

# Data Driven Air Quality Prediction based on Mobile Measurement

Enes Esatbeyoglu<sup>1,\*</sup>, Andreas Sass<sup>1</sup>, Oliver Cassebaum<sup>1</sup> and Sandro Schulze<sup>2</sup>

<sup>1</sup>Volkswagen AG Group Research, Wolfsburg, Germany

<sup>2</sup>Otto-von-Guericke-University Magdeburg, Germany

**Abstract.** The temporal and spatial prediction of nitrogen dioxide (NO<sub>2</sub>) is very essential because of its harmful impacts on the environment. Its forecasting would help, for example, to regulate predictively the traffic flow. Traditionally, air quality measurements are performed at fixed locations or dedicated mobile laboratories. In this work, we installed a measurement technology in a vehicle and connected it to the vehicle measuring system in order to be able to evaluate further parameters. To this end, we selected one route profile and continuously measured the NO<sub>2</sub> concentration in real-time traffic. We have driven this route profile several times in succession. The rationale of this approach is the idea that several vehicles are equipped with the same measurement technology and drive on the same route profile within the same time. The contribution of this work is to forecast the NO<sub>2</sub> concentration for a given route profile under constant weather conditions based on mobile measurements. To this end, we divided the recorded data into training and test data and investigated five different approaches for forecasting the NO<sub>2</sub> concentration on the respective route profile. Among other aspects, we used cross-validation methods in order to assess the prediction quality. Results show that sliding-window approaches using the averaging of previous rounds are most suitable for predicting NO<sub>2</sub> concentration. Furthermore, our data reveal that the prediction quality is improved when the test data immediately follow the training data.

## 1 Introduction

The NO<sub>2</sub> concentration is one of the harmful pollutants to the environment and public health. One of the causes for the formation of this pollutant is industry and traffic [1, 2]. Usually, official traffic air monitoring stations, located next to the roadside, measure the air quality. The hourly limit value in European cities for NO<sub>2</sub> concentration, which may be exceeded 18 times a year, is 200 µg m<sup>-3</sup> and the average annual limit value is 40 µg m<sup>-3</sup> [3]. In this regard, cities aim at improving the air quality, and thus, protect the human health and avoid financial sanctions.

In addition to classical air quality studies based on fixed locations or mobile laboratories [4, 5], there are scientific studies based on mobile measurements. For instance, Elen et al. and Liu et al. describe a vehicular based mobile approach for measuring fine-grained air quality and other pollutants in real time [6, 7]. Furthermore, bicycles have already been used to measure ultrafine particles, black carbon, and carbon monoxide [8, 9]. The advantage of this measurement method over the stationary measuring station is its proximity to potential emitters in road traffic, the dynamics of the measurement, and greater coverage.

In addition to measuring the current state, predicting air quality is also of particular importance. The development of prediction models helps to provide early warnings to the population and take actions before tolerance limits are exceeded. In this respect, scientific approaches exist to make both, temporal as well as spatial

forecasts. In [10, 11, 12 and 13], for example, attempts on the basis of stationary measurement data were made to predict the NO<sub>2</sub> concentration using various methods in the field of machine learning. Alternatively, different dispersion models based on mathematical or statistical approaches have been applied in order to investigate the spatial forecasting of pollutants [14, 15].

In order to extend the prediction models based on stationary measuring stations or models, we equipped a vehicle with NO<sub>2</sub> measuring technology for our work. The measuring device was connected to the measuring system of the vehicle so that we also recorded specific vehicle parameters from internal (CAN) messages. Afterwards, we selected a route profile in which there is also a stationary measuring station next to the roadside. Subsequently, we drove this route profile 23 times and continuously measured the NO<sub>2</sub> concentration in real-time traffic. After dividing the measured data into training and test data, we investigated the prediction quality of different techniques, for example, using cross-validation. The contributions of this work are:

- We investigate different techniques in order to forecast NO<sub>2</sub> concentration with maximum accuracy for the selected route.
- We propose a lightweight, mobile measuring system that includes both, environmental as well as car-specific parameters, and thus, allows for more precise data collection and a higher accuracy in the prediction phase.
- A comprehensive comparison of the prediction techniques used, based on real-world data.

\* Corresponding author: [enes.esatbeyoglu@volkswagen.de](mailto:enes.esatbeyoglu@volkswagen.de)

## 2 Background

In this section, we describe in more detail the selected measurement technology and measurement methods. Furthermore, we explain the driving route profile. Subsequently, the method of cross-validation is introduced to assess the forecasting approaches.

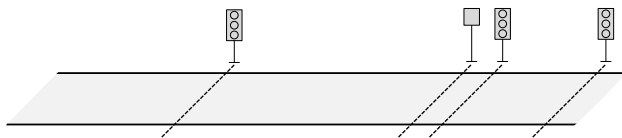
### 2.1. Measurement Technology

We installed the measuring device "NO<sub>2</sub> / NO / NO<sub>x</sub> Monitor Model 405 nm" from "2B Technologies, Inc." on the vehicle. The principle of measurement is a direct measurement of NO<sub>2</sub> by absorbance at 405 nm in the concentration range 0-10,000 ppb with an accuracy of 2 ppb.

This device is designated as a Federal Equivalent Method (FEM) for NO<sub>2</sub> compliance monitoring in the United States but not in Europe [16]. Therefore, the measuring method does not comply with the official European directives and thus is not an officially licensed measuring technique. In comparison, the measurement method of road measuring stations according to the 22nd Ordinance to the Federal Immission Control Act (BImSchV) for NO<sub>2</sub> concentration is based on the chemiluminescence method [3, 17]. However, both methods mentioned above have already been compared with each other, concluding that the data quality of our measuring device used is given [18].

### 2.2. Route Profile

We used the measuring technique described in the previous section in order to measure the NO<sub>2</sub> concentration on a route profile. This route is 680 m long and can be understood as a circular with right turn process. Fig. 1 shows schematically the respective start and end point of a circle as well as at which points traffic lights and the air measuring station are located.



**Fig. 1.** Route profile which was driven on 23 times and consists of three traffic lights and a traffic air monitoring station.

We drove the route profile a total of 23 times in succession in order to make a statement about how the NO<sub>2</sub> concentration changes on the route for following vehicles. The rationale of this procedure is the idea that several vehicles are equipped with the same measurement technology and drive on the same route profile within the same time. External conditions regarding weather and time, which have an impact on the NO<sub>2</sub> concentration [4, 5], can be seen in Table 1.

**Table 1.** External conditions to weather which we assumed to be constant.

Weather and Time Data	
Time	2-3 pm
Temperature	5.5 – 5.7 °C
Precipitation	0 ltr/m <sup>2</sup>
Wind Direction	220 – 240 °
Wind Speed	2 – 2.7 m/s

### 2.3. Cross-Validation

After we have presented the route profile on which the NO<sub>2</sub> concentration was measured, we explain in more detail one of the methods used in this work.

Cross-validation methods are test methods of data analysis that are used, among other aspects, in data mining where the goal is prediction. There are several types of this method, such as simple, stratified or leave-one-out cross-validation [19, 20]. In this work, the process of simple cross-validation is described in more detail and then applied.

In  $k$ -fold cross-validation, the forecast results are evaluated by partitioning the original data set into  $k$  equally sized (length  $m$ ) subsets  $T_1, \dots, T_k$  consisting of  $N$  elements. Furthermore, a distinction is made between a training set and test set.  $k$  passes are started in which the  $i$ -th subset of  $T_i$  is used as the test set and the remaining  $k-1$  subset of  $T_i$  is used as the training set. The total error rate is calculated as an average from the individual error rates of the  $k$  individual runs [19, 20, 21].

## 3 Forecasting Methodology

In this section, we describe in more detail the data preparation. Subsequently, we explain the forecasting approaches, used in this paper, and their application. Finally, we present an estimation for the quality of each forecasting approach.

### 3.1 Data Preparation

As described in the previous section, we connected the measurement device to the vehicle's measuring system. All channels or information stored on the measuring system have different transmission frequencies (asynchronous) because of different priorities on the CAN. This means, for example, that the value for the driving speed is transmitted more frequently in a time unit than the value for the NO<sub>2</sub> concentration. Consequently, all recorded signals must be normalized to the same equidistant form in order to be able to deduce the distance covered from the driving speed.

In this work, we chose a linear interpolation based on the time vector with an increment of 0.2 s. In order to

infer distant-equivalent signals from time-equivalent ones, asynchronous distance-based units were initially extracted from the velocity signal. Subsequently, we used the linear interpolation (based on the distance) with an increment of 1 m in order to calculate signals with an equivalent distance. Afterwards, we created a 23 x 680 matrix. The row size stands for the number of rounds and the column size for the NO<sub>2</sub> concentration recorded or interpolated per 1 m.

In addition, we have taken into account the delay time and the reaction time of our measurement device by shifting the measurement data of the NO<sub>2</sub> concentration by a certain time interval. This is particularly important to determine the location of the measured NO<sub>2</sub> concentration.

### 3.2 Forecasting Approaches

In order to predict the NO<sub>2</sub> concentration along the route as described in Section 2.2, we used different approaches to determine the NO<sub>2</sub> concentration for the subsequent round. Depending on the approach, we have taken the individual rows of all columns as test or training data sets. In Fig. 2, we show the different approaches, in particular, how differ in the training data used.

Approach 1 describes the forecasting of the following round based on the previous round. We have assumed that the next round will be the same as the previous round. We considered the one round as training data and the directly following round as test data. These data were then compared with one another in order to make a statement about the forecasting quality.

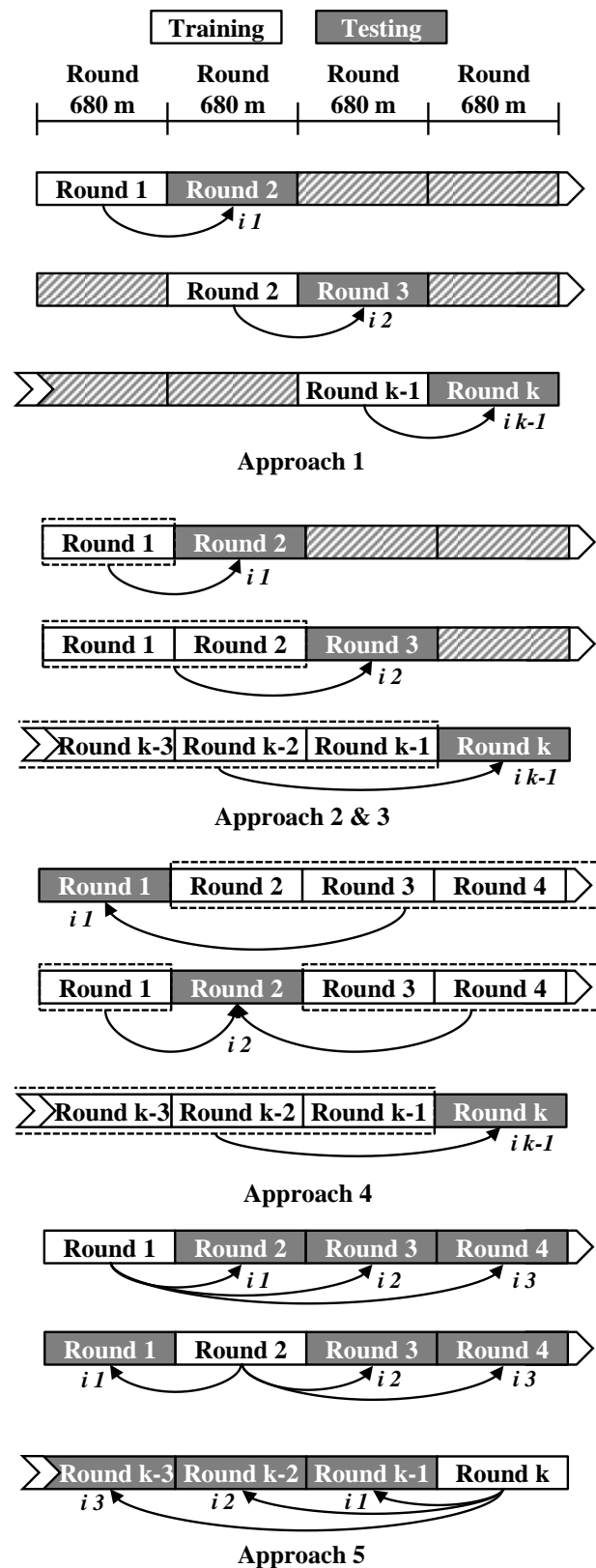
For the approaches 2 & 3 we calculated the mean over the complete distance of previous laps. For approach 2, we have not assumed the number of previous rounds to be constant. Rather, the number of rounds per iterative step increased so that we then calculated the mean value for these rounds (always starting with round 1). Finally, we compared the mean (dashed box over the previous laps) with the round directly after (i.e., the round predicted to be predicted).

Approach 3 differs from approach 2 by the number of rounds over which we have calculated the mean value. Here, we have calculated the mean over the last three, four, five and six rounds directly before the round to be predicted (sliding-window). These data were then compared with one another.

In approach 4 we assumed that the NO<sub>2</sub> concentration of each round can be forecasted by calculating the mean value of all driven rounds except the round to be predicted. Especially here we have used the method of cross-validation from Section 2.3. The round to be predicted can be seen as test data and the mean value of the remaining rounds as training data.

In the last approach, we also applied the method of cross-validation. Compared to the first approach, we compared not only the following rounds but every single round with the other rounds. For example, the first round was assumed as training data and the remaining data as test data. These data were then compared with one

another in order to make a statement about the forecasting quality.



**Fig. 2.** Different approaches to forecast the next round. The dashed lines indicate the rounds over which the mean value is calculated. Approaches 2 and 3 are similar and differ only in the size of the training data set (sliding-window). Cross-validation was used especially for the last two approaches.

### 3.3 Forecasting Quality

For all approaches, presented in the previous subsection, we discuss the forecasting quality in more detail in the following. To this end, we calculated the percentage error per approach between the training and test data set in order to be able to make a statement about the forecast quality. To this end, we have assumed the Root Mean Squared Percentage Error (RMSPE) as a measure for assessing the forecasting quality. It indicates how well test data is adapted to existing training data, or how much a forecast deviates on average from the historical data (i.e., actual observed values). The larger the RMSPE, the greater the deviation from the model. In the literature, the RMSPE is used in many prognosis error evaluations, for example in regression based and statistical methods [22, 23]. The RMSPE is calculated as follows [23]:

$$RMSPE = \left( \frac{\sum_{i=1}^N \left( \frac{T_i - P_i}{T_i} * 100 \right)^2}{N} \right)^{\frac{1}{2}}$$

- N = Number of elements in the column (680)
- T = Training data
- P = Testing data (Predicted data)

The difference between the training and test data set of each column entry (the NO<sub>2</sub> concentration per 1 m on the route profile) is calculated and set in relation to the training data set per entry. Afterwards, the ratio is squared and set in relation to the total data size. Finally, the root is calculated. Using this approach, we calculated the prediction error of a single (predicted) round against other rounds.

Afterwards, we use box-plot diagrams to assess which approach in section 3.2 can be used in order to realize the lowest prediction error (mean). The box-plot diagrams are suitable for representing the distribution and dispersion of the RMSPE. To this end, the error of each approach on mean (+) and median (dashed-line) as well 25th and 75th percentile and the minimum and maximum deviation are displayed in Fig. 3.

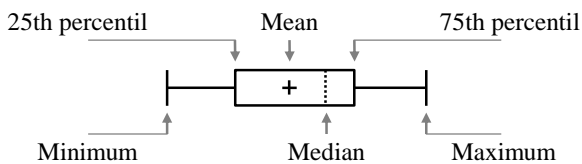


Fig. 3. General structure of the box-plot diagram

### 4 Forecasting Methodology

In this section, we present the results for our five approaches and discuss them afterwards. As an example, we will describe the results and the prediction quality for the first approach in more detail. To this end, we calculated the RMSPE between (i+1)-th round (testing data) and (i)-th round (training data). For this approach, we assumed that the next round corresponds to the

previous round. We show the errors related to this approach in Table 2.

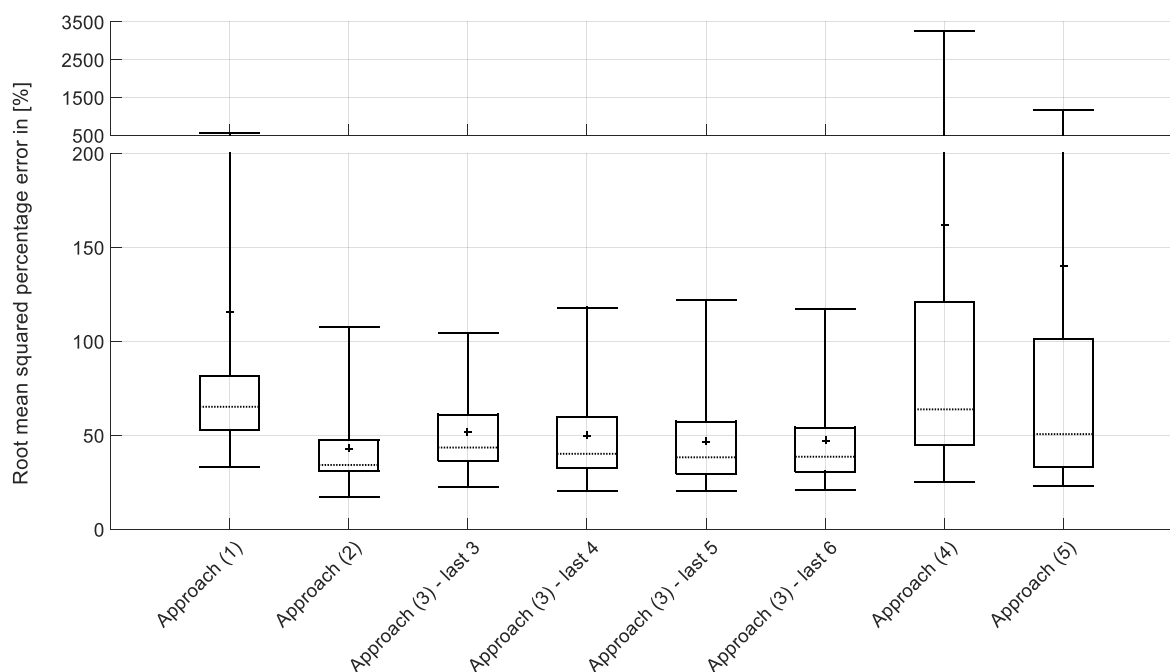
Table 2. The table shows the RMSPE for the first approach, described in Fig. 2

RMSPE based on the last round			
Round	RMSPE	Round	RMSPE
1st → 2nd	64,3 %	12th → 13th	47,6 %
2nd → 3rd	82,5 %	13th → 14th	488,3 %
3rd → 4th	566,1 %	14th → 15th	239,1 %
4th → 5th	60,3 %	15th → 16th	52,5 %
5th → 6th	65,2 %	16th → 17th	63,9 %
6th → 7th	75,4 %	17th → 18th	63,1 %
7th → 8th	78,1 %	18th → 19th	78,1 %
8th → 9th	47,0 %	19th → 20th	33,2 %
11th → 10th	38,5 %	20th → 21th	53,4 %
12th → 11th	151,1 %	21th → 22th	76,1 %
13th → 12th	34,8 %	22th → 23th	67,0 %

The left column indicates the two rounds (observed & predicted) used for prediction. The right column specifies the respective RMSPE. The specified error in percent indicates how much the NO<sub>2</sub> concentration of the following round deviates on average from the previous (actual observed) round.

The table shows that the RMSPE varies widely from round to round. This depends, among other aspects, on the amount of traffic per round. It is also important in which intensity and form a vehicle ahead emits NO<sub>2</sub>, which is then absorbed by the measuring device (exhaust gas plume). We have not classified these as outliers in our database. It can therefore be assumed that RMSPE is only low if the traffic volume is identical to the previous round under constant weather conditions.

Consequently, with this value, no reliable can be made whether the NO<sub>2</sub> concentration for the next round will be higher or lower. Since presenting all tables would not fit into the page limit, we visualize the RMSPE values for all five approaches by means of box-plot diagrams in Fig. 4. Because the third approach is based on sliding window, we took the mean values of the NO<sub>2</sub> concentration of the previous three, four, five and six rounds as training data. These were then compared with the NO<sub>2</sub> concentration of the following round.



**Fig. 4.** Results of all approaches as box-plot diagram. The different approaches are described on the x-axis and the respective RMSPE's on the y-axis.

Subsequently, the RMSPE was also calculated for each window size separately. Due to the largest maximum deviation varying considerably from the smallest one, we have divided the display into two areas with different scaling. The bottom range has an interval of 50 % ending at 200 % and the upper range has an interval of 1000 % starting at 500 %.

First of all, our data reveal that the approaches 1, 4 & 5 have the greatest dispersion. If the maximum errors are considered these results have a RMSPE of more than 500 %. The minimum, median and mean errors of these three approaches are also larger than for approach 2 & 3. Similarly, the NO<sub>2</sub> concentration of individual rounds differs considerably. Hence, we conclude that the result of predicting other rounds based on one round (including cross-validation) is worse than for the other two approaches. Based on our evaluation, we argue that approaches 1, 4 & 5 are, on average, unsuitable for forecasting subsequent rounds under the assumption of unchangeable weather parameters.

The dispersion of the second and third approach are considerably smaller. Hence, the prediction with these methods is suitable on average. However, our data reveal that the basis on which the forecast is based with at least 50% accuracy requires larger historical data series than one. The second method seems to produce lower RMSPE, but a sliding-window method leads to a similar prediction quality.

In the fourth approach, a mean is formed analogous to approaches 2 and 3. However, the crucial difference is the correlation between the training data and the test data directly following the training data. This shows that the mean value computation, based on previous rounds, only leads to a low RMSPE if the test data follows the training data directly.

## 5 Conclusions and Future Work

In this paper we have first described how we measured the NO<sub>2</sub> concentration. Subsequently, we investigated different approaches in the field of data mining, with the help of which the NO<sub>2</sub> concentration of the following rounds can be forecasted. We have used both simple approaches and methods of cross-validation. In order to show the error between the training and test data, we first calculated the RMSPE and then visualized the distribution or dispersion in a box-plot diagram.

We have shown that approaches without previous averaging of the NO<sub>2</sub> concentration over certain rounds generate higher RMSPE than approaches with averaging. We have also shown that averaging is not sufficient for significantly low RMSPE. Rather, a training data set must be generated directly before the test data set. Different traffic conditions and, among other aspects, waste gas plumes result for each round of traffic, so that approaches with no averaging are unsuitable.

The current work can be expanded by equipping several vehicles with suitable NO<sub>2</sub> measuring technology in order to make a comprehensive statement about the quantity and quality of the measurements. Based on this, predictions for route sections can then be derived using the methods described above. Furthermore, the NO<sub>2</sub> concentration should also be measured at different weather conditions and times (especially rush hour) on a selected route in order to take into account the influences of these parameters.

## References

1. W. H. Organization, Monitoring ambient air quality for health impact assessment



2. U. Gehring, A. H. Wijga, M. Brauer, P. Fischer, J. C. de Jongste, M. Kerkhof, M. Oldenwening, H. A. Smit, B. Brunekreef, Traffic-related Air Pollution and the Development of Asthma and Allergies during the first 8 Years of Life, *American Journal of Respiratory and Critical Care Medicine* 181 (6) (2010) 596-603. doi:10.1164/rccm.200906-0858OC.
3. Act on the Prevention of Harmful Effects on the Environment Caused by Air Pollution, Noise, Vibration and Similar Phenomena.
4. M. M. Rahman, B. Yeganeh, S. Clifford, L. D. Knibbs, L. Morawska, Development of a land use regression model for daily NO<sub>2</sub> and NO<sub>x</sub> concentrations in the Brisbane metropolitan area, Australia, *Environmental Modelling & Software* 95 (2017) 168-179. doi:10.1016/j.envsoft.2017.20006.029.
5. K. Bashir Shaban, A. Kadri, E. Rezk, Urban Air Pollution Monitoring System With Forecasting Models, *IEEE Sensors Journal* 16 (8) (2016) 2598-2606. doi:10.1109/JSEN.2016.2514378.
6. B. Elen, J. Peters, M. Poppel, N. Bleux, J. Theunis, M. Reggente, A. Standaert, The Aeroflex: A Bicycle for Mobile Air Quality Measurements, *Sensors* 13 (12) (2012) 221-240. doi:10.3390/s130100221.
7. W. Liu, X. Li, Z. Chen, G. Zeng, T. Leon, J. Liang, G. Huang, Z. Gao, 210 S. Jiao, X. He, M. Lai, Land use regression models coupled with meteorology to model spatial and temporal variability of NO<sub>2</sub> and PM<sub>10</sub> in Changsha, China, *Atmospheric Environment* 116 (2015) 272-280. doi:10.1016/j.atmosenv.2015.06.056.
8. J. Van den Bossche, J. Peters, J. Verwaeren, D. Botteldooren, J. Theunis, B. De Baets, Mobile monitoring for mapping spatial variation in urban air quality: Development and validation of a methodology based on an extensive dataset, *Atmospheric Environment* 105 (2015) 148-161. doi:10.1016/j.atmosenv.2015.01.017.
9. Z. Ross, P. B. English, R. Scalf, R. Gunier, S. Smorodinsky, S. Wall, M. Jerrett, Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses, *Journal of Exposure Science and Environmental Epidemiology* 16 (2006) 106-114. doi:10.1038/sj.jea.7500442.
10. S.I.V. Sousa, F.G. Martins, M.C.M. Alvim-Ferraz, M.C. Pereira, 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling & Software* 22, 97-103. doi: 10.1016/j.envsoft.2005.12.002
11. Y. Rybarczyk, R. Zalakeviciute. Machine learning approach to forecasting urban pollution. *IEEE Ecuador Technical Chapters Meeting (ETCM)*, 2016, 1-6. doi: 10.1109/ETCM.2016.7750810
12. C. Yan, S. Xu, Y. Huang, Y. Huang, Z. Zhang. Two-phase Neural Network Model for Pollution Concentrations Forecasting. *Fifth International Conference on Advanced Cloud and Big Data*, 2017. doi: 10.1109/CBD.2017.73
13. M. M. Dedovic, I. Turkovic, T. Konjic, S. Avdakovic, N. Dautbasic. Forecasting PM<sub>10</sub> concentrations using neural networks and system for improving air quality. *XI International Symposium on Telecommunications (BIHTEL)*, October 24-26, 2016. doi: 10.1109/BIHTEL.2016.7775721
14. B. Owen, H. A. Edmunds, D. J. Carruthers, R. J. Singles, Prediction of total oxides of nitrogen and nitrogen dioxide concentrations in a large urban area using a new generation urban scale dispersion model with integral chemistry model, *Atmospheric Environment* 34 (3) (2000) 397-406.
15. S. Tunlathorntham, S. Thepanondh, Prediction of Ambient Nitrogen Dioxide Concentrations in the Vicinity of Industrial Complex Area, Thailand, *Air, Soil and Water Research* 10 (2017) 1178622117700906.
16. U. G. I. (GPO), J. Orme-Zavaleta, Office of Research and Development; Ambient Air Monitoring Reference and Equivalent Methods: Designation of One New Equivalent Method, *Federal Register* 82 (90).
17. DIN EN 14211 Ambient air - Standard method for the measurement of the concentration of nitrogen dioxide and nitrogen monoxide by chemiluminescence, *Deutsches Institut für Normung eV*, Berlin.
18. S. Gilde, GAW Brief des DWD – Der CAPS-Monitor, ein neues Instrument zur Messung von Stickstoffdioxid in Umgebungslust, [https://www.dwd.de/DE/forschung/atmosphaerenbeob/zusammensetzung\\_atmosphaere/hohenpeissenberg/download/gaw\\_briefe/gaw\\_brief\\_059\\_de\\_pdf.pdf?\\_\\_blob=publicationFile](https://www.dwd.de/DE/forschung/atmosphaerenbeob/zusammensetzung_atmosphaere/hohenpeissenberg/download/gaw_briefe/gaw_brief_059_de_pdf.pdf?__blob=publicationFile)
19. R. Menard, M. Deshaies-Jacques, Evaluation of Analysis by Cross-Validation. Part I: Using Verification Metrics, *Atmosphere* 2018 (9) (2018) 86. doi:10.3390/atmos9030086.
20. P. Refaeilzadeh, L. Tang, H. Liu, Cross-validation (2009) 532-538.
21. T.-T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, *Pattern Recognition* 48 (9) (2015) 2839-2846. doi:10.1016/j.patcog.2015.03.009.
22. Alho, J. and B. Spencer. (2005). *Statistical Demography and Forecasting*. Dordrecht, The Netherlands: Springer. W. Alonso and P. Starr (Eds.). *The Politics of Numbers*. New York: Russell Sage.
23. M. V. Shcherbakov, A. Brebels, N. L. Shcherbakova, A. P. Tyukov, T. A. Janovsky, V. A. Kamaev, A survey of forecast error measures, *World Applied Sciences Journal* 24 (2013) 171-176. doi: 10.5829/idosi.wasj.2013.24.itmies.80032.