Tomasz RYMARCZYK[1,2], Tomasz CIEPLAK[3], Grzegorz KŁOSOWSKI[3], Edward KOZŁOWSKI[3]

Research & Development Centre Netrix S.A.(1), University of Economics and Innovation (2)
Lublin University of Technology (3)

doi:10.15199/48.2020.05.18

The use of LSTM networks in the detection of outliers in IoT-based air quality monitoring systems

Abstract. *Recently, several attempts have been made to build social air quality monitoring systems. Systems of this kind are solutions in the creation of which it is necessary to solve many problems concerned with the collection and analysis of data. After all, such systems are complex, extensive and multidisciplinary IT solutions. Our work focuses on creating such a system which, in addition to being a distributed social system, additionally uses low-budget and available measuring devices. The system consists of the data acquisition subsystem, then the data collection and analysis subsystem, and the communication system with the end user. In this article, we focus on describing data acquisition subsystems and on one aspect related to data analysis, namely outliers prediction using recurrent neural networks in the form of their implementation as LSTM.*

Streszczenie. *W okresie kilku ostatnich kilku miesięcy podjęto działania budowy społecznościowych systemów monitorowania jakości powietrza. Systemy tego rodzaju są rozwiązaniami, przy tworzeniu których konieczne jest rozwiązanie różnorodnych problemów związanych z gromadzeniem i analizą danych. Systemy tego rodzaju to złożone, rozbudowane i multidyscyplinarne rozwiązania informatyczne. Opisywana praca koncentruje się na działaniach związanych z stworzeniem takiego systemu, który oprócz tego, że jest rozproszonym systemem społecznościowym, dodatkowo wykorzystuje niskobudżetowe i ogólnie dostępne urządzenia pomiarowe. System składa się z podsystemu gromadzenia danych, następnie podsystemu gromadzenia i analizy danych oraz systemu komunikacji z użytkownikiem końcowym. W tym artykule skupiamy się na opisie podsystemów akwizycji danych oraz na wybranym zagadnieniu związanym z analizą danych, a mianowicie przewidywaniu wartości odstających z wykorzystaniem rekurencyjnych sieci neuronowych w postaci ich implementacji jako sieci LSTM. (**Wykorzystanie sieci LSTM w wykrywaniu wartości odstających w systemach monitorowania jakości powietrza opartych na IoT**)*

Keywords: sensor networks, internet of things, long short-term memory , outliers detection
Słowa kluczowe: sieci sensorów, internet rzeczy, sieci neuronowe LSTM, wykrywanie wartości odstających

Introduction

The article describing our work consists of two parts. The first part explains the structure and operation of the system in technical terms. In this part, we focus on describing the technical elements included in the system. We depicted a measuring device. Then, we presented a data collection and analysis model. Therefore, the natural continuation is the second part of the article, in which we presented work on the implementation of the LSTM network to detect outliers on the example of O_3 concentration, which is a substantial component of smog.

An additional premise prompting us to continue our project described in previous works [1],[2] is the announcement of changes in regulations. The new regulations will require information on air quality in cities given more frequently and accurately compared to what is happening today. An additional problem is the fact that currently the air quality monitoring area is point-based and does not cover all districts of the city. This means that the information on air quality concerns the area in which the measuring devices are located. For example, in the city of Lublin there are only two measuring stations measuring air quality parameters. Therefore, in order to obtain data from the entire city, it is necessary to use a distributed measurement system. Another aspect in favor of conducting our research is the increase in environmental awareness, which in the future will make publicly available environmental sensors "everywhere".

The above assumptions, indicating the need to use publicly available measuring devices, result in poor quality of the data received. One of the important steps carried out while the system is in operation is sensor calibration. The work [3] presents methods of applying statistical techniques to calibrate low-budget air quality sensors. The impact of measurement errors on the final results of air quality measurements can also be eliminated by redundancy of the measuring elements and then the application of data analysis methods with which the data from the sensors will be properly prepared. In previous works [1],[2] we used distance methods in detecting outliers, in this work we will describe the use of artificial neural networks to create models that allow catching anomalies in incoming data.

System Design Assumptions

Brief description of the requirements

It is obvious that state institutions operating in the field of environmental protection are compulsorily forced to use professional, certified measuring stations. Therefore, at present, the number of systems monitoring the quality of the natural environment (including air quality) is rather small in Poland and amounts to 1230 autonomous stations [4]. To change and improve this situation, along with the growing popularity of amateur electronic systems, which was a direct result of falling prices and their high availability on the market, a discussion began in the scientific community about the desirability of using low-budget solutions for air quality monitoring purposes[5],[6]. A number of supranational institutions such as WMO (World Meteorological Organization) and European Commission have published in recent years reports on the use of low-budget sensors for air monitoring [7], [8]. The usefulness of this kind of sensors is also discussed in scientific articles[9], where the utility aspect of such devices has been tested. The results show that we are still in the initial phase of using such devices to create systems on a larger scale than just proof of concept.

Functional assumptions of the system

Designing and building a system consisting of many tens, hundreds or thousands of sensory devices, you can meet many technical issues that need to be rethought and solved to achieve success. First of all, the IoT sensor system is a distributed platform, vertically and horizontally scalable. The system is built based on the architecture of integrated web services using various technologies, software languages and methods of intra-system communication, as well as with external devices and other systems.

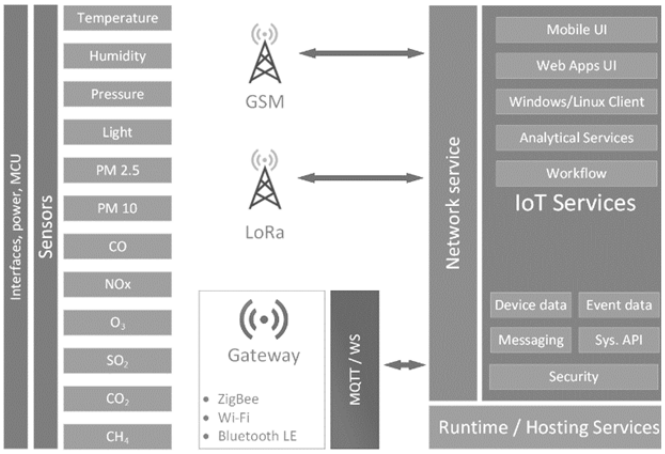

Fig. 1. Schema of sensors IoT operational blocks.

A diagram of a designed system (see Fig. 1) depicts a platform that is based on the paradigm of Internet o Things. The schema contains some number of fundamental modules (like sensors, connectivity, and the main subsystem). Data is collected using a collection of sensors with appropriate power and device management components, then the data is sent to the analysis and storage subsystem, and various types of network technologies are used (Bluetooth, LoRa Network, Wi-Fi, GSM). An important feature of the IoT sensor system is the possibility of using a variety of communication technologies (see Fig. 1).

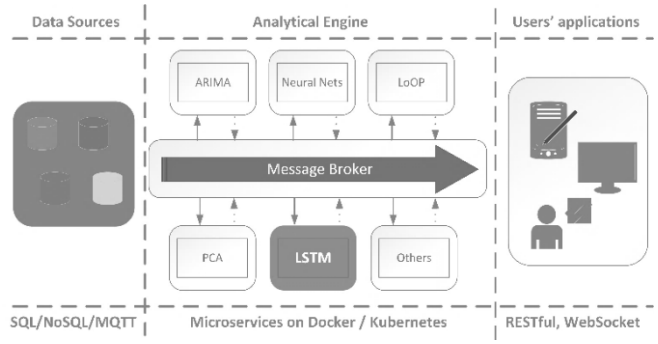

Fig. 2. Data flow within the system

The data analysis and preservation subsystem supports a large number of data sources, analytical methods and the data itself (see Fig. 2, above).

Analytical Methods

The use of commonly accessible air quality monitors is related to their low price, and sadly, to their lower quality compared to professional air quality monitoring stations. Hence, it is vital to develop a system that will be able to separate variable measurement values from measurement errors with a large variability of atmospheric conditions.

Measuring device and data transfer module

Devices applied to the research include a set of sensors like humidity, temperature, and air quality sensor (see Fig. 3). Raspberry Pi was used as a base of the measurement device. The device contains three blocks: A – sensors block with a very basic optical dust sensor GP2Y10; B - control and communication block consisting of the Raspberry Pi Zero device; C – batteries and power module.

Fig. 3. and transfer of measurement data.

The device was transferring data via Wi-Fi in series. MQTT was used as the communication protocol. The assumption related to the communication method, was the use of the JSON protocol for data exchange. The data received from the sensors was sent via the data bus to the database.

Fig. 4. The internal structure of the device for the measurement Data processing model

Data kept in the database constitute a repository of historical data used for batch analysis. In case of the systems, models are frequently created by knowledge transfer (obviously, if a given method allows it) and continuous enhancement of previously developed models. That's why historical data is a central part that permits one to continually create and refine models using a variety of statistical, machine learning or other even more advanced analytical tools. In contrast to batch analysis, stream analysis allows one to inspect incoming data from devices in real-time. The combination of these two methods of analysis, called Lambda architecture [10], lets the use of models created on historical data to explain phenomena occurring in real time. In the case of the low budget devices, an important step in the implementation and production use of such systems is the development of a method that allows increasing the quality of the acquired data [11].

A similar application of the Lambda architecture was used in the analysis of outliers in the measurement of electricity consumption [12].

Methods of anomaly detection

Process monitoring can use traditional techniques that use statistical measures, such as cumulative sum (CUSUM) and exponential moving average (EWMA) in a time window, to detect changes in the base distribution [13]. However, for a statistical measure to correctly identify changes in a process, one must first determine the length of the time window. The size of the time window has a decisive influence on the results of the statistical method.

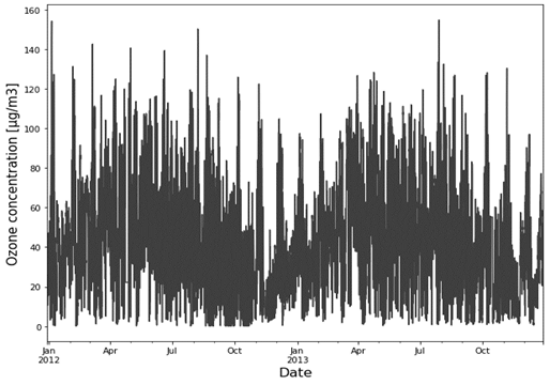

Fig. 5. Historical observations of ozone concentration in the air in Lublin from January 1, 2012 to the end of Decembre 2014.

Data preparation and analysis

Pollutants that reduce air quality depend on many variables, but they show seasonality and other patterns. Studies show [14], [15] that the concentration of PM 2.5,

PM 10 increases to dangerous levels in the autumn and winter months. In case of O_3, seasonality also depends on the season of the year. With O_3, the highest concentrations are recorded in the dry summer months when there is high sunshine.

Statistical tests (see Table 1) show that the time series is stationary. This enables value prediction using statistical methods.

Table 1. Stationarity tests results.

Augmented Dickey-Fuller Test Results:	
ADF Test Statistic	-26.210305
P-Value	0.000000
No. Lags Used	73.000000
No. Observations Used	131422.000000
Critical Value (1%)	-3.430400
Critical Value (5%)	-2.861562
Critical Value (10%)	-2.566782
Date type:	float64
Stationarity test:	True

Application of LSTM networks for outlier detection

Research has shown [16], [17], [18], [19] that Long Short Term Memory (LSTM) networks can be particularly used to detect sequences containing long-term patterns of unknown length. When outliers are detected, a model should be developed that can predict ozone concentration with some degree of uncertainty. To achieve this we used quantile regression [20]. There are many numerical methods that you can use to solve this problem to solve the problem [21-28]. Generally speaking, the determination of the outlier can be carried out in the following stages:

1. Training the LSTM network so that it can predict the next v value of $\{x_{t+1},...,x_{t+v}\}$ from previous p values in the $\{x_{t-p+1},...,x_t\}$ series.
   Using the time series $\{x_1,...,x_T\}$, the input of the LSTM network is the sequence of M dimensional vectors $\{x_{t-p+1},...,x_t\}$, and the output yields v vectors of the M dimensional vector $\{x_{t+1},...,x_{t+v}\}$ that is predicted simultaneously.
2. Calculation of error vectors values $e = x_{true} - x_{pred}$ using a trained model using the LSTM network.
   The x_{true}, x_{pred} values, respectively, are the observed and predicted values.
3. In the next action, the multidimensional Gaussian distribution should be adjusted to the determined error vectors calculated on the basis of test data using the maximum probability estimation (*MLE*) method.
   Calculation of the error vector at the point where the anomaly probably occurred. If the determined vector is at the ends of the Gaussian distribution estimated in the previous stage, it can be concluded that an anomaly has occurred.

Results

The data set was prepared on the basis of values obtained from the IMGW archive, the number of the data set is about 130,000 readings taken at intervals of one hour. The set was divided into subsets: learning and test in a 50/50 ratio.

LSTM network configuration

The network that detects outliers consists of connected LSTM layers and dense networks. LSTM layers allow the prediction of subsequent v values. The task of dense networks is to assign values to the appropriate quantile range. The network structure is shown in the figure below. The model is created on a daily basis, therefore there are 24 neurons at the input. And on the output return the predictor values for the variable values from the test set.

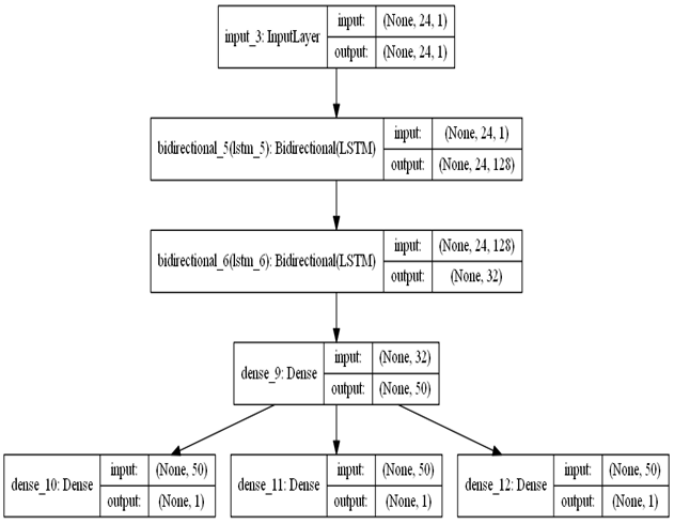

Fig. 6. The structure of the LSTM network used during model learning.

Table 2. Mean squared logarithmic error (MSLE).

90 quantile	0.31
50 quantile	0.11
10 quantile	0.24

When calculating the 90th and 10th quantile, we take into account the most likely values that the measured values can take. The width of this range between 10 and 90 quantile can vary widely. When the learning model has "certainty" about the future, then this range is relatively small, comparing it to the range when the model cannot correctly predict the changes. Therefore, the quantile interval indicates outliers in predicting ozone concentration in air.

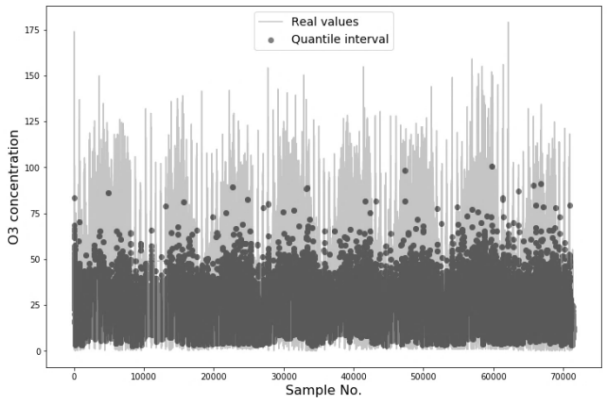

Fig. 7. The degree of uncertainty vs. actual data.

Conclusions

The article we described IT tools and analytical methods used to build and operate an air quality monitoring system. Our goal was to build a neural network with the help of which we will teach a model that allows determining outliers in measurement data. The tests show that the use of the LSTM network is a good starting point for further research related to data analysis in the presented measurement system.

Authors: *Tomasz Rymarczyk, Ph.D. Eng., University of Economics and Innovation, Projektowa 4, Lublin, Poland, E-mail: tomasz@rymarcyk.com*
Tomasz Cieplak, Ph.D., Lublin University of Technology, Nadbystrzycka 38A, Lublin, Poland, E-mail: t.cieplak@pollub.pl*;*
Grzegorz Kłosowski, Ph.D. Eng., Lublin University of Technology, Nadbystrzycka 38A, Lublin, Poland, E-mail: g.klosowski@pollub.pl*;*
Edward Kozłowski, Ph.D., Lublin University of Technology, Nadbystrzycka 38A, Lublin, Poland, E-mail: e.kozlovski@pollub.pl*;*

REFERENCES

[1] Cieplak T., Rymarczyk T., Tomaszewski R., A concept of the air quality monitoring system in the city of Lublin with machine learning methods to detect data outliers, MATEC Web Conf., 252 (2019), 03009

[2] Rymarczyk T., Cieplak T., Klosowski G., Kozlowski E., Monitoring the natural environment with the use of IoT based system, in 2019 Applications of Electromagnetics in Modern Engineering and Medicine (PTZE), 2019, 151–155.

[3] Cordero J. M., Borge R., Narros A., Using statistical methods to carry out in field calibrations of low cost air quality sensors, Sensors Actuators B Chem., 267 (2018), 245–254,

[4] Główny inspektorat ochrony środowiska, Informacje ogólne - GIOŚ, 2019. [Online]. Available: https://powietrze.gios.gov.pl/pjp/content/measuring_air_assessment_measurings. [Accessed: 01-Sep-2019].

[5] Castell N. et al., Localized real-time information on outdoor air quality at kindergartens in Oslo, Norway using low-cost sensor nodes, Environ. Res., 165 (2018), 410–419

[6] Popoola O. A. M. et al., Use of networks of low cost air quality sensors to quantify air quality in urban settings, Atmos. Environ., vol. 194, pp. 58–70, Dec. 2018.

[7] Lewis A. C. et al., Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications Draft copy for public review, 2018.

[8] Gerboles M., Spinelle L., Borowiak A., Measuring air pollution with low-cost sensors | EU Science Hub, European Commission, 2017. [Online]. Available: http://ec.europa.eu/environment/air/ pdf/Brochure lower-cost sensors.pdf. [Accessed: 03-Jun-2019].

[9] Robinson J.A., Kocman D., Horvat M., Bartonova A., End-User Feedback on a Low-Cost Portable Air Quality Sensor System-Are We There Yet?, Sensors (Basel), 18 (2018), no. 11

[10] Yamato Y., Kumazaki H., Fukumoto Y., Proposal of Lambda Architecture Adoption for Real Time Predictive Maintenance, in 2016 Fourth International Symposium on Computing and Networking (CANDAR), 2016, 713–715.

[11] Bun B., Calibration using supervised learning for low-cost air quality sensors., 2017.

[12] Liu X., Iftikhar N., Nielsen P.S.,Heller A., Online Anomaly Energy Consumption Detection Using Lambda Architecture, Springer, Cham, 2016, 193–209.

[13] Chatfield C.,The Analysis of Time Series. Chapman and Hall/CRC, 2016.

[14] Chen W., Yan L., Zhao H., Seasonal Variations of Atmospheric Pollution and Air Quality in Beijing, Atmosphere (Basel)., 6 (2015), No. 11, 1753–1770

[15] Cichowicz R.,. Wielgosiński G., Fetter W., Dispersion of atmospheric air pollution in summer and winter season., Environ. Monit. Assess., 189 (2017), No. 12, 605

[16] Malhotra P., Vig L., Shroff G., Agarwal P., Long Short Term Memory Networks for Anomaly Detection in Time Series. .

[17] Yang H., Pan Z., Tao Q., Robust and adaptive online time series prediction with long short-term memory, Comput. Intell. Neurosci, 2017 (2017), 1–9

[18] Singh A., Anomaly Detection for Temporal Data using Long Short-Term Memory (LSTM), 2017.

[19] [1] Kłosowski G., Rymarczyk T., Gola A., Increasing the reliability of flood embankments with neural imaging method. Applied Sciences, 8 (2018), No. 9, 1457.

[20] Rodrigues F., Pereira F. C., Beyond expectation: Deep joint mean and quantile regression for spatio-temporal problems., preprint arXiv:1808.08798, 2018

[21] Rymarczyk T, Kłosowski G. Innovative methods of neural reconstruction for tomographic images in maintenance of tank industrial reactors. Eksploatacja i Niezawodnosc – Maintenance and Reliability, 21 (2019); No. 2, 261–267

[22] Rymarczyk, T.; Kozłowski, E.; Kłosowski, G.; Niderla, K. Logistic Regression for Machine Learning in Process Tomography, Sensors, 19 (2019), 3400.

[23] Romanowski A., Big Data-Driven Contextual Processing Methods for Electrical Capacitance Tomography, IEEE Transactions on Industrial Informatics, 15 (2019), No. 3, 1609-1618

[24] Kozlowski E.; Mazurkiewicz D., Kowalska B. et al., Binary Linear Programming as a Decision-Making Aid for Water Intake Operators, 1st International Conference on Intelligent Systems in Production Engineering and Maintenance, Wroclaw, Poland 28-29.09.2017,

[25] Kryszyn J., Smolik W., Toolbox for 3d modelling and image reconstruction in electrical capacitance tomography, *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska (IAPGOŚ)* , 7 (2017), No. 1, 137-145

[26] Majchrowicz M., Kapusta P., Jackowska-Strumiłło L., Sankowski D., cceleration of image reconstruction process in the electrical capacitance tomography 3d in heterogeneous, multi-gpu system, *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska (IAPGOŚ)*, 7 (2017), No. 1, 37-41

[27] Galazka-Czarnecka, I.; Korzeniewska E., Czarnecki A. et al., Evaluation of Quality of Eggs from Hens Kept in Caged and Free-Range Systems Using Traditional Methods and Ultra-Weak Luminescence, Applied sciences-basel, 9 (2019), No. 12, 2430.

[28] Szczęsny A., Korzeniewska E., Selection of the method for the earthing reistance measurement, Przegląd Elektrotechniczny, 94 (2018), No. 12, 178-181.