Prediction of Air Pollution Levels using Meteorological Inputs through a Neural Network

Loudoun Academy of Science

Darsh Patel

Air pollution is a major problem faced by rapidly developing cities in the modern world. Places like Delhi, India and Shanghai, China regularly exceed the healthy standards set by the World Health Organization. According to the WHO, worldwide air pollution accounts for 4.2 million deaths worldwide, and a staggering 91% of the world’s population lives in places that exceed the WHO’s air quality guidelines. Therefore, it is imminent that a means of reducing air pollution be discovered.

One way to do so is to predict air quality more efficiently and accurately. This presents a variety of benefits such as motivating changes in behavior and public policy and raising awareness on the dangers of air pollution. Another benefit could be a prediction of whether the air is healthy/unhealthy down to a specific location, as cities with varying terrain features have varying amounts of smog. An expanding field of study has been predicting air quality using machine learning methods, which have many benefits over traditional predicting methods including consideration of non-linear factors. There are numerous studies that deal with the prediction of air pollutants through a variety

The first study under consideration deals with predicting air pollution in Hong Kong. Air pollutants in Hong Kong are emitted largely by industry and automobiles. More accurate prediction of the air pollution would help with understanding of air pollution and development of pollution control strategies by Hong Kong officials (Zhang, 2017).

Two methods currently exist for the prediction of air pollutants: deterministic and stochastic. The deterministic method models physical and chemical transportation process in terms of meteorological variables, and it can either do short-term or long-term predictions. Some researchers use these to develop integrated air quality models with source, dispersion, and environmental impacts. However, it is difficult to predict the concentration of air pollutants with high accuracy due to complexity of how different meteorological variables affect each other and air pollutant concentration, as well as multiplicity of sources (Zhang, 2017).

The statistical approach learns from historical data, and its methods include time series analysis, Bayesian filter, Artificial Neural Networks (ANNs), and multiple linear regressions (MLR). However, linear regression models are not ideal due to reactions between air pollutants and influential factors being nonlinear. An ANN has the advantage of incorporating nonlinear relationships between pollutants and meteorological factors but has several drawbacks including poor generalization and time to train the model (Zhang, 2017).

Therefore, in this study the researchers used an extreme learning machine (ELM), which performs better when there is more noise in the data. ELMs are the same thing as traditional feed forward networks, but they have random connections between neurons in each layer. ELMs also do not use backpropagation. Instead, they start with random weights and biases and train the weights based on a single step with the least-squares method of error minimization. They been widely used from biomedical engineering to computer vision and are also used for a variety of purposes including classification, regression, clustering, and feature selection (Zhang, 2017).

Feed-forward neural networks based on back propagation (FFANN-BP) is the conventional and most popular way to predict smog levels, but it requires a “desired output.” It also needs an activation function (such as unit step, linear, sigmoid, sinh, etc.) Drawbacks of FFANN-BPs include them being extremely time consuming (all parameters tuned iteratively) and prone to getting caught in local minimums (Zhang, 2017).

In this study, the inputs included high temperature, low temperature, difference between high and low temperature, average temp, wind speed, wind direction (in radians), relative humidity, and time variables: day of week and month of year. The used 10-fold cross validation, and the average accuracy for 10 iterations recorded. The number of hidden nodes was 20. They used evaluation metrics such as mean absolute error (MAE), root mean square error (RMSE), index of agreement (IA) and R2 value. The results with the highest R2 and lowest RMSE was determined to be the “best” method (Zhang, 2017).

The ELM scored well over linear regression and FFANN-BP for coefficient of determination. It also yielded lowest RMSE and the fastest speed of training. The tests were run again but with a limit value of 50 ug/m3. Based on the data, this value was exceeded 38% of days (Zhang, 2017). In conclusion, they found that ELM performed better in terms of precision, robustness, and generalization, and provided the best performance on indicators such as R2 and RMSE However, there were no significant differences between the prediction accuracies of each model (Zhang, 2017).

A second study modeled PM2.5 (particulate matter) levels in Quito, Ecuador, a city surrounded by mountains on each side that do not allow for dispersion of smog. The atmosphere has a “mixing layer”, where pollutants tend to collect. The depth of this mixing layer depends on solar radiation and temperature. The shallower the mixing depth, the less dilution of emissions.

There are three current methods of predicting smog levels: statistical, chemical transport, and machine learning. The chemical transport method is used most often nowadays but requires an updated source list that is difficult to produce. Also, especially in a city like Quito, topography can complicate chemical transport models. New research has indicated that the ANN is most accurate (Deters, 2017).

The study used three variables: wind speed, wind direction, and precipitation. They used Boosted Trees (BT) and Linear Support Vector Machines (LSVM) along with the traditional neural network to determine any differences between methods. BTs combine simple rules to create classification algorithm, where each misclassified data point gains weight, while a LSVM aims to classify data by separating data into two or more categories using a line (Deters, 2017).

Quito is often under direct sunlight, since it is located on the equator, and is located on a long plateau near a volcano. Due to the complex terrain, wind speed is almost unpredictable. Data from June 2007 to July 2013 was used, and weekend values were removed to minimize discrepancies in the data. To obtain general trends, the data was used to generate convolutional based spatial representations. From this analysis, they found that strong winds resulted in low PM2.5 concentrations, and that winds generally come from a similar direction (Deters, 2017).

Two classes are used (above 15 ug/m3 and below), and a high difference in performance was yielded between the two sites of measurement. A ROC curve was used to evaluate the binary classifier by plotting true negatives vs. true positives. Using three classes yields that values of 10–25 μg/m3 and >25 μg/m3 are not influenced by meteorological parameters. The RMSE graph echoes the same concept, suggesting that prediction of PM2.5 is more reliable for extreme than for moderate climactic conditions. Meteorological inputs were found to accurately describe PM2.5 levels if they are less than 20 μg/m3, while additional factors such as festivals, wildfires, and accidents could explain the PM2.5 levels exceeding 20 μg/m3 (Deters, 2017).

In conclusion, regression analysis showed that prediction is possible for levels less than 20 ug/m3. The accuracy improved in conditions of strong winds and high precipitation. Future research could include a hybrid between machine learning and chemical transport models (Deters, 2017).

And finally, a last article described the development of an ANN-based forecasting system with sensitivity analysis in Auckland, New Zealand. The purpose of this study is to determine effect of different meteorological parameters on pollutant concentrations to construct an ANN model based on pattern recognition. It forecasted NO2 concentrations, one of many particles that constitutes polluted air (Elangasinghe, 2014).

The site in Auckland, New Zealand was chosen due to a complex mixture of traffic, home heating, and industrial emissions. One year of data was chosen for training and two weeks for testing. The researchers used sensitivity analysis, which provided extra knowledge on response of the network to meteorological and emission parameter changes. They also tested three linear models developed on the same data set for a comparison against the neural network. These included one with inputs of time scales, one with time scales, wind speed, and wind direction, and one with time scales and all meteorological inputs. Evaluation metrics used include Mean Bias (MB), Root Mean Square Error (RMSE), coefficient of determination (R2), and index of agreement (IA) (Elangasinghe, 2014).

The researchers found a clear inverse relationship between NO2 concentrations and wind speed, which means as wind speed increased NO2 concentrations decreased. Also, NO2 was linearly related to solar radiation and relative humidity. As for their neural network, the number of hidden layers was 37. They used genetic optimization, which took the longest out of all models tested, but got the most accurate results. From their sensitivity analysis, they found that wind speed, wind direction, and hour of the day showed greatest sensitivity, and that wind speed plays most important role and shows a strong inverse relationship. From their results, they also determined that the model performs worst in spring and best in autumn (Elangasinghe, 2014).

The proposed study involves using a variety of meteorological inputs such as temperature, wind speed, wind direction, relative humidity, and average precipitation to predict smog levels. Smog levels will be measured by PM2.5, PM10, and NO2. The difference between the proposed study and the discussed studies is that the proposed study will predict air quality in Los Angeles, a topographically complex city surrounded by the Santa Monica Mountains and the San Gabriel Mountains, both of which tend to concentrate air pollution in the San Fernando Valley. Air quality data will be obtained from the United States Environmental Protection Agency (EPA), and weather data will be obtained from Weather Underground for a station approximately 1 mile away from the pollution meter. A study by USC found that the most polluted area of Los Angeles is Chinatown (Mackovich, 2018). Therefore, the pollution meter closest to Chinatown was chosen, at 1630 North Main Street, Los Angeles, CA 90012.

Data will be organized in a CSV file, and Excel and Pandas will be used to clean up the CSV files so that they can be inputted into the neural network. Previous research pairs meteorological data and air pollution data on the basis of time to apply machine learning methods (Zhu, 2018). The proposed research will be using a similar approach, the only difference being that measurements are recorded each day, 365 days of the year. Jupyter Notebook will be used as the environment to run the neural network. TensorFlow is an open-source library from Google that is used for machine learning applications such as neural networks. TensorFlow will be used to make the neural network.

EPA and Weather Underground data from LA

Run SVM/NN/LinReg w/ threshold level

Clean data, by removing null values and formatting correctly

Visualize data to determine patterns

Select inputs (weather data) and separate data into training and testing sets

Build linear regression and neural network using TensorFlow

Compare metrics and draw conclusions

Run linear regression and neural network

Matplotlib, Seaborn are relevant libraries

Sample data table shown below

ROC Curve/confusion matrix for binary classification

RMSE, r2, index of agreement

Data from other countries will also be used

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Index | Avg. PM 2.5 (µg/m³) | Temp (F) | Wind Speed (mph) | Wind direction (rad) | Cloud Cover (%) | Atmospheric Pressure (mb) | Precipitation (in) |
| 0 |  |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |
| 5 |  |  |  |  |  |  |  |
| 6 |  |  |  |  |  |  |  |
| 7 |  |  |  |  |  |  |  |
| 8 |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |

As indicated in the flow chart, a future investigation could be a binary classification with threshold to predict if air quality is healthy/unhealthy. When a binary classification is used, the neural network would output a confusion matrix, which displays false positives, false negatives, true positives, and true negatives in a table format. It would also output a ROC (Receiver Operating Characteristics) curve which shows the true positive rate vs. false positive rate with different thresholds. In order to evaluate the binary classification, the area under the ROC curves for traditional neural networks vs. support vector machines and other algorithms would be compared. Additionally, pollution and weather data from different regions of the world could be used if time permits.

The prediction of air pollution using machine learning is a popular field of research and serves a variety of purposes including influencing public policy and raising awareness about pollution hazards. The accurate prediction of air quality will help those sensitive to pollutants and public officials in rapidly developing cities to best combat the important problem of air pollution.

**References**

Cambria, Erik, et al. “Extreme learning machines [trends & controversies].” *IEEE Intelligent Systems* 28.6 (2013): 30-59.

Cortes, C & Vladimir, V. “Support-vector networks.” *Machine learning 20.3* (1995): 273-297.

Deters, J. K., Zalakeviciute, R., Gonzales, M., & Rybarczyk, Y. (2017) Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. *Journal of Electrical and Computer Engineering*. https://doi.org/10.1155/2017/5106045

Elangasinghe, M. A., Singhal, N., Dirks, K. N., & Salmond, J. A. (2017) Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis. *Atmospheric Pollution Research*. https://doi.org/10.5094/APR.2014.079

Mackovich, R (2018). “USC study calculates air pollution levels in every L.A. neighborhood.” *USC Dornslife*: (July 2018).

Zhang, J. & Ding, W. (2017). Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of Hong Kong. *International Journal of Environmental Research and Public Health*. <https://doi.org/10.3390/ijerph14020114>

Zhu, D., Cai, C., Yang, T., Zhou, X. (2018). A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. *Big Data and Cognitive Computing* (2018): 2, 5. https://doi.org/10.3390/bdcc2010005