Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2020) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 11th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 6-9, 2020, Warsaw, Poland

Analysis and processing of environmental monitoring system

Viktor T. Pyagay*[a], Zhibek N. Sarsenova[a], Kulyanda S. Duisebekova[a], Nurzhan T. Duzbayev[a], Nurtai Albanbai[b]

[a]*International Information Technology University, 34-1 Manas str., Almaty 050040, Kazakhstan*
[b]*Kazakh National University, 71 al-Farabi Ave, Almaty 050040, Kazakhstan*

Abstract

In this article, the problem of monitoring of climatic and ecological condition of the region is considered. The authors propose an approach to the construction of such systems using the LoRaWAN (Long Range Wide Area Network) technology. This will make it possible to create easily scalable low-cost systems with high energy efficiency through the use of modern communication technologies. During the experiment, the authors have identified certain patterns in collected data, such as dependence on the time of year, weather, and the location of certain industrial facilities near the observed zone.

Keywords: LoRaWAN; Environmental monitoring; Big Data; IoT; Ecology;

Introduction

Environmental pollution by enterprises and industrial vehicles, causing habitat degradation and damaging the health of the population, is the most acute problem of priority social and economic importance. Most of the air is poisoned by automobile exhausts, exhaust gases and power plants, and fires.

The main air pollutants in the Republic of Kazakhstan are manufacturing enterprises, their share in the total volume of emissions is 50%, enterprises for the production and distribution of electricity, gas and water 28%, mining enterprises occupy 14% of the total structure, the remaining industries account for about 8%. At the same time, air

* Corresponding author. Tel.: +7-707-220-1190; fax: +7-727-320-0000.
E-mail address: v.pyagai@edu.iitu.kz

pollution is associated with emissions from non-ferrous and ferrous metallurgy enterprises, thermal power engineering, and the oil and gas complex. At present, on average, in the Republic of Kazakhstan, per inhabitant, 200 kg of various chemical compounds are released into the atmosphere per year, while in 2000 this indicator was 163 kg [1].

The number of vehicles is growing every year, private-sector heating systems are improving, the emissions of which are carried out in the surface layer of the atmosphere above the territory of industrial centers, new technologies for controlling harmful emissions appear, but despite this, the quality of the air leaves much to be desired. Therefore, the development of methodologies for reducing pollutant emissions, means of monitoring and controlling the level of pollution in order to reduce the technogenic impact on the atmosphere is currently relevant.

1. Problem, relevance

The object of the study is the ecological state of the air environment of the city of Almaty and the region, emissions and variability of concentrations of pollutants.

The subject of the study is the anthropogenic and meteorological conditions of air pollution, which determine the concentration of the main pollutants of the air basin.

The study of large amounts of data, in modern terminology defined as "big data", is a key element in the theory of the "Fourth Paradigm of Science" proposed by the Turing Prize winner, American scientist Dr. James Nicholas Gray. [2] Gray suggests that the collection and analysis of large volumes of statistical data, made possible by the widespread dissemination of computer systems, provides the world with a fundamentally new approach to research and will allow a fresh look at many problems of modern science. A mathematical analysis of the data will help to identify patterns and increase the accuracy of the final conclusions, which would be impossible for smaller samples. That is why it is important to provide both a large number of sources of such data and the most simplified access to them. The latter also reflects such a common concept in the scientific community as “open data”, i.e. providing free access to data to everyone, so that anyone can conduct research on them and thereby accelerate the solution of a specific scientific problem. For such studies in ecology in foreign sources, the term "macro system ecology" (macrosystems ecology, MSE) is used [2]. One of the approaches used in the ecology of macrosystems to obtain large amounts of data is precisely the use of extensible eco-monitoring systems that collect data on the current state of the environment.

In cities, the environmental sustainability index is usually calculated from four sub-indices: air quality, CO2 emissions, energy and indoor pollution.

Existing ground-based information systems use data from meteorological and upper-air stations, data from water and aircraft, weather radar centers, and various atmospheric sounding systems. An important role is played by information obtained from artificial satellites of the Earth, as well as monitoring data of the cryosphere (state of snow and ice cover and permafrost zones). Also, climatic and ecological changes in the environment require monitoring of internal and external factors. External factors are factors caused by the influence of solar and cosmic radiation. The intensity of external factors of influence depends on solar activity, the parameters of the Earth’s orbit, and the Earth’s rotation speed. The effects of impacts are determined by the intensity of the impact factors, the properties and composition of the Earth’s atmosphere, and the properties of the earth’s surface. Internal factors influencing climate variability include thermal emissions and emissions of various substances into the biosphere, as well as their redistribution between different environments.

Currently, in terms of analysis and assessment of environmental and technological hazards, an exclusive role is given to the environmental monitoring system. In this area, to predict the development of environmentally hazardous situations, the old practice based on observation, accumulation of data and compilation of bulletins of environmental pollution is not enough. Environmental monitoring is an information system for observing, assessing and forecasting changes in the state of the environment, created with the aim of highlighting the anthropogenic component of these changes against the background of natural processes. Speaking about the environmental monitoring system, we mean that it should accumulate, systematize and analyze information: on the state of the environment; the reasons for the observed and probable changes in the state (i.e., the sources and factors of influence); on the permissibility of changes and loads on the environment as a whole. In accordance with the above definitions and the functions assigned to the system, three main areas of activity can be distinguished, which include environmental monitoring[3]:

- monitoring of impact factors and the state of the environment;
- assessment of the actual state of the environment;
- prediction of the state of the environment and assessment of the forecasted state.

It should also be noted that the monitoring system itself does not include environmental quality management activities, but it is a source of information necessary for making environmentally significant decisions. The systematic method of environmental monitoring is based on the examination of the environmental impact of harmful emissions on the environment and provides a comprehensive accounting of measurements and comparing them with standard indicators expressed through qualitative and quantitative characteristics of environmental safety. The method of environmental impact assessment in the development of the system contains a set of measures, including identification, analysis, tracking and monitoring of environmental risks from their planned values. Environmental monitoring, implied in this work, is a system of mobile observation points for the state of air pollution in the territory of Almaty, in particular for the spread of pollutants from stationary sources of pollution, i.e. from industrial enterprises. The largest volume is observed in manufacturing enterprises (45,9%).

The project involves the creation of a software and hardware complex for environmental monitoring.

2. Problem, relevance

To solve the problems of forecasting and monitoring the ecology, climate and meteorological conditions, it was decided to develop a platform for collecting and transmitting heterogeneous data in Almaty and Almaty region.

The main goal of the system is monitoring air quality, forecasting, assessing and identifying trends in the state of the atmosphere to prevent negative consequences that adversely affect human health and the environment. This allows real-time monitoring of the environment and meteorological conditions, as well as monitoring of all major pollution sources for making subsequent management decisions. When designing the system, the following requirements were established: low cost of creation and operation, autonomy of the end measuring stations, scalability, online access to the collected data, data security and stability of the system as a whole.

A diagram of a fragment of the network of the system was developed on the basis of an analysis of the parameters necessary for assessing the ecological situation, specific terrain, and the development of the city. The scheme consists of data collectors, which include several local and one base station. Data collectors are located in different parts of the city. Within the region, data collection is based on LoRa (Low Range) technology. Data transmission from data collectors is carried out using cellular mobile communication systems. Each local station is mobile. The local station includes sensors for measuring the necessary parameters, electronic processing devices, a LoRa radio module with technology. The base station is stationary. It also contains sensors for measuring parameters, electronic processing devices, as well as equipment for transmitting information over cellular networks (GSM or LTE standard) and a LoRaWAN gateway. Data received from local and base stations is transmitted to the server and can be additionally available to the system user via the web interface or mobile application, as well as to third-party software products via API (Application Programming Interface).

The core of each local station is a board based on the ARM (Advanced RISC Machines) Cortex-M3 controller, and the core of each base station is a single-board computer based on the ARM Cortex microprocessor - A53. Gas analyzers and other measuring equipment connected to them. A list of air pollutants was determined, the concentrations of which should be controlled. This list includes substances such as carbon monoxide, nitric oxide, ozone, methane and suspended particles in the air. Also a station for connecting temperature, humidity, pressure, GPS sensors. Stations can be equipped at the request of the user, depending on the installation location.

As a data transmission technology, a network based on the LoRa standard was chosen. As mentioned earlier, the low data rates in the LoRa network are offset by the large coverage area with low power consumption. Among other things, LoRa networks use the unlicensed radio frequency range LPD433 (Low Power Device) and the ISM (industrial, scientific and medical) band 915 MHz [4], which will allow you to quickly deploy the network without additional permissions, as well as expand it with new devices as needed without a monthly fee.[4].

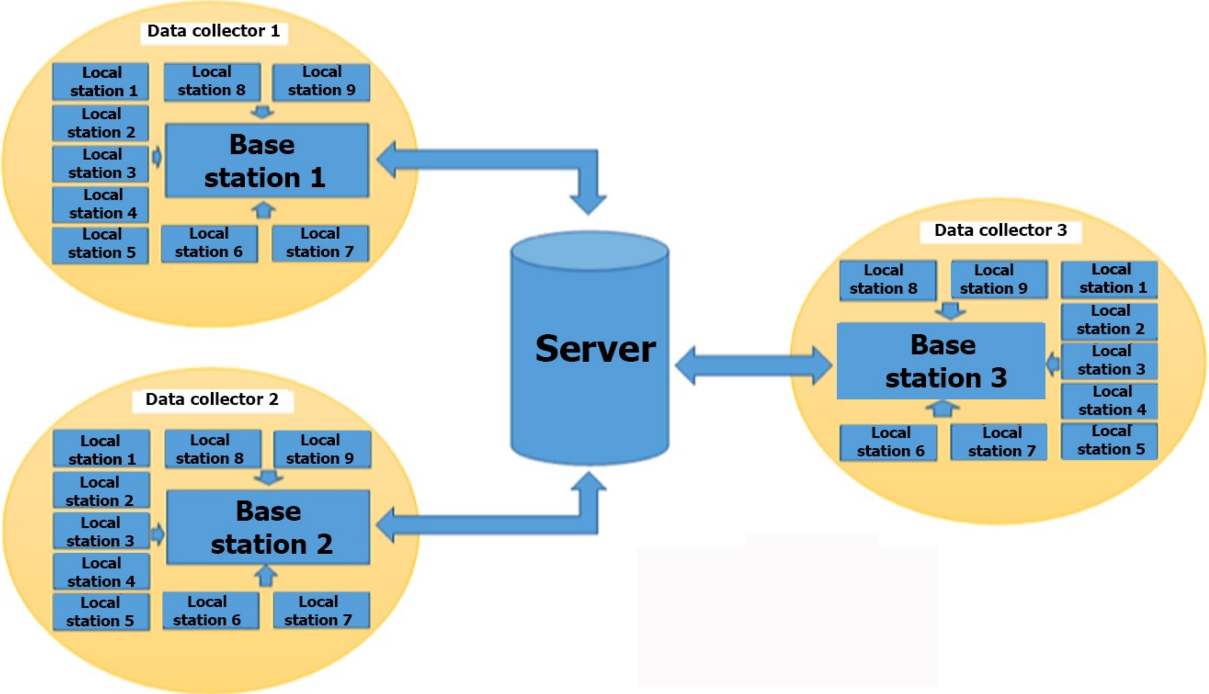

Fig. 1 - Block diagram of a network fragment

Local stations collect data from the sensors and transmit them to the nearest base station (LoraWAN gateway). If no base station is within range of the local station, data is stored on a removable SD card. As soon as the local station re-enters the coverage area of the base station, all the accumulated information will be automatically sent to the base station. The base station also collects readings at the installation site and transfers the received information to the server, and not for processing. In addition to stations, the LoRa network has a management server that organizes the transfer of data from end devices to the storage server and vice versa. The network server resolves network collisions, changes the transmitter power and data transfer rate, monitors the battery power of the end devices, and sends data to the storage server. On the server for processing and storage of data is the extraction, storage and processing of data received from measuring stations.

Third-party applications receive environmental data from the data processing and storage server, bring them in accordance with the units of measure adopted in the region, and display them to the end user both in the form of summary information from all stations and for each station on the network separately. Interactive maps, pivot tables, histograms are used to visually display information available both through the website and through the mobile application.

The developed scheme, implemented on a modern elemental base, will provide the statistical data necessary for the analysis of an automatic network to collect environmental and meteorological information at different times of the day, under different weather conditions.

LoRa Node (End device) consists of several block diagrams listed below:

- Block diagram for determining the coordinates of devices
- A block diagram for storing information in memory
- Block diagram of a real-time clock
- LoRa radio module

The RAK811 radio module was used as a radio module. Radio module allows you to create a radio network type "star". The module is equipped with a radio frequency transceiver, this transceiver has the ability to transmit data in urban areas at a distance of up to 5 km (line of sight-15 km). the module used a microcontroller STM32L151, this microcontroller has the same advantage as the amount of power consumption. The module can be controlled with simple AT commands.

Different wireless technologies meet the needs of a particular application with changes in modulation schemes and frequency. Long-range devices with low bandwidth requirements (Fig. 2), which are typical for IoT applications, are poorly supported by these existing technologies. LPWAN is designed for similar applications and devices. [5]

One possible solution to the problem of wireless network connections with cellular topology is ZigBee technology. The main disadvantage of this technology is the limited bandwidth supply due to the high data rate and low receiver sensitivity. Some ZigBee connections have problems sending data over a distance of more than 20-30 meters, as the energy coming from the transmitter is lost too quickly.

Instead of a mesh network, most LPWAN technologies use a star-shaped network. As with Wi-Fi, star network endpoints connect directly to the access point. It is also possible to use a repeater to easily fill gaps in coverage, which for most applications is a good intermediate solution in terms of latency, reliability, and coverage.

To achieve a long distance in wireless communication, a large channel reserve is required. In other words, when a signal is transmitted, it needs enough energy to detect when received. Because a certain amount of energy is lost in the process of spreading through space and the materials in between.

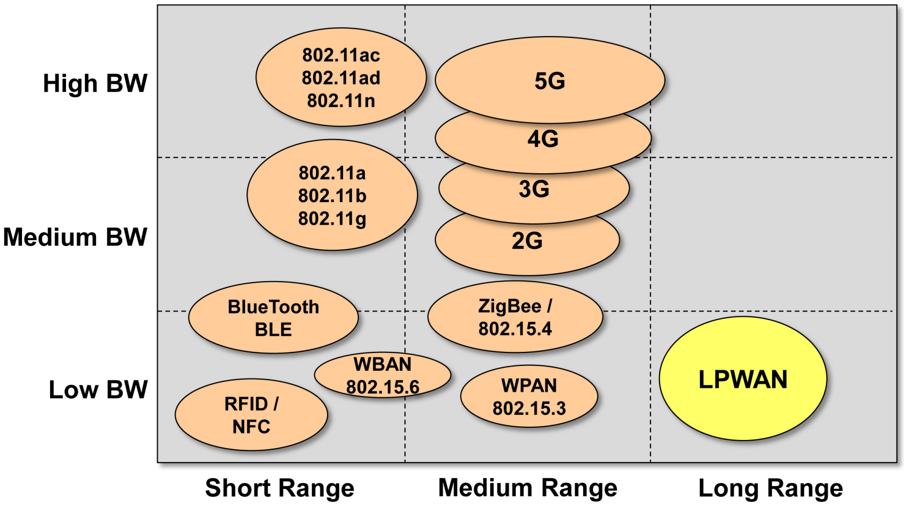

Fig. 2 - Range of different wireless technologies

LPWAN (Low Power WAN) technologies typically operate at 140-160 decibels (dB) of the total path, which can increase the range to several kilometers under suitable conditions. This is primarily achieved due to the high sensitivity of the receiver. Receiver sensitivity of more than -130 dBm is common for LPWAN technologies compared to the -90 to -110 dBm seen in many traditional wireless technologies. The -130 dBm technology can detect signals 10,000 times weaker than the -90 dBm technology, which is a significant advantage of LPWAN[5].

3. Results and Discussions

In statistical applications, data analysis can be divided into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). Descriptive statistics is a summary of statistics that quantitatively describes or sums up a feature of information collection, while descriptive statistics in the mass sense is the process of using and analyzing these statistics. The EDA focuses on discovering new features in the data, while the CDA focuses on confirming or falsifying existing hypotheses. In this article, we used descriptive statistics because the data contains a simple summary of the sample and the observations made. Also since descriptive analysis includes one-factor analysis, where one-factor analysis includes a description of the distribution of a single variable, including its Central trend (including mean, median, and mode) and variance (including the range and quartiles of the dataset, as well as propagation measures such as variance and standard deviation)).

Authors [7] proposed a solution for semi-empirical equation of turbulent diffusion to calculate the average of the impurity concentration in the boundary layer of the atmosphere from an instantaneous point source.

For visual statistics, the Google Collaborator tool was used, aimed at simplifying research in the field of machine and deep learning. The advantages of this service, you can get remote access to the machine with a connected video card, and completely free of charge, which greatly simplifies life when you have to train deep neural networks.

Air pollution data is presented as a csv file it has 11403 instances of training data.

In figure 3, all 24 attributes have missing values, 5 more than 50% of all data. In most cases, NA means the absence of the subject described by the attribute, for example, the absence of data on certain days for technical or other reasons.

As it can be seen in figure 3, the largest amount of data is missing in columns 15 and 17, and columns 2 and 6 have almost all the values. It is also worth noting that 100 percent of the completed data in this file is not. Therefore, it is necessary to remove all empty cells, that is, to get rid of hidden data. 11403 is clean, cleaned data.

Figures 4 and 5 show the distribution of the data in 3 forms. Grand Total is the average for each row. Figure 4-a used the Johnson distribution, figure 4-b used the normal distribution, figure 5 the lognormal distribution, where the x coordinate represents the values of pollutants, and the y coordinate represents the percentage of occurrence of the corresponding indicators.

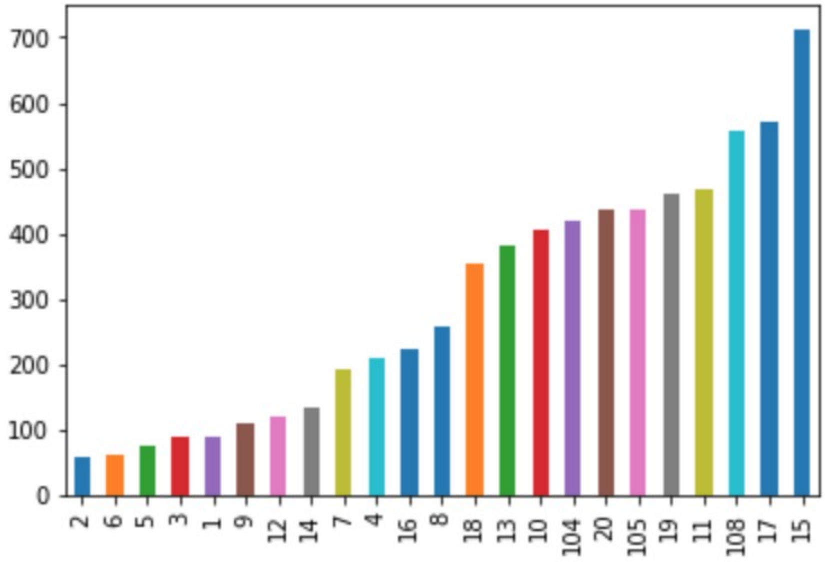

Fig. 3 – Percentage of missing metrics in the file

The Johnson's SU-distribution is a four-parameter family of probability distributions first investigated by N. L. Johnson in 1949.

The normal distribution of data is a pattern of occurrence of its different values. A lognormal distribution in probability theory is a two-parameter family of absolutely continuous distributions. If a random variable has a lognormal distribution, then its logarithm has a normal distribution.

Obviously, Grand Total does not follow the normal distribution, so it must be converted before performing the regression. While log conversion works pretty well, Johnson's unrestricted distribution works best.

The mass concentration of PM2.5 (particulate matter) is a key parameter for assessing air quality and its threat to human health. According to the standards of the World Health Organization (WHO), the average annual level of PM2.5 should not exceed 10 µg / m3, and the average daily level should not exceed 25 µg / m3. In the meantime, as it can be seen on the Table 1, concentration of PM2.5 within the city of Almaty perceptibly exceeds the recommended values which affects on the health of citizens of the region in particular infectious diseases like tuberculosis[8].

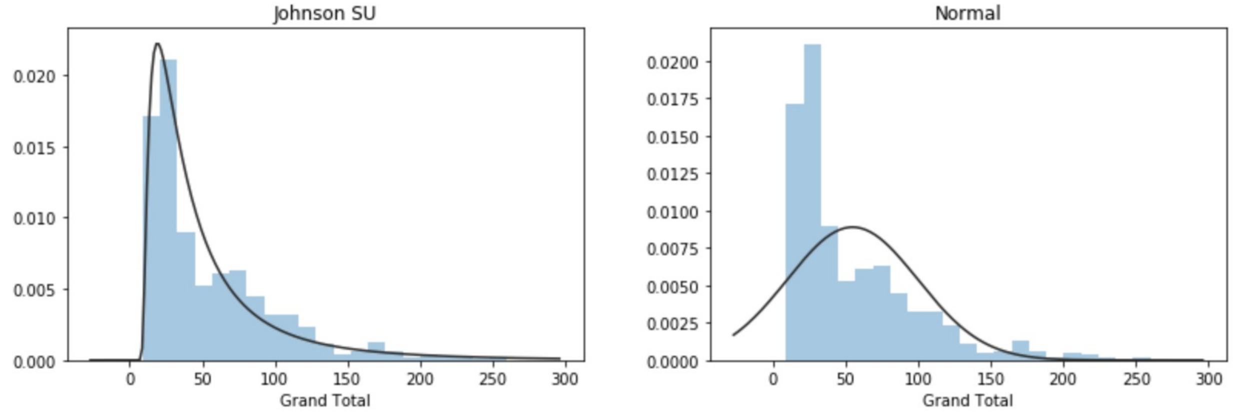

Fig. 4 - (a) Johnson Distribution; (b) Normal distribution

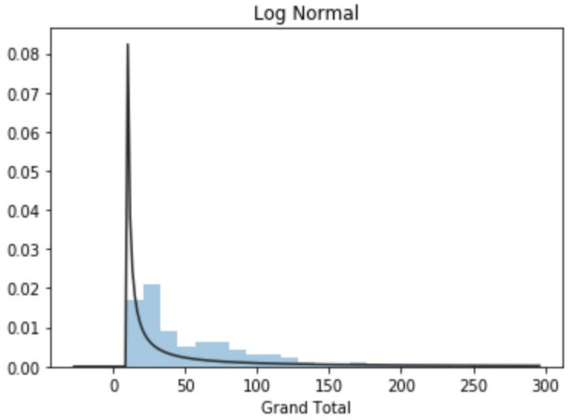

Fig. 5 - Lognormal distribution

The system developed during the research will help optimize the process of taking readings from various remote sensors installed in the studied region using modern LPWAN radio access technologies. The mobile monitoring system is implemented based on the construction of a radio access network of the LPWAN technology using various remote sensors of climatic and environmental conditions in the territory of the metropolis.

The study provides a detailed analysis based on Google Colaboratory cloud services to identify deviations from the standard value of monitoring climatic and environmental conditions. The authors came to certain regularities, such as the dependence on the time of year, weather and location of certain industrial facilities near the observed zone. It is also revealed that the concentration of suspended particles in the air exceeds the norm by 2 times most of the observed time period, and the possible consequences of this are considered. Of the 23 points, it was also determined which were more polluted and which were less. The authors also came to the conclusion that in the winter months showed air pollution revealed the highest rate than in the summer months.

Table 1. Average annual level of PM2.5 within the city of Almaty in 2018

Sensor №	Average value of PM2.5
1	71.21013873276992
2	51.85856141279108
3	54.5784159142277
4	42.46047823660232
5	57.493546891926606
6	47.461127920254135
7	83.72271849117314
8	51.139981386440645
9	37.697744085226546
10	68.394826476
11	39.033943278327
12	27.41208597187028
13	46.17060518731989
104	48.203698908774186
108	109.48883671872834

From Table 1 we can conclude that the most air polluted points are:

1. Point 108 (Jean Kairat) with a value of 109.4
2. Point 7 (Kok Kainar) with value 83.7
3. Point 1 (Seifullina - Dulatova) with a value of 71.2

Future research on data prediction will be concluded as well as development of SMS or other ways of notification of citizens on the environmental composition of the area.

Acknowledgments

The research was carried out within the framework of the project №AP05134597 "Software-hardware complex for analysis and monitoring of climatic and ecological changes of the environment".

References

[1] Valihan Bishimbaev, Ph.D. 2010. South-Kazakhstan State University, Kazakhstan Fatima Ermahanova,Ph.D Eurasian National University, Kazakhstan - Rational use of associated petroleum gas and benefits of electrostatic gas cleaning. - ISSN:1804-0527.

[2] Soranno, P.A.. 2014. Macrosystems ecology: big data, big ecology / P.A. Soranno, D.S. Schimel // Frontiers in Ecology and the Environment. – **1** – p. 3.

[3] Bushmeleva K.I. Plyusnin I.I. Sysoev S.M. Bushmelev P.E. Elnikov A.V. 2007. *Modern high technology*. - **3** - S. 41-43

[4] Kulyanda S. Duisebekova, Zhibek Sarsenova, Viktor Pyagay, Saule T. Amanzholova. 2018. Environmental monitoring system for analysis of climatic and ecological changes using LoRa technology // *The 5th International Conference* DOI: 10.1145/3330431.3330446.

[5] LoRa Alliance. Accessed – 15.03.2018. Available at:URL: https://www.lora-alliance.org/;

[6] Álvaro Gómez-Losada, Francisca M. Santos, Karina Gibert, José C.M. Pires. 2019. A data science approach for spatiotemporal modelling of low and resident air pollution in Madrid (Spain): Implications for epidemiological studies. *Computers, Environment and Urban Systems*, Available at: https://www.sciencedirect.com/science/article/pii/S0198971518304447#bb0010

[7] Duysebekova, K., Serbin, V., Kuandykov, A., Kozhamzharova D. 2016. The Solution of Semi-empirical Equation of Turbulent Diffusion in Problems of Polluting Impurity Transfer by Gauss Approach 2016 *Procedia Computer Science*. CHARMS-2016. p. 372-379

[8] Rakhmetulayeva S.B., Duisebekova K.S., Mamyrbekov A.M., Kozhamzharova D.K., Astaubayeva G.N., Stamkulova K. 2018. Application of Classification Algorithm Based on SVM for Determining the Effectiveness of Treatment of Tuberculosis. // *9th International Conference on Ambient Systems, Networks and Technologies*, ANT-2018 and the 8th International Conference on Sustainable Energy Information Technology, SEIT 2018, Porto, p. 231-238, Scopus.