Neural network model for the prediction of PM2.5 daily concentrations in Cracow

Karolina Alicja Sala

Abstract— The aim of study is to develop artificial neural networks for the air quality prediction, based on pollutants observations for the 2017 year, in Cracow, Poland. Briefly overviewed the stages of development of the neural network models and description of the features. Several parameters such as PM1, PM2.5, PM10, temperature, air pressure and humidity are considered in this study. LSTM neural network models have been found to provide sufficient reliable predictions, which means they are an effective tool for analyzing and predicting air pollution. The performance of the developed model was assessed through a measures.

Index Terms— Air Pollution, Long Short-Term Memory, Neural Network, Prediction

1 Introduction

Air quality has recently become a serious environmental problem in many southern Polish cities. Standards defined by European Union and Polish law are exceeded many times. The disturbing results of studies conducted by the WHO makes a lot of people are asking about the reason for the poor condition of the air in the southern regions of the country [1]. Pollutants come from various sources. Sulfur oxides are mainly produced by cars, while the emission of nitrogen oxides is responsible for both vehicles and the energy industry. In the case of nitrogen dioxide contamination, the problem occurs in large cities, with heavy transport arteries. In the air there are also polycyclic aromatic hydrocarbons, which are from dust produced as a result of combustion of fossil fuels and wood. Carbon monoxide is produced by the combustion of fossil fuels with limited oxygen availability. In the air there are also dusts visible with the naked eye. The biggest problem is in the case of polycyclic aromatic hydrocarbons and particulate matter PM10, consisting of particles with a diameter below 10 micrometres, as well as a finer fraction of dust, PM 2.5, consisting of particles with a diameter below 2.5 microns.

This issue affects environmental degradation and human health. Nitrogen oxides contribute to the development of bronchial and circulatory diseases, while sulfur oxides have a negative effect on lung function. Carbon monoxide affects the work of the heart and brain. Airborne dusts can cause throat and respiratory diseases as well as allergies. Particle size of aromatic hydrocarbons is critical in determining the particle deposition location in the human respiratory system [2]. Referring to particles with a diameter less than or equal to 2.5 m, has been an increasing concern, as these can be deposited into the lung gas-exchange region, the alveoli [3]. Therefore, air quality forecasting can have a significant impact on the steps taken by the government to reduce this issue.

2 Relevant Work

Forecasting the concentrations of air pollutants represents a challenging errand due to the complexity of the physical and chemical processed involved. The majority common forecasting approaches are numerical models and statistical models. Numerical models need knowledge of pollution sources, the chemical composition of the exhaust gases, and the physical processes in the atmospheric boundary layer [4]. However, statistical models have been applied on the basis of meteorological data. Mostly require a huge amount of measurement data under a large multiplicity of atmospheric conditions. For several years, lots of researchers used these methods for predicting concentrations of air pollutants.

There have been many applications of neural networks since the 1990s, and researchers have obtained practically good results. Artificial neural networks structures have been developed to make better forecasts. Models based on artificial neural networks are based on the study of processes occurring in the human brain. During the process of learning neural network can reveal the complex relationship between the data input and output. After completing the training, the neural network is able to predict the future values of the given sequence based on the values given previously and the existing variables. In the paper Xiao Fen et al. (2015) an artificial neural network was used to predict daily average PM2,5 concentrations two days earlier. The model was developed from 13 different air pollutants monitoring stations in China. Meteorological data was used as inputs to a multi-layer perceptron (MLP) type of backward propagation neural network. It turned out that the model created could be an effective tool to improve PM 2.5 prediction accuracy [5]. Junxiang Fan et al. (2017) proposed a spatiotemporal prediction framework based on missing value processing algorithms and deep recurrent neural network (DRNN). Evaluated and analysed three different missing value fixing algorithms. It turned out that the DRNN framework handles this issue best in comparison to the other tested algorithms, which were deep feed forward neural networks and gradient boosting decision trees [6]. In the above experiment, the tested algorithms were incorporated into deep neural network that consists of LSTM (Long Short-term Memory). This technique was also proposed by Xiang Li et al. (2017). Experiments performed using the spatiotemporal deep learning (STDL) model, the time delay neural network (TDNN) model, the autoregressive moving average (ARMA)

- *Karolina Alicja Sala is currently pursuing masters degree program in computer science in University of Warmia and Mazury in Olsztyn, Poland, E-mail: karolinaalicjasala@gmail.com*

model, the support vector regression (SVR) model, and the traditional LSTM NN model. The LSTME model proved to be better compared to the statistics-based models [7]. Brief review of existing researches illustrated huge potential of long short-term memory neural network purpose in air quality prediction issues.In this application, examined RNN approach for air quality forecast based on data generated by the sensor network located in Cracow. This paper aims to extend a special RNN architecture referred to a long short-term memory neural network (LSTM NN).

3 Data Collection and Preprocessing

3.1 Data Collection

The study area is one of the most popular Polish city, Cracow. Used dataset consists air quality data, includes meteorological parameters and pollutant concentrations matter PM1, PM2.5 and PM10, temperature, air pressure and humidity from 2017, generated by network of 56 sensors located in Cracow. Each had its own location (6 of them where replaced during this time period and have almost the same latitude and longitude). Downloaded pollutant data starting from 1th Jan 2017, till 31th Dec 2017. The data resolution is 1 hour. Dataset contained 670246 records for each station. Due to the significant amount of missing data for many sensors, four of them were selected for this experiment. Two factors were taken into account when choosing the sensors. Selected sensors with the largest amount of recorded data, and which were located at a considerable distance from each other to reduce the interference of distinct geographic areas on the monitoring meteorological data.

3.2 Data Collection and Preprocessing

In this first part, we'll try to get a grasp of how polluted was the air all over the Cracow during 2017. The descriptive statistics of air pollution for the dataset are shown in Table 1. The values (mean, standard deviation, minimum, maximum) are based on the average daily values in each selected monitoring station.

Krakow is located in the basin (valley) of the Vistula River, due to which natural conditions of ventilation are limited. On three sides it is surrounded by high elevations of terrain, that causes limiting the movements of air masses. About 30% of days are windless, for the next 30-40% of the time the wind does not exceed 2 m/s. Fig. 2 shows the relationships between the hourly concentration of PM2.5 at Cracow in 2010 and the variables temperature taken into account. It has been observed that the temperature has a significant affect on PM25 levels. The temperature and concentration of PM25 are directly correlated. High temperature caused high potential for air pollution, whereas the low temperature caused low air pollution potential. Observation puts forward the hypothesis that the increase in pollution is related to the decrease of temperature. Temperature, consistent with summer/winter cycles, shows an increasing trend in the first half of the year and a decreasing

one in the second half. Humidity is one of the atmospheric elements that absorbs solar radiation, preventing it from inter

TABLE 1
DESCRIPTIVE ANALYSIS OF THE DATASET

Feature	Station	Mean	Std	Min	Max
PM2.5	212	25.042283	28.011725	0	221
	214	32.057186	35.591432	0	272
	220	31.737531	35.785010	0	286
	226	30.49050	37.38188	0	304
PM10	212	40.775306	41.698919	0	314
	214	51.663665	51.969130	0	425
	220	50.903866	52.411169	0	451
	226	48.214244	54.388198	0	448
PM1	212	25.227373	24.439209	0	189
	214	32.408724	32.566478	0	256
	220	32.181287	32.433222	0	245
	226	30.379226	31.784032	0	239
Temperature	212	8.994176	8.332834	-15	35
	214	8.542042	8.350891	-15	33
	220	9.085504	8.175138	-14	33
	226	8.635272	8.286543	-16	33
Humidity	212	68.926724	20.428986	2	121
	214	74.978947	21.952261	17	145
	220	72.735600	21.226801	16	131
	226	69.997901	19.193049	9	120
Pressure	212	101491.706542	736.828704	98295	103038
	214	101804.586207	794.818701	98468	104176
	220	101649.274953	766.953423	98325	103843
	226	101632.577625	742.311130	98381	103503

acting with pollutants. Therefore, the humidity level is also a significant factor. A pressure increase was also observed. The overview of the datasets seeks to set up the connection between the hourly mean concentration of PM2.5 and explanatory variables. The average off all pollutions far exceeds accepted standards. Therefore, this variable has been assumed as the output data in this research.

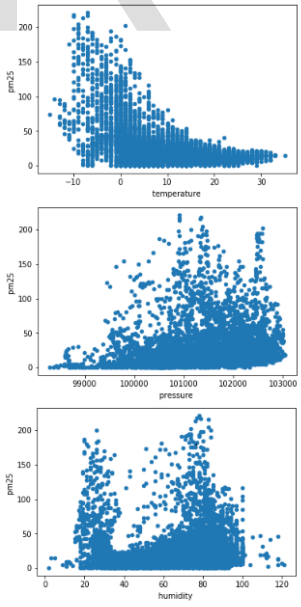

Fig. 1. Scatter plots for the set of hourly observations recorded in 2017, the relationships between the PM2.5 concentration and the following features: temperature, humidity, pressure.

2.3 Data Preprocessing

The following features were used as the input data: PM1,

PM10, temperature, pressure and humidity. Data have been merged hourly in chronological order. Many data were missing. In case of a hourly observations, the missing data might be filled up. Used a non-parametric algorithm called k-nearest-neighbors (KNN) to replace missing values. Nearest Neighbor (NN) assignment algorithms are effective methods to supplement missing data, where each value is missing some records is replaced by the value obtained from related case units in the whole set of records [8]. A filled set of data was were preprocessed and transformed into a time series so that it could be used for a supervised learning problem. All features were rescaled normalized and standardized before giving as input to the network. The input data scaled between 0 and 1 because LSTM model works best on values in 0-1 range. Data normalization helps to remove noisy data, creating a more homogeneous distribution, in order to get better prediction performance.

4 METHODS

In this paper, used a specific RNN architecture, long short-term memory (LSMT), to explore time series pollution in Cracow. This section details the modelling approaches. To avoid the raising of the air pollution under such conditions, it is crucial to predict the content of pollutants in air taking into account the meteorological conditions. The observing statistics characterized by the multidimensional as well as multivariate properties. This way make the environmental forecast challenging. Conducted a comprehensive examination of pollutant concentration as well as meteorological factors through correlation analysis. The analysis was aimed at determining the degree of impact of each meteorological factor on pollutants, and thus on the selection of factors with the greatest impact on pollution. Then choosing the input parameters of the neural network.

The data were partitioned into training and testing sets; 70 percent of the data as the training set, and the remaining 30 percent was used as the test set. After some trial and error, the neural network consisting of one hidden layer of 16 neurons: the structure 32-16-1. The model trained using Adam optimizer with three metrics to evaluate the models accuracy, including Root Mean Squared Error (RMSE), Mean absolute error (MAE), Correlation coefficient (R).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(yi - \bar{y})^2}}$$

5 RESULTS

To get the results in a most objective way, repeated the experiments of training and testing with different activation functions. Applied the early stopping method to avoid overfitting in training, which guarantees the similar results in training and testing. Training during 40 epochs with the patience option of 10 and set the batch size to 72, the dropout rate to 0.1. The validation error used to perform "early stopping" by choosing the number of epochs where its mean error is minimum. The experiment results for four selected stations are presented below in Table 2. The best prediction is achieved for 2th day at station 226 and for 1th day at station 220.

TABLE 2
EXPERIMENT RESULTS

Station	Forecast measure	RSME	MAE	Correlation coefficient
212	+1day	19.050	3.749	0.506
	+2 days	18.447	3.861	0.472
214	+1day	21.457	4.105	0.501
	+2 days	16.120	3.508	0.676
220	**+1day**	**16.421**	**3.451**	**0.705**
	+2 days	18.909	3.860	0.550
226	+1day	18.452	3.645	0.436
	+2 days	**12.509**	**3.039**	**0.695**

Figures 2 and 3 presents the comparisons of predicted concentrations and observed concentrations of on data for PM2.5 concentrations for stations 220 and 226 (with the best predictions), respectively. A regression plots shows the correlation between the actual and the predicted result based on the value of the correlation coefficient. Performance of summer model gave better results than winter season.

For this project, used Python packages, including NumPy to implement model, scikit-learn to rescale data with MinMaxScaler, Matplotlib to do visualization and Keras to build the neural network

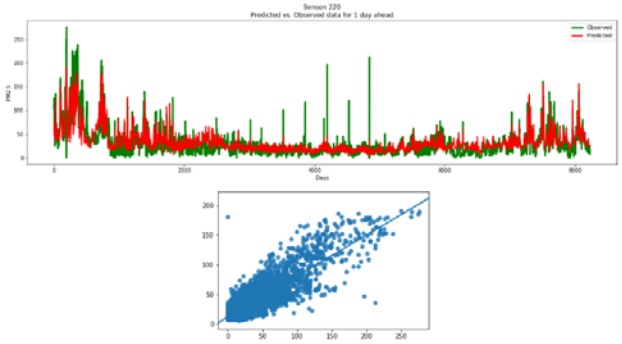

Fig. 2. Predicted and observed concentrations of on data for PM2.5 concentrations for station 220.

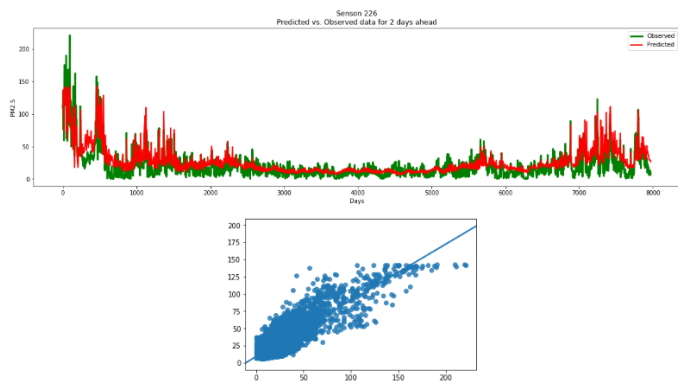

Fig. 3. Predicted and observed concentrations of on data for PM2.5 concentrations for station 226.

6 Conclusion

In this paper, based on long short-term memory neural network, a model of forecasting daily PM2.5 concentrations in Cracow was developed. This research explored the association between the concentration along with meteorological conditions. The relationship has been modelled to explore PM2.5 hourly concentration in terms of influence weather conditions. It turned out that temperature was the most determinant factors affecting air pollution potential compared to other meteorological factors. At the next stage of this scientific research, was to developed a neural network model for short-term prediction of content of air pollutants. The aim of the experiments is to examine the effectiveness of the LSTM model for the prediction of PM2.5 concentrations two days in advance. During modeling of the neural network tested the number of input neurons and the number of hidden layers. Model for the forecasting of the air pollution level described in this paper is sufficiently effective for short-term data.

Over the last few years, artificial neural network have been applied to many environmental engineering problems. This severe PM2.5 pollution problem requires long-term policy as well as effective air quality prediction models. The right models can help to alert against exceedances of the legal concentration.

Acknowledgment

I gratefully acknowledge to my thesis supervisor for valuable guidance.

References

[1] "Air quality in Europe — 2017 report", https://www.eea.europa.eu/publications/air-quality-in-europe-2017

[2] Hinds, W.C. Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles; John Wiley & Sons: Hoboken, NJ, USA, 2012.

[3] Forecasting Urban Air Quality via a Back-Propagation Neural Network and a Selection Sample Rule Yonghong Liu 1,*, Qianru Zhu 2, Dawen Yao 1 and Weijia Xu

[4] Forecasting Urban Air Quality via a Back-Propagation Neural Network and a Selection Sample Rule Yonghong Liu 1,*, Qianru Zhu 2, Dawen Yao 1 and Weijia Xu

[5] Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation, XiaoFeng, Qi Li, Yajie Zhu, Junxiong Hou, Lingyan Jin, Jingjie Wang, Atmospheric Environment Volume 107, April 2015, Pages 118-128

[6] A Spatiotemporal Prediction Framework for Air Pollution Based on Deep RNN Junxiang Fan, Qi Li, Junxiong Hou, Xiao Feng, Hamed Karimian, Shaofu Lin, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume IV-4/W2, 2017 2nd International Symposium on Spatiotemporal Computing 2017, 7–9 August, Cambridge, USA

[7] Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation Article in Environmental Pollution, September 2017

[8] Nearest neighbor imputation algorithms: a critical evaluation, Lorenzo Beretta, Alessandro Santaniello, BMC Med Inform Decis Mak. 2016; 16(Suppl 3): 74.