Comparison of Different Machine-Learning Approaches Using a Visual Interface

Yao Shen[1], Stephan Lehmler[2](✉), Syed Monjur Murshed[2], and Till Riedel[1]

[1] KIT, Karlsruhe, Germany
uivnz@student.kit.edu, riedel@teco.edu
[2] EIFER, Karlsruhe, Germany
{lehmler,lehmler}@eifer.org

Abstract. Air quality is one of the most important topics in our urban life, as it is of great significance for human health and urban planning. However, accurate assessment and prediction of air quality in urban areas are difficult. In major cities, typically only a limited number of air quality monitoring stations are available, and inferring air quality in the un-sampled areas throughout the city is challenging. On the other hand, air quality varies in the urban areas non-linearly; it is highly spatially dependent and considerably influenced by multiple factors, such as building distribution, traffic situation and land uses.

In this research, we model air quality in the city of Augsburg using spatial features and high quality sensor data. We identify spatial features such as types and areas of different land uses, road networks with high resolution.

We integrate open available data to the air quality prediction. In this regard, we compare a simple baseline model with linear regression models (Ordinary Least-Squares and Ridge Regression) and tree-based machine-learning models (Gradient Boosting and Random Forest). In our evaluation, given the non-linearity of the data, tree-based models outperform all linear models, which are commonly used in literatures.

In addition, we created an interactive and visual dashboard. This dashboard demonstrates the analytical workflow, gives insight into model performance and uncertainty and visualizes the results.

Keywords: Air quality · Land-use regression · Dashboard · Machine-learning

1 Introduction

Smart Cities can be defined as cities that predict and accommodate citizens' needs using different types of data and sensors to provide information and applying advanced information technologies [1]. They will contribute to the efficient management of assets

H. Santos et al. (Eds.): SmartCity 360 2019, LNICST 323, pp. 115–126, 2020.
https://doi.org/10.1007/978-3-030-51005-3_12

and resources. In addition, as mentioned by [2], “all activities for a smart city related to public services can be managed by developing a platform (dashboard) for monitoring all relevant data.” Visualization of data is important in Smart City applications, since it can communicate information clearly and efficiently, which supports people’s decision-making process in the city planning.

In this context, advanced analysis tools have become essential for a Smart City development. Nowadays, machine-learning techniques play a key role in data analysis, predictive modeling and visualization [3]. In regard to machine-learning, one of the most complicated problems is the diverse types of input data and the amount of available data. In particular, for a smart city data analysis, spatial information of events and changes around the city is required [4].

Relevant spatial data (also known as geospatial data or geographic information) around a city is usually collected using ground surveying, remote sensing, more recently through mobile mapping, geo-located sensors, geo-tagged web contents, Volunteered Geographic Information (VGI) and so on [5]. In this study, we aim to tackle the smart city issues by using machine-learning based spatial data analysis.

Air quality information, such as the concentration of certain particulate matters like PM2.5 and PM10, is important for the quality of our urban life, as it is of great significance for human health and city management. However, in major cities, there is typically only a limited number of air quality monitoring stations. Inferring air quality in the areas not covered by measurements is challenging as air quality varies in the urban areas non-linearly, it is highly spatial-temporal dependent and considerably influenced by multiple factors, such as meteorology, traffic volume and land uses [6–8].

Compared to the previous works, the contribution of this paper lies in several aspects. We build land-use regression (LUR) models on mobile measurement data of air pollution, we can conclude that mobile monitoring data is suited for LUR modelling at a higher spatial resolution and therefore they can be used to characterize and prove the spatial variability of air quality in the complex city area. We identify more spatially related features with higher resolution, such as types and areas of different land uses, information regarding road networks. By extracting and utilizing data from VGI projects such as OpenStreetMap (OSM), we evaluate the usefulness of the crowd-sourcing data and the contribution of the open spatial data. We integrate these features to the air quality prediction; our proposed approach can be applied to assess air quality in any new urban areas. Afterwards, a visual interface is developed to demonstrate the work-flow of the analysis, including the data exploration, correlation analysis, model comparison and the inference of air quality for a new city area with a finer granularity.

2 Related Work

The LUR is one popular approach for predicting spatial variations in air pollution. As stated by [9], the general concept of a LUR is based on two steps: first, the dependencies between explanatory variables and monitored pollution levels are evaluated using ordinary least-square (OLS) regression for all measurement locations; second, the relationships found between concentrations and the explanatory variables are used to infer concentration levels at locations without measurements but with available land-use data.

LUR modeling requires air pollution measurements at multiple locations across the study area, according to a review by [10], stationary monitoring used by LUR is typically at 20 to 100 locations, spread over the study areas. [11] investigated in the effect of the number of monitoring sites on the LUR performance and their result suggested that for complex urban settings, a LUR model should be based on a large number of measurements (> 80 in their study), higher R^2 achieved from smaller number of monitoring sites for LUR models do not reflect the true predictive ability. A cost-effective alternative way to collect data at a high spatial resolution is to use mobile measurements, however, only few studies use mobile measurements as a basis for LUR modelling [12]. Based on surveyed papers [9, 12–16], the OLS regression models built on mobile measurement data explained between 0.40 and 0.60 of the observed variability in concentrations (R^2 tested on the training datasets). The time resolution of monitoring is usually 1 s.

To summarize the predictor variables used in the LUR models, frequently used data include: area of land-use, road network or traffic information, physical geography such as elevation and slope, and meteorological data. Some studies [15, 17] also used demographic data such as number of inhabitants and population density. Study of [15] brought the functions of year, day and hour into their modeling. Most studies have assessed a large number of potential predictor variables in modeling air quality concentrations and selected a smaller set of variables to the final models. Because of data availability, extraction and definition of predictor variables differ substantially between studies. For further development of LUR methods that can be transferred to other areas, open spatial data is worth getting more attention.

In addition, machine-learning approaches such as ensemble regression methods have been utilized to handle complex and nonlinear relationships that exist within data and produce forecasting models with comparable performance in practice. Based on our review of papers [6, 7, 18, 19] from the domain of spatial data analysis, since the prediction accuracy follows algorithm design, the machine-learning algorithms are crucial for building air quality prediction models, whereas statistical models have not been heavily used recently. Moreover, the random forest based approach is a prominent technique in selecting variables and inferring air pollution values.

3 Study Area and Required Data

3.1 Study Area and Description

The study site is the city of Augsburg, Swabia, Bavaria, Germany, the third-largest city in Bavaria (after Munich and Nuremberg) with a population of 300,000 inhabitants ($N48^{\circ}22^{'}$, $E10^{\circ}54^{'}$, 2000 inhabitants km^{-2}). The municipal area of Augsburg covers 147 km^2 and the city border is 78 km long. The widest point north to south is 23 km and east to west is 15.5 km. Residential and traffic areas make up only 36% of the city's land-use; one-third is devoted to agriculture and nearly 24% is forestland. The inner city of Augsburg covers approximately 6.8 km^2 and it is within the primary highway B 300 at the south and the primary highway B 2 at the east. Multiple railways locate at the west border, a tertiary highway borders the inner city at the north. The study area covers approximately 4 km^2, data was collected mostly in the inner city area of Augsburg, especially within the inner city borders at the south and east. There is no primary highway located across the study area, however, multiple railways pass through it at the southwest, shown in Fig. 1.

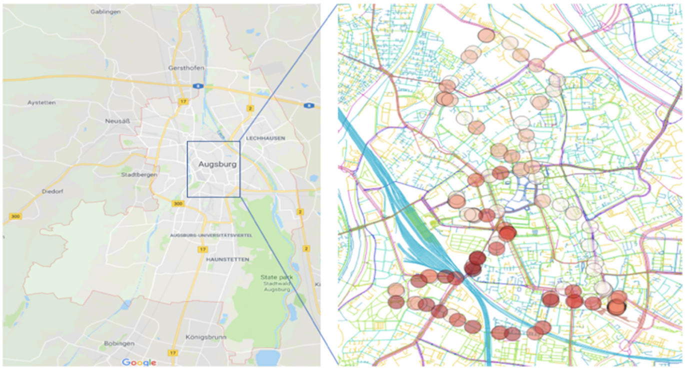


Fig. 1. The study area and mobile measurements during the first day of the IOP project in the city of Augsburg. To give a clear display of the data points and the buffer (50 m) around the measurements, we applied time-series re-sampling's approach to take the median PM1 value during a 5 min period (original data with 1-s resolution). The color gradient represents the height of PM1 and the data is displayed on a map, which shows different types of railways and roads extracted from OSM for the study area. (Color figure online)

3.2 Mobile Measurement Data

The mobile measurements were taken from the Intensive Operation Period (IOP) of the particulate matter measurement project SmartAQnet[1]. During the first day of the IOP, on the 26th Sep. 2018, a 11.4 km long route was done on foot 3 times in the study area. Measurements were taken from 12:16:02 to 23:10:14, about 11 h in total, approximately 4 h for each walk. The route was then repeated during the next day and one month later. The air pollution data were measured using the DustTrak DRX, which has a 1 s time resolution. The data types (unit originally in mg/m^3, multiplied by 1000 to unit $\mu g/m^3$) are: PM1, PM2.5 and PM10.

3.3 Geographic Data

The OSM project[2] is a repository that provides user-generated street maps. It is a powerful source of information that can be used free to understand and to model the built environment. OSM is available as a vector data collection comprising point features (nodes), line features (ways) and polygon features (ways and relations). Each feature has at least one "tag" (key-value-pair) describing it. See [20, 21] for more detailed description.

[1] https://smartaqnet.github.io/.

[2] https://en.wikipedia.org/wiki/OpenStreetMap.

The OSM data are downloaded and extracted from Geofabrik's free download server[3]. The file used for our study area is the file for Swabia, Bavaria, Germany[4]. The downloaded data is in a number of ESRI compatible shapefiles[5].

4 Implementation and Results

4.1 Aggregation of Concentrations

Following the study of [9], the measured air pollution concentrations of all three days are projected on a grid with 6100 cells, each of size 50 m * 50 m, covering the complete region of interest. For this purpose, the maximum and minimum values of the coordinates are utilized to define the bounding box for the grid building. We use this method to develop models for mean pollutant concentrations at a high spatial resolution. All measurements are performed in each cell. To assure data quality, we require the model input to be based on grid cells containing at least 50 mobile measurements, this removes approximately 4.6% of the data. Within the 6100 grid cells, 760 cells cover data points and 442 cells of them cover at least 50 measurements. The original datasets of the three days contain 131050 data points in total, 124985 data points remained for building our model after the selection of grid cells.

Considering the temporal aspect of the measurements, the information of the hour h is brought into the model to capture the temporal patterns within a day, such as the pattern during the rush hours. After the manipulation of aggregation, 3363 data points are used in the final dataset for the further analysis.

4.2 Feature Generation

We use the geographic data available from OSM, which includes land-use, buildings, traffic, railways and roads. After the aggregation of concentrations, the centroids of grid cells are used to draw 50 m buffers and to extract geographic features from OSM. Two types of features from OSM are considered: the polygon features and the line features. We generate the buffers and intersect these buffers with the OSM geographic layers. More specifically, the intersected areas for each values of keys are calculated for polygon features whereas we extract the intersected lengths for each types of the line features, such as road and railway. Based on the study of [18], we generate a vector as geographic abstraction for each aggregated mobile measurement location. In the next step, we quantify and evaluate the importance of individual components in the geographic abstraction vectors.

4.3 Preprocessing

In the experiment, we started with randomly taking left-out samples in a small size from the data and using the remaining data as the training set to predict the PM values for the

[3] http://download.geofabrik.de/ 2018/12/12 17:02.

[4] http://download.geofabrik.de/europe/germany/bayern/schwaben-latest-free.shp.zip.

[5] Geospatial vector data format for storing geometric location and associated attribute information.

left-out samples drawn before. That means, the model fits on the training dataset, then one uses the left-out samples as the ground truth to calculate the prediction accuracy. After we split the dataset randomly, the number of the observations in the training dataset is 2690, the percentage of data in training set is approximately 79.99%; the number of observations in the left-out set, i.e., test dataset is 673, the percentage of data in test set is approximately 20.01%.

For geographic features, we standardize them by removing the mean and scaling to unit variance, calculated from the training set. Standardization is useful when one of the variables has a very large scale, since this might lead to regression coefficients of a very small order of magnitude.

Specifically, for the sake of the interpretation of variables in linear models, the feature *hour* is one-hot encoded. They are therefore not treated as numerical but as categorical variables. This was done to improve the performance of the linear models, because air quality varies non-linear with time.

In order to assess the relative importance of the features we generated, we apply the means of importance measure based on random forest algorithm, namely Mean Decrease Impurity on the training dataset. The impurity (residual sum of squares) decreases from each feature can be averaged for a forest and the ranking of features is obtained according to this measure. Following the proposed approach of [18], we construct the weighted features by multiplying the values of all aforementioned preprocessed features by their relative importance. In this way, we can particularly penalize trivial features.

4.4 Experimental Result

To predict PM concentrations for a target location that does not have air quality measurements, we train different machine-learning models. The most commonly used LUR models in the literature apply ordinary least-square regression (OLR). We examine OLR and ridge regression. In addition, we examine two tree-based machine-learning algorithms: random forest and gradient boosting. The tree-based models are applied particularly for handling nonlinear relationships.

We tune the hyper-parameters by 5 folds cross-validated grid-search of each model to further improve their performance (tuned parameters for random forest: n_estimators = 256 for prediction on PM1 and PM2.5; n_estimators = 512 for prediction on PM10, min_samples_split = 2; for gradient boosting: n_estimators = 1024, learning_rate = 0.25; for ridge regression: alpha = 0.03125). We use the central tendency, namely, the mean of the output value observed in the training data, as a baseline to compare the results of all of our regression models.

From the Table 1, Table 2 and Table 3 we can compare the prediction's results for the three pollution types. The best prediction is achieved on the PM1. According to Table 1, gradient boosting regression generated the best training score whereas random forest performed the best on the test dataset. Tree-based models outperformed linear regression models and all the models performed better than the baseline.

Table 1. Result of prediction on PM1.

Regressors	R^2 (Train)	R^2 (Test)	RMSE	MAE	MAPE
RandomForest	0.755999	0.41941	13.013	4.61543	15.5181
GradientBoosting	0.801266	0.413835	13.0754	5.03747	17.3187
RidgeRegression	0.263594	0.177901	15.4848	7.05862	27.4429
LinearRegression	0.28177	0.190075	15.3697	6.92811	26.5321
Baseline	0.00	0.00	17.0807	9.12583	38.3164

Table 2. Result of prediction on PM2.5.

Regressors	R^2 (Train)	R^2 (Test)	RMSE	MAE	MAPE
RandomForest	0.747725	0.408235	13.3858	4.86004	16.3528
GradientBoosting	0.792927	0.399515	13.4841	5.3184	18.1369
RidgeRegression	0.244739	0.173651	15.818	7.50331	29.2644
LinearRegression	0.264099	0.187799	15.682	7.34138	28.0926
Baseline	0.00	0.00	17.4022	9.53591	39.8149

Table 3. Result of prediction on PM10.

Regressors	R^2 (Train)	R^2 (Test)	RMSE	MAE	MAPE
RandomForest	0.752592	0.299473	26.6166	7.05102	18.4924
GradientBoosting	0.849792	0.305144	26.5086	7.69487	20.9269
RidgeRegression	0.147497	0.0545743	30.921	9.5087	29.5942
LinearRegression	0.162844	0.0625943	30.7895	9.33308	28.4433
Baseline	0.00	0.00	31.811	11.2264	37.669

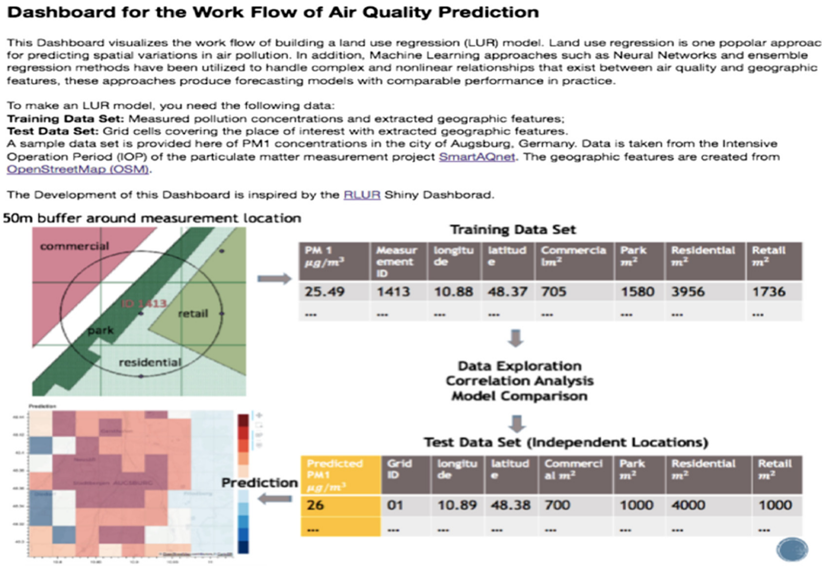
5 Visualization and Dashboard Development

We present an application to simplify the LUR modeling process. We develop a user-friendly dashboard using the Python (3.6) programming language, particularly, the visualizations of all parts of this application have been built with the Python package Bokeh[6] (1.0.4). This application is developed as a processing pipeline to model air quality based on sensor data and spatial information. The main goal of this dashboard is to provide an introduction of the work-flow for predicting air quality using LUR. Our model uses openly available data, which also offers the possibility to use it on other study area. The

[6] https://bokeh.pydata.org/en/latest/.

development of the dashboard is inspired by the Smart City applications introduced by [22] and the RLUR Shiny Dashboard [23].

To make a LUR model on the dashboard, users will need a training dataset with measured pollution concentrations and extracted geographic features, and a test dataset, which contains grid cells covering the place of interest with extracted geographic features. A sample dataset for training is provided here of PM1 concentrations in the city of Augsburg, Germany. The data description and complete approach for feature extraction is described in previous sections. A sample test dataset is provided for the whole city area of Augsburg. The bounding box of Augsburg is defined using the Nominatim API (3.2). A short description is also provided on the first page of the dashboard (Fig. 2).



Dashboard for the Work Flow of Air Quality Prediction

This Dashboard visualizes the work flow of building a land use regression (LUR) model. Land use regression is one popolar approach for predicting spatial variations in air pollution. In addition, Machine Learning approaches such as Neural Networks and ensemble regression methods have been utilized to handle complex and nonlinear relationships that exist between air quality and geographic features, these approaches produce forecasting models with comparable performance in practice.

To make an LUR model, you need the following data:
Training Data Set: Measured pollution concentrations and extracted geographic features;
Test Data Set: Grid cells covering the place of interest with extracted geographic features.
A sample data set is provided here of PM1 concentrations in the city of Augsburg, Germany. Data is taken from the Intensive Operation Period (IOP) of the particulate matter measurement project SmartAQnet. The geographic features are created from OpenStreetMap (OSM).

The Development of this Dashboard is inspired by the RLUR Shiny Dashborad.

50m buffer around measurement location

Training Data Set

PM 1 $\mu g/m^3$	Measurement ID	longitude	latitude	Commercial m^2	Park m^2	Residential m^2	Retail m^2
25.49	1413	10.88	48.37	705	1580	3956	1736
...	...	...	...	...	...	...	...

Test Data Set (Independent Locations)

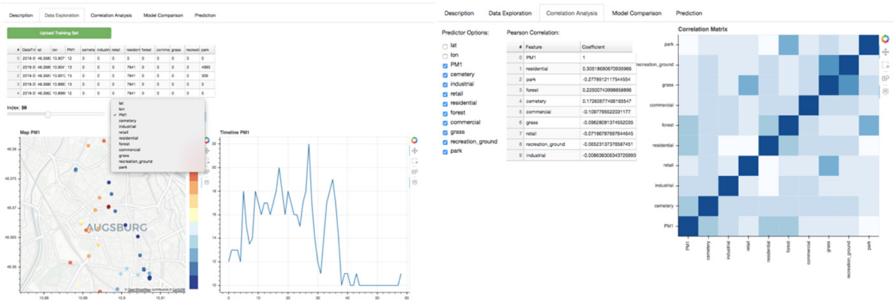
Predicted PM1 $\mu g/m^3$	Grid ID	longitude	latitude	Commercial m^2	Park m^2	Residential m^2	Retail m^2
26	01	10.89	48.38	700	1000	4000	1000
...	...	...	...	...	...	...	...

Fig. 2. Description page of the dashboard

5.1 Data Exploration

The first step of the analysis is data exploration. A sample dataset for training is provided, however, we also allow users to upload a training dataset from a local data source using Upload Training Set tool to make the dashboard more flexible to use, see Fig. 3. To do that, we utilize the CustomJS module to supply a snippet of JavaScript code that should be executed in the browser to open a file dialog. The uploaded data table should be saved as a csv text file and as another format restriction, the uploaded data table should contain two columns named as “lat” and “lon” respectively with the WGS 84 coordinates. Moreover, object data type is excluded for further development of the dashboard.

The target variable can be selected, e.g., PM1. Indexes are added to the data table as row labels. With an index slider, users can explore the data nicely, the map shows the



#	DateTi	lat	lon	PM1	cemete	industri	retail	resident	forest	comme	grass	recreati	park
0	2018-0	48.358	10.907	12	0	0	0	0	0	0	0	0	0
1	2018-0	48.358	10.904	13	0	0	0	7841	0	0	0	0	4990
2	2018-0	48.358	10.901	13	0	0	0	7841	0	0	0	0	309
3	2018-0	48.358	10.898	13	0	0	0	7841	0	0	0	0	0
4	2018-0	48.356	10.896	12	0	0	0	7841	0	0	0	0	0

Predictor Options:

- □ lat
- □ lon
- ☑ PM1
- ☑ cemetery
- ☑ industrial
- ☑ retail
- ☑ residential
- ☑ forest
- ☑ commercial
- ☑ grass
- ☑ recreation_ground
- ☑ park

Pearson Correlation:

#	Feature	Coefficient
0	PM1	1
1	residential	0.3051869067093966
2	park	-0.2778512117544554
3	forest	0.2250074399858896
4	cemetery	0.17260677468165547
5	commercial	-0.1097795522031177
6	grass	-0.09628081374552035
7	retail	-0.07186787697644645
8	recreation_ground	-0.06523137376587451
9	industrial	-0.008638306343726993

Fig. 3. Data exploration and correlation analysis (Color figure online)

location of the measurements and color code signifies the value of the target variable. In this sample dataset, the collection's time is recorded. We therefore sorted the table by time and plot the time series graph to show the temporal aspect of the data.

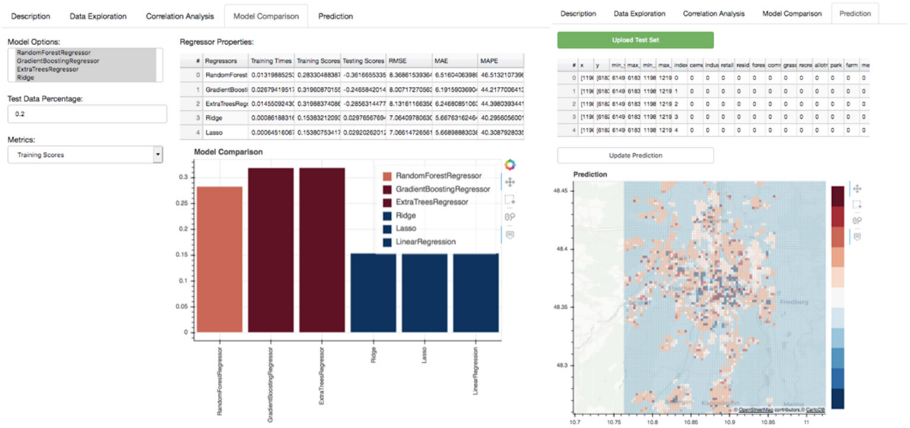
5.2 Correlation Analysis

As the next step, we apply correlation analysis and help users to identify which types of variables are more important for predicting the target variable. We use the function of pandas (0.23.4) DataFrame to compute pairwise Pearson correlation of columns, excluding NA/null values. As initial state, the target variable, in this case PM1 is selected automatically. The Correlation Matrix shows the correlation of all other variables except PM1 to detect the multicollinearity. As illustrated in Fig. 3, the Pearson Correlation table shows the correlation of the selected features with the target variable and this table is sorted in a descending order by the absolute value of the correlation. Selected features will be brought to the next step and will be used for training different models and comparing the model performance.

5.3 Model Comparison

The dashboard offers six different machine-learning algorithms to predict the air pollution levels using selected features from the last tab. The six algorithms are random forest, gradient boosting, extra trees, ridge, lasso and linear regression. To train the models, we build a function to execute each algorithm through a pipeline, which will fit the regressors on the training dataset, test them on the validation dataset and record performance metrics. For applying the algorithms, we use the standard methods from Python library scikit-learn (0.19.1).

As shown in Fig. 4, the Model Options is a multi-selection's tool, initially, all the models are selected for comparison. The Test Data Percentage can also be given by users. As the metrics, we record six measurements in total: Training Time, Training Score, Testing Score, RMSE, MAE and MAPE. The Training Score and Testing Score specify the R^2 on the training set and on the validation set respectively. Using the Regressor Properties table and the Model Comparison bar chart, we can get the best model according to the selected metrics. After the comparison, one can use the Model Options tool again to manually identify the best model, which will be applied for the prediction in the next step.



Model Options:
RandomForestRegressor
GradientBoostingRegressor
ExtraTreesRegressor
Ridge

Test Data Percentage:
0.2

Metrics:
Training Scores

Regressor Properties:

#	Regressors	Training Times	Training Scores	Testing Scores	RMSE	MAE	MAPE
0	RandomForest	0.01319885253	0.28330488387	-0.3616655335	8.36861539364	6.51604063986	46.5132107396
1	GradientBoosti	0.02679419517	0.31960870155	-0.2465842014	8.00717270563	6.19159036904	44.2177006413
2	ExtraTreesRegr	0.01455092430	0.31988374086	-0.2856314477	8.13161166356	6.24680851063	44.3980393441
3	Ridge	0.00086188316	0.15383212093	0.0297656769	7.06409780630	5.66763162464	40.2956056001
4	Lasso	0.00064516067	0.15380753417	0.02920262012	7.06614726561	5.66898883038	40.3087928035

#	x	y	min_	max_	min_	max_	index	ceme	indus	retail	resid	fores	comr	grass	recre	allotr	park	farm	me
0	[119	[618	6149	6183	1198	1219	0	0	0	0	0	0	0	0	0	0	0	0	0
1	[119	[618	6149	6183	1198	1219	1	0	0	0	0	0	0	0	0	0	0	0	0
2	[119	[618	6149	6183	1198	1219	2	0	0	0	0	0	0	0	0	0	0	0	0
3	[119	[618	6149	6183	1198	1219	3	0	0	0	0	0	0	0	0	0	0	0	0
4	[119	[618	6149	6183	1198	1219	4	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 4. Model comparison and prediction (Color figure online)

5.4 Prediction

In the last step, we visualize the prediction for the place of interest, see Fig. 4. To this end, we will need a test dataset. In the sample dataset, we built grid cells covering the city of Augsburg, each grid cell has the size 200 m * 200 m and land-use features are extracted for each grid cell. After that, we apply the model selected from last tab and predict the target variable, in this case, the PM1 value. We plot our prediction using color code on the map. The Upload Test Set button extend the flexibility of making predictions on this dashboard. We allow users to apply new test dataset for any city area. The Update Prediction button is used to make new predictions when any parameters of previous tabs have been changed, such as the target variable and the model.

6 Conclusion

6.1 Summary

This paper modeled air quality in the city of Augsburg using spatial features and mobile measurement data with high quality. We extracted and utilized data from OSM and identified spatial features such as types and areas of different land uses, road networks with high resolution. The advantages of our approach include that it used publicly available open data to construct the geographic predictor variables instead of using expensive datasets. Therefore, the built model can be easily used to infer air quality for other urban areas. In addition, our approach quantified the importance of geographic features on air quality prediction, enabled us to select features and integrate the important spatial factors automatically to the modeling, without using domain knowledge of air quality. We applied appropriate machine-learning approaches and compared the model performance using a visual interface (dashboard). A dashboard was developed at the end of this study to demonstrate the work-flow of the analysis, including the data exploration, correlation analysis, model comparison and the inference of air quality for a new city area with a fine granularity.

6.2 Limitations

The applicability of the LUR models obtained in this study is restricted by the characteristics of the input (air pollution) data, such as that the data points are collected using a single mobile sensor and they are only captured on the walked route. Due to the available data only covering 3 days, we were not able to include weather or seasonality effects into our model. Including additional measurements taken throughout the year should improve the relevance of our predictive model. Furthermore, the LUR models are only applied in a relatively small study area. How well the model would perform at a larger scale (e.g., including the peripheries and not only the city center) or even in another city area is still an open question. For instance, there is no primary highway located across the study area, however, the highway traffic could be an interesting factor to our study. As stated by [12], the generalization of the LUR model to areas where no measurements were made is limited, especially in predicting absolute concentrations. While this study showed some potential of mobile sensors and spatial features for air quality prediction, there is still more data needed for the evaluation of this approaches further applicability.

References

1. Hashem, I.A.T., et al.: The role of big data in smart city. Int. J. Inf. Manag. **36**(5), 748–758 (2016)
2. Suakanto, S., et al.: Smart city dashboard for integrating various data of sensor networks. in ICT for Smart Society (ICISS). In: 2013 International Conference (2013)
3. Gangappa, M., Mai, C.K. Sammulal, P.: Techniques for Machine Learning based Spatial Data Analysis: Research Directions (2017)
4. Bermudez-Edo, M., Barnaghi, P.: Spatio-temporal analysis for smart city data. In: Proceedings of WebConf 2018 (2018)
5. Li, S., et al.: Geospatial big data handling theory and methods: a review and research challenges. ISPRS J. Photogrammetry Remote Sens. **115**, 119–133 (2016)
6. Yu, R., et al.: RAQ–a random forest approach for predicting air quality in urban sensing systems. Sensors **16**(1), 86 (2016)
7. Zheng, Y., Liu, F. Hsieh, H.P.: U-Air: when urban air quality inference meets big data. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1436–1444. ACM (2013)
8. Kang, G.K., et al.: Air quality prediction: big data and machine learning approaches. Int. J. Environ. Sci. Dev. **9**(1), 8–16 (2018)
9. Hasenfratz, D., et al.: Deriving high-resolution urban air pollution maps using mobile sensor nodes. Perv. Mob. Comput. **16**, 268–285 (2015)
10. Hoek, G., et al.: A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. **42**(33), 7561–7578 (2008)

11. Basagaña, X., et al.: Effect of the number of measurement sites on land use regression models in estimating local air pollution. Atmos. Environ. **54**, 634–642 (2012)
12. Van den Bossche, J., et al.: Development and evaluation of land use regression models for black carbon based on bicycle and pedestrian measurements in the urban environment. Environ. Model Softw. **99**, 58–69 (2018)
13. Weichenthal, S., et al.: A land use regression model for ambient ultrafine particles in Montreal, Canada: a comparison of linear regression and a machine learning approach. Environ. Res. **146**, 65–72 (2016)
14. Hankey, S., Marshall, J.D.: Land use regression models of on-road particulate air pollution (particle number, black carbon, PM2.5, particle size) using mobile monitoring. Environ. Sci. Technol. **49**(15), 9194–9202 (2015)
15. Patton, A.P., et al.: An hourly regression model for ultrafine particles in a near-highway urban area. Environ. Sci. Technol. **48**(6), 3272–3280 (2014)
16. Kanaroglou, P.S., et al.: Estimation of sulfur dioxide air pollution concentrations with a spatial autoregressive model. Atmos. Environ. **79**, 421–427 (2013)
17. Habermann, M., Billger, M., Haeger-Eugensson, M.: Land use regression as method to model air pollution. Previous Results Gothenburg/Sweden. Procedia Eng. **115**, 21–28 (2015)
18. Lin, Y., et al.: Mining public datasets for modeling Intra-City PM2.5 concentrations at a fine spatial resolution. In: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM (2017)
19. Sun, L., et al.: Impact of land-use and land-cover change on urban air quality in representative cities of China. J. Atmos. Solar-Terrestrial Phys. **142**, 43–54 (2016)
20. Wiki, O.: Main Page – OpenStreetMap Wiki (2014)
21. Schultz, M., et al.: Open land cover from OpenStreetMap and remote sensing. Int. J. Appl. Earth Obs. Geoinf. **63**, 206–213 (2017)
22. Lehmler, S., et al.: Usability of open data for smart city applications–evaluation of data, development of application and creation of visual dashboards. In: REAL CORP 2019–IS THIS THE REAL WORLD? Perfect Smart Cities vs. Real Emotional Cities. Proceedings of 24th International Conference on Urban Planning, Regional Development and Information Society (2019)
23. Morley, D.W., Gulliver, J.: A land use regression variable generation, modelling and prediction tool for air pollution exposure assessment. Environ. Model Softw. **105**, 17–23 (2018)