Method for Environmental Monitoring in the Incomplete Data Conditions

Nikita Tursukov[a], Ilya Viksnin[b], Iuliia Kim[c] and Evgenii Neverov[d]

ITMO University, Saint-Petersburg, Russia

Abstract

In this paper, we propose a method for analyzing and processing incomplete data obtained in the environmental monitoring process. Incomplete and inaccurate data often occur during the operation of environmental monitoring sensors. As a result, these data contribute to the deterioration of the environmental pollution forecast. In the developed method, data is processed, analyzed, and then a model for predicting environmental pollution is generated. This approach is effective for applying to incorrect data, as it increases the accuracy of further forecasts. In this paper, we analyze various approaches to the prediction, and implement the appropriate method implemented using neural networks mechanisms.

Keywords

Neural networks, environmental pollution, data forecasting

1. Introduction

With the development of industrial enterprises production capacities, the pollutants concentrations detection issue increases. In order to reduce the environmental risks, enterprises invest in early warning systems. These systems, involve predicting the values of certain substances concentrations at potentially dangerous objects. When the number of sensors collecting information on the environmental condition increases, the issue of predicting the values when data is incomplete arises. Due to the partial lack of information collected by the sensors, it is impossible to accurately understand whether the local environmental situation is safe for the ecosystem. At the same time, it is important to accurately determine the concentration of potentially dangerous substances at critical infrastructure facilities, and not to confuse them with other substances located within a certain area.

In this paper, we propose a method that allows to analyze incomplete data on the environmental condition, thereby increasing the accuracy of further forecasts. We start with the subject area overview, in the next step a description of the approach is provided, than an empirical study using real environmental monitoring data is conducted and the results obtained are described.

Proceedings of the 12th Majorov International Conference on Software Engineering and Computer Systems, December 10–11, 2020, Online & Saint Petersburg, Russia

stepingnik@gmail.com (N. Tursukov); wixnin@mail.ru (I. Viksnin); yulia1344@gmail.com (I. Kim); datnever@ya.ru (E. Neverov)

0000-0003-3848-1981 (N. Tursukov); 0000-0002-3071-6937 (I. Viksnin); 0000-0002-6951-1875 (I. Kim); 0000-0003-0733-1294 (E. Neverov)

2. Related Work

In industrial facilities, it is crucial to accurately determine the concentration of potentially dangerous substances and not confuse it with others located within the same enterprise area. In other words, it is necessary to clearly distinguish dangerous substances from harmless ones. Methods that use machine learning to analyze environmental parameters and provide a concentration forecast for unreliable and incomplete data were proposed. There are classic machine learning tasks that are usually applied for critical infrastructure facilities monitoring:

- clustering - determining how harmless a substance is, as well as to specify the release source location;
- classification - determining the concentration increase possibility.

An approach to assessing the environmental situation of various natural resources using machine learning methods was demonstrated in [1].The article [2] predicts the level of the territory contamination based on data obtained from several monitoring stations and transmitted via the Internet of Things. For example, a classifier based on Bayesian networks was developed to assess the probability of air pollution by PM2.5 particles. In [3] special attention was paid to the air monitoring system in order to predict the appearance of pollutants based on retrospective data. To perform this, the researchers tested three machine learning algorithms that predicted an increase in the concentration of ground-level ozone, nitrogen dioxide, and sulfur dioxide.

In most of the considered machine learning methods, classification is used to determine whether the situation is critical. For instance, many projects create alarm systems that generate a warning signal in case of detecting the state that is not regulated by the system [4]. Based on the collected data, the model is trained, and the concentration thresholds are determined. If such thresholds are exceeded, the alarm is activated.

Most studies involve detecting critical situations on an object by performing a classification task. Those methods use retrospective data for long-term forecasting, and do not consider incomplete data that may prevent the detection of increased pollutant concentrations [5]. At the same time, machine learning techniques are being increasingly used for detecting the a contaminant appearance.

The developed method involves the use of a neural network that eliminates the incompleteness of the data. Further, using machine learning methods, a more accurate forecast of the concentration of pollutants is made.

3. Materials and Methods

To solve the mentioned problem, we propose to use regression models and neural networks that allow analyzing time series containing information on the pollutant concentration level in the environment, and other factors that may potentially affect its content. The regression allows to analyze the time scale and allows to obtain approximate values for the pollutants concentration. At the same time, the use of neural networks is gaining momentum, since they can both classify the danger of a pollutant, and generate forecasts, considering the sensors' location and the information collected by them.

In contrast to the back propagation neural network, which is standard for solving prediction problems, deep neural networks is considered for predicting data when processing long time intervals [6]. Such networks form a directed sequence between elements, which allows to process a series of events over time, and to link previous information to the current task.

Software data analysis implementation is performed using the Python 3.7 and R programming languages.

4. The Environmental Monitoring Method

The environmental monitoring method represents the order of actions and operations to be performed with the input data. Input data, in general, are parameters obtained from the sensors that collect data the environmental condition. Data is taken for a period of time that is determined by the operator.

As a result of data analysis, a forecast of the pollutant values for the time period n is obtained. The forecast is both numerical concentration indicators and a graph that visualizes retrospective data and data that is adjusted by the model.

Initial data processing involves analyzing the data obtained in order to identify parameters that affect the concentration indicators. Historical data is checked for correctness, by escaping of abnormal data jumps, in order to more accurate further indicators prediction. A final data set is generated, and predictors are selected-indicators that can affect the final predicted concentration value of the predictor substance.

In case of large variations in indicators, the collected time series data can be normalized for more accurate analysis. Standard fields of the generated predictors and responses data set for air analysis is described below:

- Date Time – date and time;
- WSW - wind speed (m/s);
- WDW - wind direction (degrees);
- Sigma – standard deviation of wind direction (degrees);
- Ambient Temp – temperature (degrees Celsius);
- Press - the atmospheric pressure (the atmosphere);
- Amb RH - relative humidity (%);
- NO - concentration of nitric oxide II (ppb);
- NO2 - concentration of nitric oxide IV (ppb);
- NOx - concentration of other nitrogen oxides (ppb);
- SO2 - concentration of sulfur oxide IV (ppb);
- CO-concentration of carbon monoxide (ppm);
- O3-ozone concentration (ppb-billionth part);
- PM10-class 10 ultrafine particle concentration (mcg/m3) ;
- PM2.5-concentration of ultrafine particles of class 2.5 (mcg/m3).

Regression analysis is performed by constructing linear and logistic regression models, described by the expression (1).

$$y{=}g(\mathrm{b}_0{+}\sum(\mathrm{b}_\mathrm{i}\mathrm{x}_\mathrm{i})+\varepsilon)\ , \tag{1}$$

where y is a continuous dependent variable; $\mathrm{b_0}$ is a free term of line assessment; $\mathrm{b_i}$ is an angular regression coefficient; $\mathrm{x_i}$ - factors continuous model, g is a sigmoid function for implementing a logistic regression model.

The autoregressive model, in turn, can also be supplemented with logistic regression, and is described by (2).

$$x_{\mathrm{t}}{=}\mathrm{b_0}{+}\sum\left(\mathrm{b_i}\mathrm{x_{t-i}}\right)+\varepsilon_{\mathrm{t}}\ , \tag{2}$$

where x_{t} is the series value at time t; $\mathrm{b_0}$-free term of line assessment; $\mathrm{b_i}$-angular regression coefficient; $\mathrm{x_{t-i}}$ - value of time series at time $t-1$.

To evaluate the constructed models, the following metrics were used:

- Mean Absolute Error (MAE);
- Mean Squared Error (MSE);
- Root Mean Squared Error (RMSE);

These metrics calculation is represented by (3)-(5).

$$MAE{=}\frac{1}{n}\sum_{i=1}^{n}\left|y_i{-}\hat{y}_i\right|\ , \tag{3}$$

$$MSE{=}\frac{1}{n}\sum_{i=1}^{n}\left(y_i{-}\hat{y}_i\right)^2, \tag{4}$$

$$RMSE{=}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i{-}\hat{y}_i\right)^2}, \tag{5}$$

where $\mathrm{y_i}$ is the predicted value of the I-th ultrafine particle concentration indicator; $\widehat{\mathrm{y}}_{\mathrm{i}}$ is the real value of the i-th ultrafine particle concentration indicator.

As a result, a timeline is formed with the results of the regression forecast, as well as the necessary predictors that affect the concentration. Further data analysis and prediction is performed using the recurrent neural network Long-short term memory (LSTM). LSTM is able to identify significant information when processing sufficiently long time intervals and sequences [7]. This is most effective for working with incomplete data in order to restore and include it in a further forecasting task.

The operation of a recurrent neural network is described by (6).

$$h_{\mathrm{t}}{=}\mathrm{f_w}\left(h_{\mathrm{t-1}},\mathrm{x_t}\right)\ , \tag{6}$$

where h_{t} is the new state that the data processing unit outputs; $\mathrm{f_w}$ is the processing function with parameters w; $h_{\mathrm{t-1}}$ is the state obtained from the previous step; $\mathrm{x_t}$ is the incoming data.

As a result of constructing recurrent neural network LSTM model, a graph is generated that displays the retrospective pollutant indicators and the predicted ones. The correctness of the model's operation is evaluated using evaluation metrics, such as: MAE, MSE, RMSE.

5. Empirical Study

To conduct an empirical study, we analyzed data from open sources on the environmental condition. Data collected by the Stoke Hills station in Darwin, Australia, was selected for the present study. According to open data of the Northern territory of Australia environmental protection office, excess of the PM10 and PM2.5 particles number is observed in the air at this site.

The choice of data depended on the location of the monitoring station. The data used in the experiment were collected near the coal transportation station. This allowed to record a large number of concentration spikes in the test data set, as well as data losses due to sensor failures.

Data collected by the station include meteorological: wind direction and speed, temperature, pressure and humidity, and the concentration of particles (PM10 and PM2.5). Data is collected every hour. For the sample, we took data for a year (∽9000 indicators). An example of a concentration display graph is shown in Figure 1.

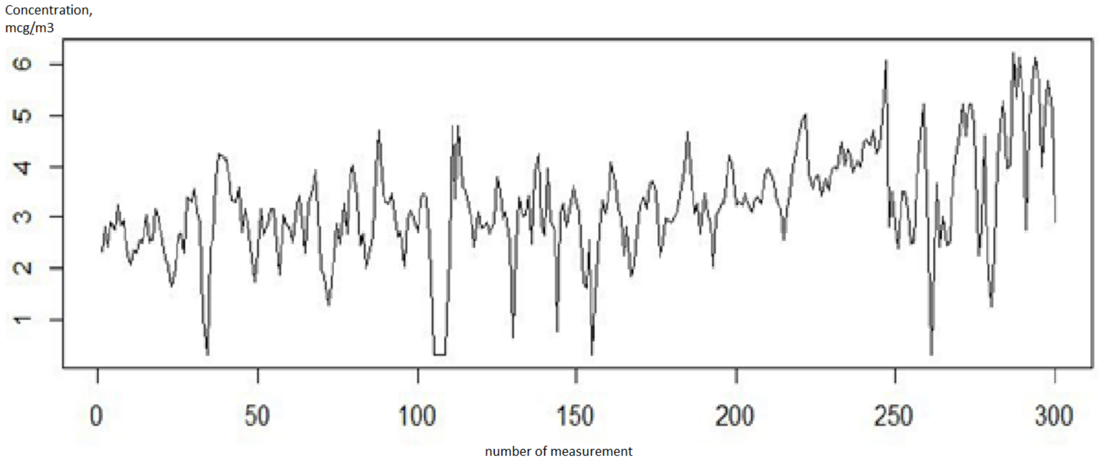

Figure 1: PM10 concentration.

6. Results

Figure 2 shows the retrospective data collected for PM10 in the air using a timeline. At the same time, there is a gap in the collected data that needs to be filled with data that is close to real values in order to make further predictions more accurate.

To solve the prediction problem, we use a model built by the LSTM neural network. For instance, Figure 3 graphically shows the results of incomplete data recovery, as well as its further prediction.

A regression analysis was performed, during which the response and predictors were transformed. Some of the results of the regression analysis are shown in Table 1.

As a result of analyzing the data set used in the experiment, it was found that the combination

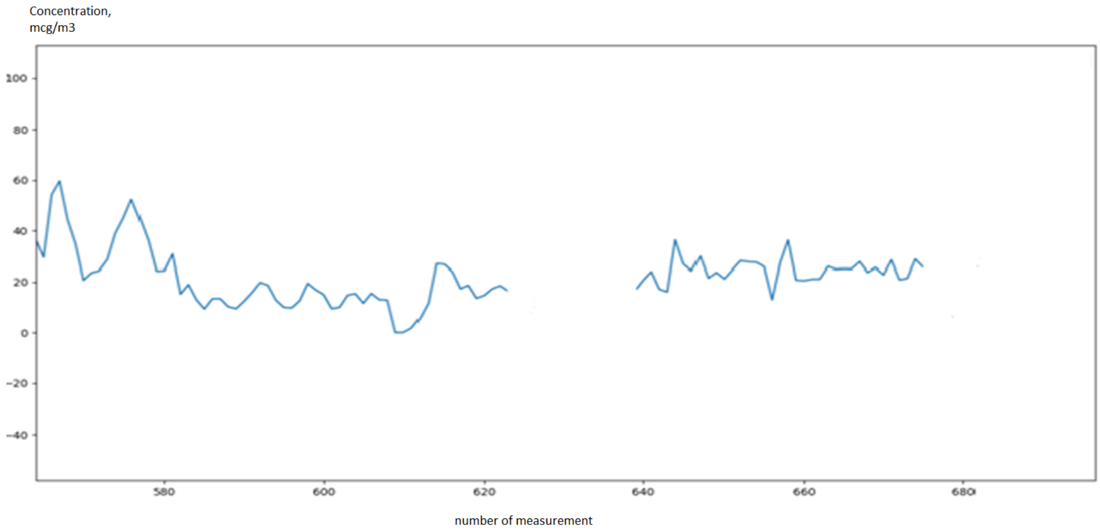

Figure 2: PM10 concentration.

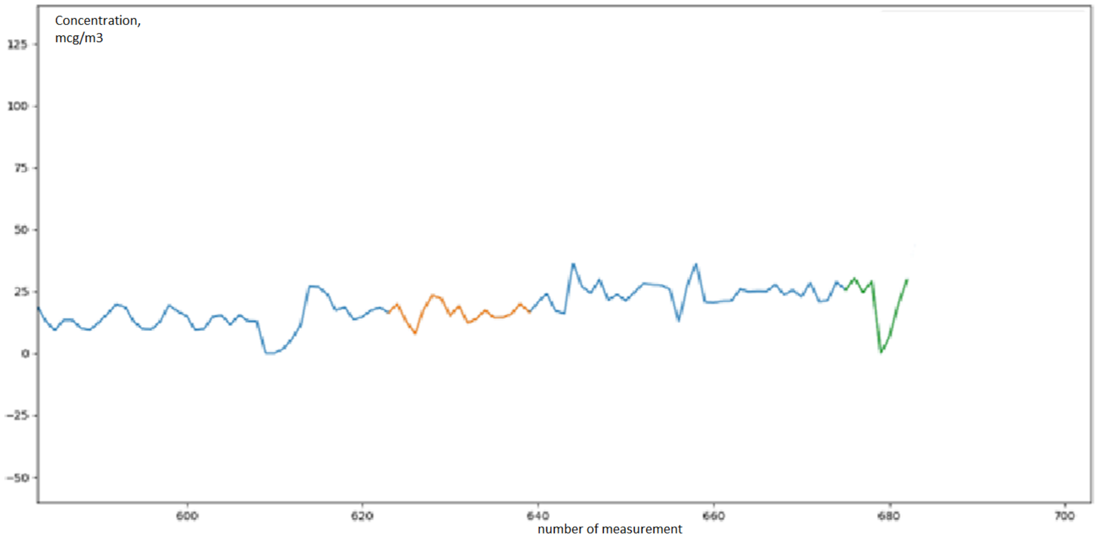

Figure 3: Recovered data.

of predictors describing temperature and humidity positively affects the determination adjusted coefficient value, which was used for data processing.In addition to temperature and humidity, the dependence of the concentration of substances on the seasons was revealed. This allows to more effectively use the LSTM network to restore data.

Thus, using metric estimates, the most successful sets of input data were selected, including predictors necessary for forecasting. Further evaluation is performed after the implementation

Table 1
Results of regression analysis.

Regression	Predictors and response transformations	The value of metrics		
		R-squared	RMSE	p-val
Autoregression	PM10(t-1),PM10(t-2),T(t-1),RH(t- 1)	0.686	8.275	<2.2e-16
Autoregression	PM10(t-1),PM10(t- 2),PM10(t-3)	0.66	8.6	<2.2e-16
Linear	PM10(t-2)	0.1572	14.2	<2.2e-16
Linear	RH, T, direction,log(PM10+2)	0.137	0.6	<2.2e-16

of the prediction model via neural networks. The estimation is performed both by analyzing the results metrics and using graphs, comparing retrospective and predicted data.

If the problem of incomplete data occurs, when factors affecting the polluting parameter cannot be considered, the predicted data should be brought closer to the actual one. To perform this, the timeline is modeled on more retrospective information. Figures 4-5 show graphs of fitting and predicting concentrations over a time series, obtained via the recurrent neural network (LSTM) model. Network training and further prediction were performed on a concentration data set, which was the only input parameter.

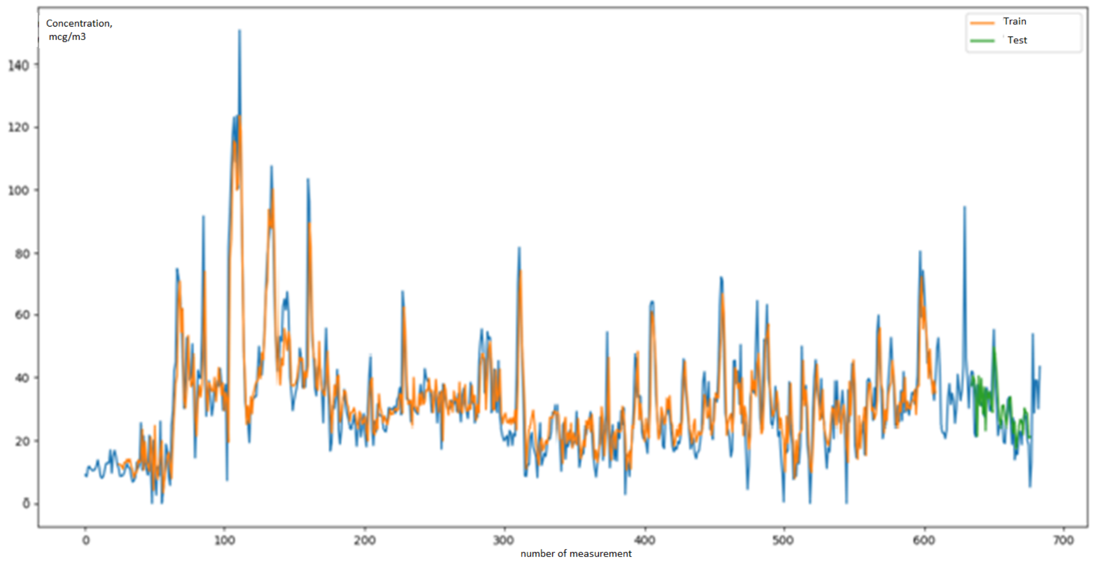

Figure 4: The fit of the model.

7. Discussion

The results obtained using the data analysis method show that the generated pollutant concentration forecast is close to real values. In addition, the data obtained allow to analyze future deviations in the substances concentration over long time periods. However, incomplete concentration data can be restored based on retrospective measurements.

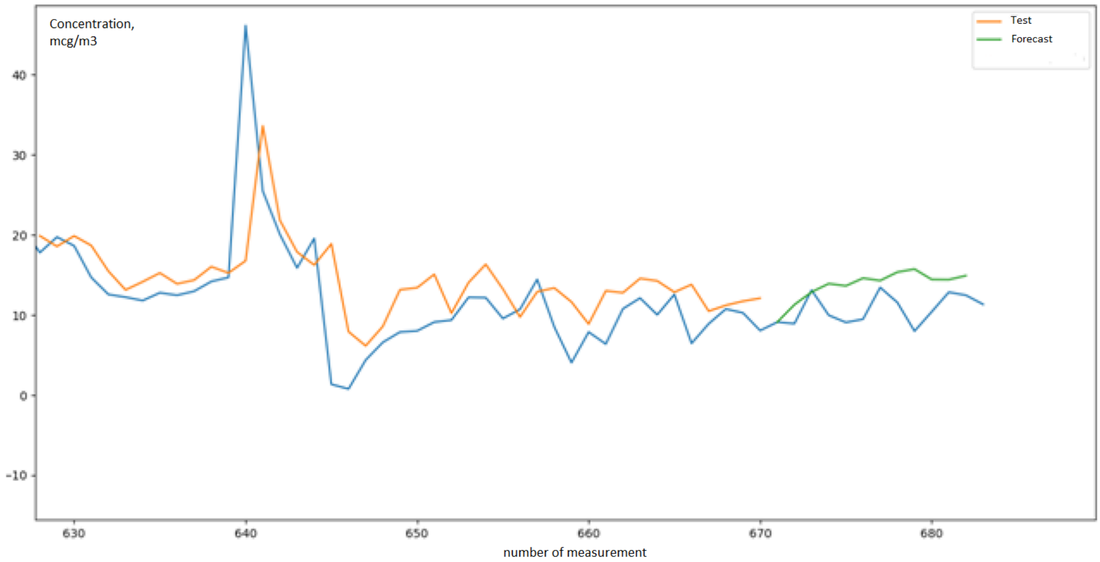

Figure 5: A certain vector of concentration growth.

Since the forecast accuracy were stable even in the incomplete data conditions, the proposed method allows to being implemented with the systems where sensors may fail due to different technical problems or malfunctions.

8. Conclusion

In this paper, we proposed and implemented a method for data processing and analysis that allows to predict deviations in the pollutants content, in the unreliable and incomplete data conditions. The method was implemented using the R and Python 3.7 programming languages, and was tested on real data on the environmental conditions obtained from public sources. The data was inaccurate and contained omissions in the measurements. Using the developed method, the missing data was restored, as well as the necessary parameters were evaluated and selected, on the basis of which the data forecast was performed. The predicted concentration values were close to the actual data. The industrial enterprises can benefit from implementing of such approach, where it is necessary to correctly predict the pollutants concentration in the atmosphere, since the proposed method allows to efficiently process the data that might be damaged or inaccurate.

Acknowledgements

This paper is supported by the Government of Russian Federation (grant 08-08).

References

[1] Pandey S. K., Kim K. H., Tang K. T. A review of sensor-based methods for monitoring hydrogen sulfide //TrAC Trends in Analytical Chemistry. – 2012. – pp. 87-99.

[2] Chiwewe T. M., Ditsela J. Machine learning based estimation of Ozone using spatio-temporal data from air quality monitoring stations //2016 IEEE 14th International Conference on Industrial Informatics (INDIN). – IEEE, 2016. – pp. 58-63.

[3] Shaban K. B., Kadri A., Rezk E. Urban air pollution monitoring system with forecasting models //IEEE Sensors Journal. – 2016. – №. 8. – pp. 2598-2606.

[4] C. Kühnerta, T. Bernarda, I. Montalvo Arango, R. Nitsche, "Water Quality Supervision of Distribution Networks Based on Machine Learning Algorithms and Operator Feedback" // Procedia Engineering, 89, 2014, pp. 189-196.

[5] Bianchi F. M. et al. Recurrent neural networks for short-term load forecasting: an overview and comparative analysis. – Springer, 2017.

[6] C. Plant, C. Böhm, "INCONCO: Interpretable clustering of numerical and categorical objects" // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011, pp. 1127-1135.

[7] K. Frederix, M. V. Barel, "Sparse spectral clustering method based on the incomplete Cholesky decomposition" // Journal of Computational and Applied Mathematics, 237(1), 2013, pp. 145-161.