УДК 005
МРНТИ 20.53.19

DESCRIPTIVE STATISTICS IN ECOLOGICAL MONITORING

ZH.N. SARSENOVA, V.T. PYAGAI, Z.Ye. TUYAKOVA

International Information Technology University

***Abstract**: This article discusses the impact of suspended particles on human health, as well as the analysis of the level of pollution in Almaty over the past 2 years using descriptive statistics. Pollution of the environment by industrial enterprises and vehicles, causing degradation of the environment and causing damage to public health, remains the most acute environmental problem of priority social and economic importance. The problem of pollution of the environment of large cities is very significant and complex, requiring first of all long-term monitoring, then a deep and competent analysis of the assessment of the situation on the data obtained, for the subsequent prevention, localization and investigation of environmental disasters and incidents, making management decisions for further work on the development of improving the quality of atmospheric air, as well as forecasting the state of the environment. In this paper, we consider a specific case, namely particles PM2.5 — air pollutant, which consists of solid particles and liquid droplets ranging in size from 10 nm to 2.5 microns. The fact that air pollution by small particles is a global killer is already widely known, but these statements have not yet been confirmed by specific figures. The authors have identified certain patterns, such as dependence on the time of year, weather, and the location of certain industrial facilities near the observed zone. The indicators with the values in the form of graphs And revealed that the concentration of suspended particles in the air exceeds the norm by 2 times most of the observed period of time, and considered the possible consequences of this.*

***Keywords**: air pollutant, environmental monitoring, descriptive statistics*

ЭКОЛОГИЯЛЫҚ МОНИТОРИНГТЕГІ СИПАТТАМА СТАТИСТИКАСЫ

***Аңдатпа**: Бұл мақалада өлшенген бөлшектердің адам денсаулығына әсер ету мәселесі қарастырылды, сондай-ақ Алматы қаласының соңғы 2 жылдағы ластану деңгейіне сипаттама статистикасын пайдалана отырып талдау жүргізілді. Қоршаған ортаны өнеркәсіп кәсіпорындары мен көлік құралдарының тіршілік ету ортасының тозуын тудыратын және халықтың денсаулығына зиян келтіретін ластануы басым әлеуметтік және экономикалық маңызы бар неғұрлым өткір экологиялық проблема болып қала береді. Ірі мегаполистердің қоршаған ортасының ластануы проблемасы ең маңызды және күрделі болып табылады, ол ең алдымен ұзақ мониторингті, содан кейін экологиялық апаттар мен инциденттерді кейіннен болдырмау, оқшаулау және тергеу, атмосфералық ауаның сапасын жақсартуды әзірлеу жөніндегі жұмыстарды одан әрі жүргізуге және басқарушылық шешімдер қабылдау үшін алынған деректер бойынша жағдайды бағалауды терең және сауатты талдауды талап етеді, сондай-ақ қоршаған ортаның жай-күйін болжау. Бұл жұмыста нақты жағдайға ерекше көңіл бөлінді, атап айтқанда, PM2 бөлшектері.5-құрамына көлемі 10 нм-ден 2,5 мкм-ге дейінгі қатты бөлшектер мен сұйықтық тамшылары кіретін ауа ластағышы. Ауаның ұсақ бөлшектермен ластануы жаһандық кісі өлтіруші болып табылатыны белгілі, бірақ бұл мәлімдемелер әлі күнге дейін нақты сандармен расталмаған. Авторлар жыл мезгіліне, ауа райына және байқалатын аймаққа жақын белгілі бір өнеркәсіптік объектілердің орналасуына тәуелділік сияқты белгілі бір заңдылықтарды анықтады. Кесте түріндегі мәндері бар көрсеткіштер келтірілген. Сонымен қатар, ауадағы өлшенген бөлшектер концентрациясының көрсеткіштері байқалатын уақыт аралығының көп бөлігі нормадан 2 есе артық екені анықталды және осының ықтимал салдары қарастырылды.*

***Түйінді сөздер**: ауа ластаушы, экологиялық мониторинг, сипаттама статистика*

ОПИСАТЕЛЬНАЯ СТАТИСТИКА В ЭКОЛОГИЧЕСКОМ МОНИТОРИНГЕ

***Аннотация**: В данной статье рассмотрен вопрос влияния взвешенных частиц на здоровье человека, а также проведен анализ уровня загрязнения города Алматы за последние 2 года с использованием описательной статистики. Загрязнение окружающей среды предприятиями промышленности и транспортными средствами, вызывающее деградацию среды обитания и наносящее ущерб здоровью населения, остается наиболее острой экологической проблемой, имеющей приоритетное социальное и экономическое значение. Проблема с загрязнением окружающей среды крупных мегаполисов является весьма значимой и сложной, требующей в первую очередь длительного мониторинга, затем глубокого и грамотного анализа оценки ситуации по полученным данным, для последующего предотвращения, локализации и расследования экологических катастроф и инцидентов, принятия управленческих решений для дальнейшего ведения работ по разработке улучшения качества атмосферного воздуха, а также прогнозирования состояния окружающей среды. В данной работе рассмотрен конкретный случай, а именно частицы PM2.5 – воздушный загрязнитель, в состав которого входят твердые частицы и капли жидкости размером от 10 нм до 2,5 мкм. То, что загрязнение воздуха мелкими частицами является глобальным убийцей, уже широко известно, однако эти заявления до сих пор не подтверждались конкретными цифрами. Авторами выявлены определенные закономерности, такие как зависимость от времени года, погоды, и расположения определенных промышленных объектов вблизи наблюдаемой зоны. Приведены показатели со значениями в виде графиков. А также выявлено, что показатели концентрации взвешенных частиц в воздухе превышают норму в 2 раза большую часть наблюдаемого промежутка времени, и рассмотрены возможные последствия этого.*

***Ключевые слова**: загрязнитель воздуха, экологический мониторинг, описательная статистика*

INTRODUCTION

Currently, much attention is paid to the purity of the surrounding air. In connection with this problem, the concept of suspended particles comes to the fore. Suspended particles (PM - particulate matter) are a widespread air pollutant comprising a mixture of solid and liquid particles in the air in suspension [1].

Particles with a mass concentration of particles with a diameter of less than 10 microns (PM10) and particles with a diameter of less than 2.5 microns (PM2.5) affect human health. PM2.5 particles are fine suspended particles, ultrafine particles with a diameter of less than 0.1 microns are also included in this category.

The main problem with fine particles is that particles with a diameter of 0.1 μm to 1 μm can be in the air for many days and weeks, and as a result of this, the particles are transported over long distances through the air [1].

It is known that PM by type of origin are of two types: primary and secondary [2].

Primary particles are emitted into the atmospheric air in the "finished" form - these are the smallest pieces of soot, automobile tires and asphalt; heavy metal compounds (for example, oxides), mineral salt particles (such as sulfates, nitrates), as well as biological pollutants (some allergens and microorganisms).

Secondary suspended particles are formed directly in the atmosphere as a result of chemical reactions of gaseous pollutants. For example, oxides of nitrogen and sulfur are introduced into the air, which, when in contact with water, form acids, and solid particles of salts (nitrates and sulfates) are obtained from the acids.

In addition to the origin, suspended particles differ in the type of source: artificial (man-made) and natural (non-anthropogenic). The main source of anthropogenic particles is transportation (erosion of the road surface, erasing brake pads and tires) and industrial processes with burning solid fuels (coal, lignite, oil), construction, mining, many types of production (especially the production of cement, ceramics, brick smelting). Sources of natural particles include such phenomena as soil erosion in arid areas and organic evaporation.

Suspended particles are dangerous because they are able to penetrate deep into the lungs and settle there [3].

Particles of PM2.5 pass through the biological barriers of the body: the nasal cavity, upper respiratory tract, bronchi. Particles of PM2.5 together with air fall directly into the alveoli - the bubbles in which gas exchange occurs between the lungs and blood vessels. Therefore, not only the respiratory system, but also the cardiovascular disease is associated with suspended particles.

The World Health Organization (WHO) has conducted a study that deals with the effects of suspended particles on human health. WHO concluded that between 1999 and 2010, 3.1 million people died from the causes of PM2.5 particles. In addition, it was found that PM2.5 particles lead to a decrease in life expectancy by an average of 8.6 months [4]. Thus, PM2.5 particles are associated with 3% of deaths from diseases of the cardiovascular and respiratory systems and 5% of deaths from lung cancer.

The World Health Organization has concluded that the harm is caused by the chronic effects of these particles on the human body. To date, there are no data indicating a safe level of exposure or a threshold level below which there are no health effects.

THE CASE OF ALMATY: AIR QUALITY MEASURES

The problem of air pollution in Almaty rises not the first year. It starts from the mid-90s, when everyone is hoping to attract the attention of the authorities and the public. And perhaps everything becomes more obvious and tangible.

Polluted air is the world's largest environmental risk to human health. Almaty ranks first in the Republic of Kazakhstan for respiratory and endocrine system diseases, blood diseases, cancer and asthma, although there are no large industrial facilities in the region.

To improve the air quality in the city, a set of measures is needed, both from the government and from residents. The main sources of air pollution are emissions from heat and power plants, road transport and the private sector.

The PMx fine dust measurement devices (PM2.5, PM10) PMS5003, which are used in the civil air quality monitoring network, in 2018 were included in the register of the state system for ensuring the uniformity of measurements. These sensors are used in the largest civilian air monitoring network in Kazakhstan: on the site airkaz.org.

This type of equipment is also used by the largest international project launched in China "Air Quality Index", which provides information on the degree of air pollution in more than 80 countries with more than 10 thousand stations located in 800 main cities of the world, in real time.

Conducting studies to assess the impact of the environment on human health is an important tool for demonstrating the need to take measures to improve air quality and reduce the negative impact of environmental factors.

Thus, it was decided to conduct a similar study in the city of Nice (France). It is based on data on air pollution in the region, obtained using the AirPaca service [5], which provides sensor readings in the public domain. The atmospheric air was measured in 6 areas: Contes 2, Aéroport de Nice, Nice Promenade des Anglais, Nice Arson, Peillon, Nice Ouest Botanique. For information about patients suffering from dyspnea, was used the archive of the hospital Pasteur (Nice). Dyspnea (shortness of breath) - a violation of the frequency and depth of breathing, accompanied by a feeling of lack of air.

For processing and storing the data used in this study, an information analysis system was developed using the Big Data approach, which made it possible to increase the efficiency of data analysis.

Big Data is information technologies for processing various structured and unstructured data of very large volumes in the context of continuous growth of data volumes and their distribution across numerous nodes of a computer network [6].

IMPLEMENTATION AND EXPERIMENTAL RESULTS

In statistical applications, data analysis can be divided into descriptive statistics , analytical data analysis (EDA), and supporting data analysis (CDA). Descriptive statistics is a summary of statistics that quantitatively describes or sums up a feature of information collection, while de-

scriptive statistics in the mass noun sense is the process of using and analyzing these statistics. EDA focuses on discovering new features in the data, while CDA focuses on confirming or falsifying existing hypotheses . In this article, we applied descriptive statistics because the data contains a simple summary of the sample and the observations made. Also since descriptive analysis involves univariate analysis, where univariate 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 dissemination measures such as variance and standard deviation).).

For visual statistics, we used Google Colaboratory, which is not so long ago appeared cloud service 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.

Data on air pollution are presented in the form of a csv file, it has 11403 instances of the training data.

In Figure 1, 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 due to technical or other reasons.

As you can see in Figure 1, columns 15 and 17 do not have the largest data, and columns 2 and 6 have almost all values. It is also worth noting that 100 percent of the filled data in this file is not. Therefore, we have to remove all empty cells, that is, get rid of hidden data. 11403 are clean, cleared data.

In Figure 2, 3, 4 are presented in 3 types of data distribution. The Grand total is the average for each row. 2 used the figure of the Johnson distribution, in figure 3 used a normal distribution 4 log-normal distribution, where x-coordinate, the values presented in pollutants. A on the coordinate d represents the percentage of occurrence of the relevant indicators.

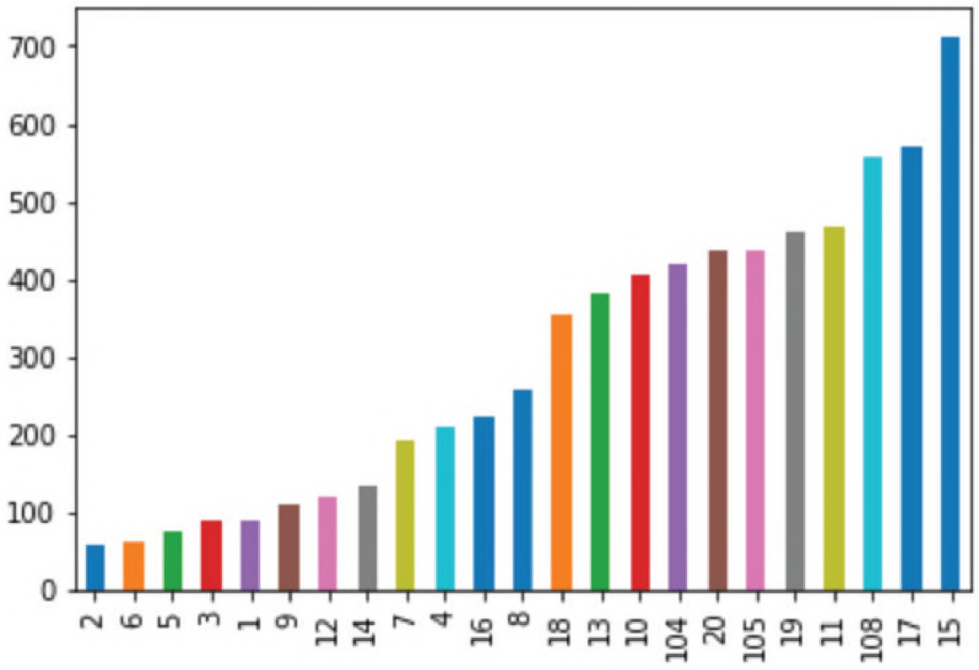

Figure 1 – Percentage of missing metrics in the file

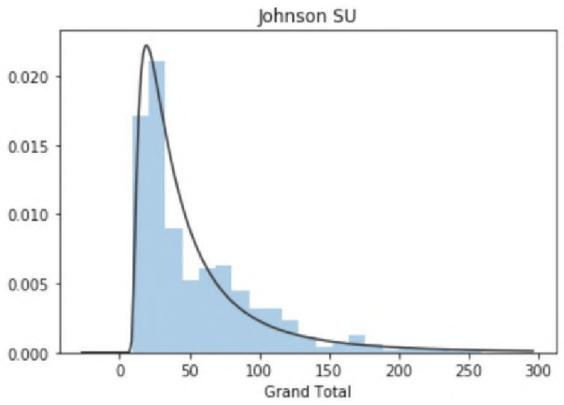

Figure 2 – Johnson SU distribution

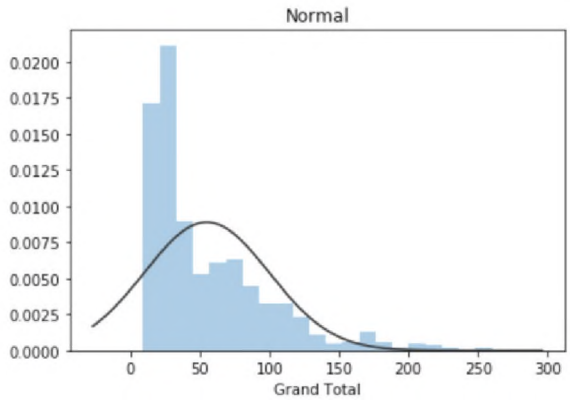

Figure 3 – Normal distribution

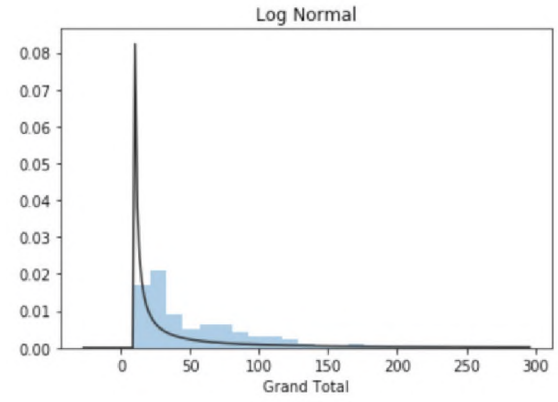

Figure 4 – Log Normal distribution

Johnson's SU-distribution for a parameter of a family of probability distributions was first investigated by L. N. Johnson in 1949. Johnson proposed it as a transformation of the normal distribution. [7]:

$$z = \gamma + \delta \sinh^{-1}\left(\frac{x-\xi}{\lambda}\right)$$

Where $z \sim \mathcal{N}(0,1)$.

The normal distribution of data is a pattern of occurrence of its different values.

The lognormal distribution in probability theory is a two-parameter family of absolutely continuous distributions. If a random variable has a lognormal distribution, its logarithm has a normal distribution.

It is obvious that Grand Total does not follow a normal distribution, so it must be transformed before regression can be performed. While the log conversion works pretty well, Johnson's unrestricted distribution works best.

In Figure 5,6,7,8 at points 1, 2, 3, 4 that is, at street intersections for points 1 - Seifullin – Dulatov; 2 - Alfarabi – Markov; 3 - Abay–Tlendiev; 4 - Gorky Park; values range from 0 to 750, where the x-coordinate is the value of air pollution(PM2.5), the y-coordinate percentage of the occurrence of values. As can be seen at 1-4 points, the percentage of occurrence of the value of 40-2%, 50 -17.5% , 50 - about 2%, 50 –35%, respectively. It is worth noting that values greater than 360 are available at point 2.you can see that Grand Total does not follow the normal distribution, so you must convert it before you can perform the regression. While the log conversion works pretty well, Johnson's unrestricted distribution works best.

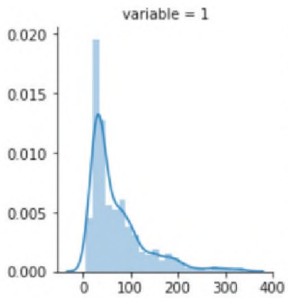

Figure 5 – point 1

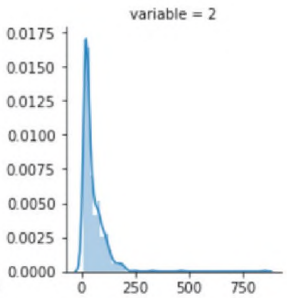

Figure 6 – point 2

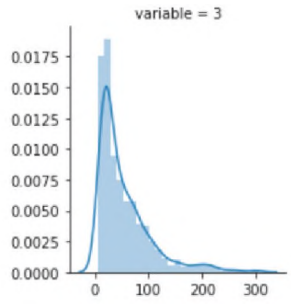

Figure 7– point 3

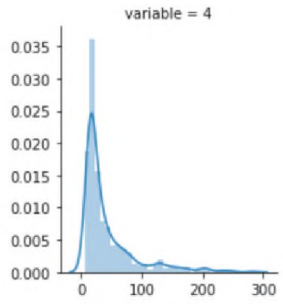

Figure 8 – point 4

In Figure 9,10,11,12 at points 5, 6, 7, 8 that is at the intersection of point 5-Tolebi - Baizakov; 6 - Rozybakiev - baykadamova; 7 - the Kok Kainar; 8 - Ryskulov - Momyshuly; values range from 0 to 600. Values greater than 300 are less likely to occur. From these 4 points we can say that values from 40-50 occur about 2%.

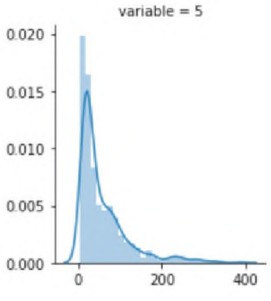

Figure 9 – point 5

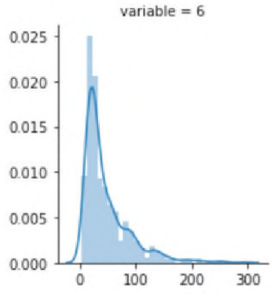

Figure 10 – point 6

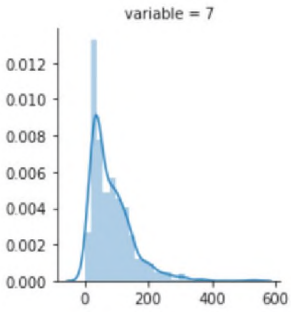

Figure 11 – point 7

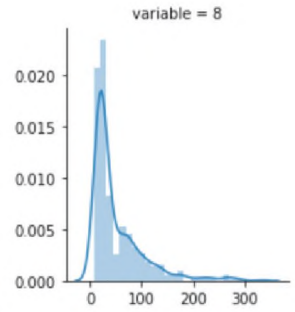

Figure 12 – point 8

In Figure 13, 14, 15, 16 at points 9, 10, 11, 12 that is, at street intersections for points 9 - Ermensay; 10 - Tulebaeva-Dzhambul; 11 - Askarova; 12 - Kamenskoye plateau; values range from 0 to 400 maximum. But more than 200 isolated cases. At point 12, mostly values between 10 and 45 occur above 3%.

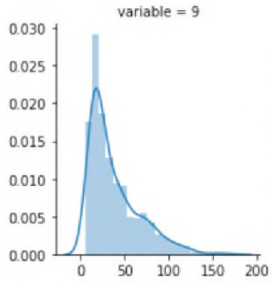

Figure 13 – point 9

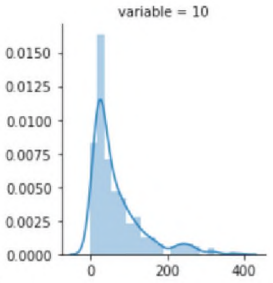

Figure 14 – point 10

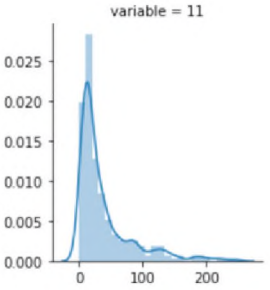

Figure 15 – point 11

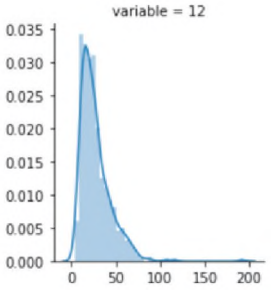

Figure 16 – point 12

In Figure 17, 18, 19, 20 at points 13, 14, 15, 16 that is, at the intersection of streets for points 13 - Furmanova-Tashkent; 14 - Mamyr; 15 - LCD "ASYL Arman"; 16 - Jean Kuat; Obviously immediately catches the eye point 15, where the value reaches as much as 10%.

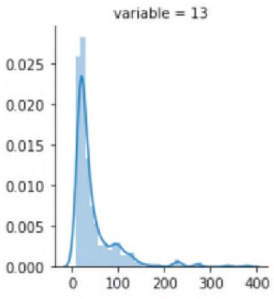

Figure 17 – point 13

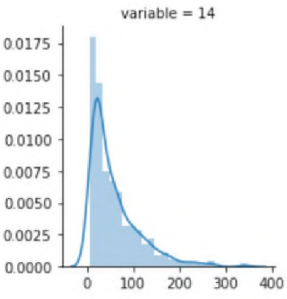

Figure 18 – point 14

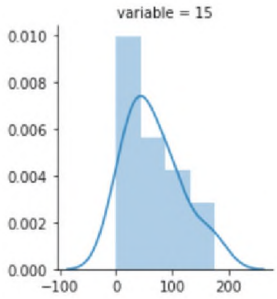

Figure 19 – point 15

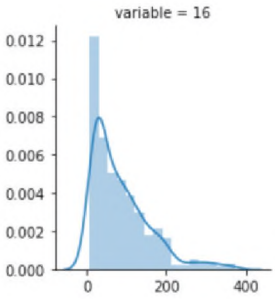

Figure 20 – point 16

In Figure 21, 22, 23, 24 at the points 17, 18, 19, 20 that is, at intersections of streets, for points 17 - Baganashil; 18 - Kyrgauldy; 19 - a Military camp; 20 - 8 - microdistrict Karasu; At the point 19 the percentage of occurrence of the values is smallest than the other graphs.

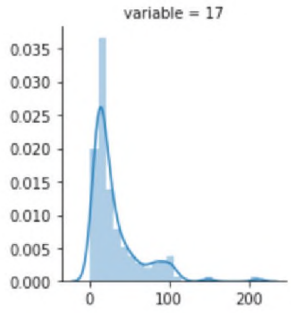

Figure 21 – point 17

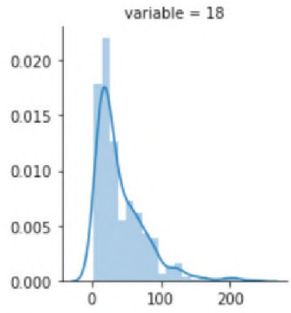

Figure 22 – point 18

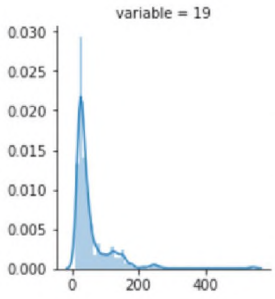

Figure 23 – point 19

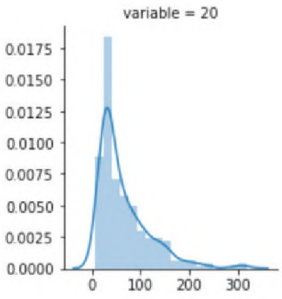

Figure 24 – point 20

In Figure 25, 26, 27 at points 104, 105, 108 that is, at intersections of streets for points 104 - Satpayev-Lugansk; 105 - Abay; 108 - Zhanna Kairat; At point 108 percent of occurrence of values the greatest than in other graphs.

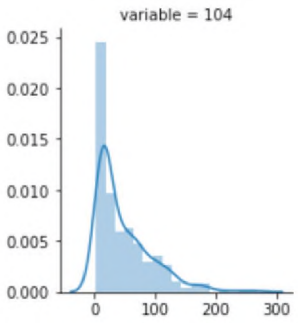

Figure 25 – point 104

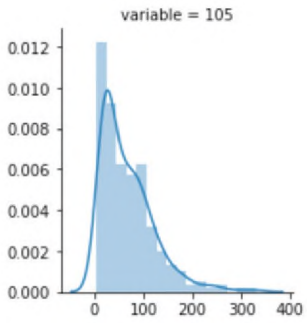

Figure 26 – 105

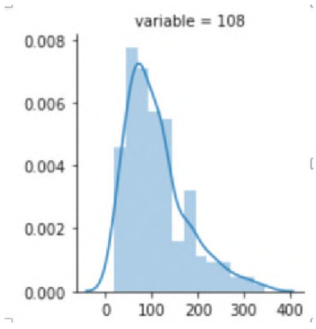

Figure 27 – point 108

In Figure 28 the Grand Total graph is an average value indicator for all points. Here we can see the average value between 10 to 50 meets 2%, and the from 50 to 80 meets at 5%.

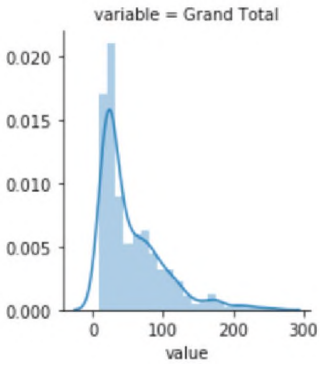

Figure 28 – Grand Total

DISCUSSION OF RESULTS

Mass concentration of PM2.5 is a key parameter for assessing air quality and its threat to human health. According to the norms of the world health organization (who) the average annual level of PM2.5 shall not exceed 10 μg/m3 and the average daily level shall not exceed 25 μg/m3.

Knowing the above standards and shown the following indicators of the city of Almaty for the last 2 years (2017(from 22 March)-2018-2019(to 25 March)) with the average, after the analysis we can draw the following conclusion:

As we can see in Table 1 have shown in every column number of values, mean value, standard deviation, minimum value, maximum value, each cell has 25, 50, 75 percent of the corresponding data.

Table 1. The arithmetic mean, maximum and minimum values, standard deviation

	1	2	3	4	5	6	7	8	9	10
count	639.000000	670.000000	641.000000	518.000000	654.000000	669.000000	537.000000	472.000000	618.000000	325.000000
mean	71.210139	51.858561	54.578416	42.460478	57.493547	47.461128	83.722718	51.139981	37.697744	68.394826
std	58.804719	54.739380	49.963944	45.358275	59.671261	42.403176	65.036166	51.922523	28.752380	66.959850
min	7.000000	5.000000	6.000000	5.000000	5.000000	3.600000	0.000000	7.000000	5.000000	0.000000
25%	30.350000	20.000000	20.000000	15.000000	18.000000	20.000000	35.000000	19.000000	16.600000	22.000000
50%	47.000000	32.000000	37.000000	24.000000	33.513649	30.267733	64.000000	27.000000	27.000000	41.000000
75%	92.100000	72.000000	73.000000	50.000000	77.000000	61.448923	115.000000	69.000000	51.000000	90.000000
max	340.000000	850.900000	304.000000	283.000000	387.000000	292.000000	530.000000	331.000000	170.600000	380.000000

Table 2. Mean values for every column

```
(729, 23)
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
14   :  58.84347897166664
15   :  66.398498331875
16   :  85.5542956104753
17   :  31.728662420382157
18   :  42.476366717759994
19   :  54.98243131453536
20   :  64.44527031788395
104   :  48.203698908774186
105   :  67.01166209601377
108   :  109.48883671872834
```

From Table 2 we can conclude the following:

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

And the more pure air are the points:

1) Point 12(Kamenskoye plateau) with a value of 27.4

2) Point 17 (Baganashyl) with a value of 31.7

3) Point 9 (Ermensay) with a value of 37.6

CONCLUSION

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 paragraphs, 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

REFERENCES

1. Chung, Y., Dominici, F., Wang, Y., Coull, B. and Bell, M. (2015). *Associations between Long-Term Exposure to Chemical Constituents of Fine Particulate Matter (PM 2.5) and Mortality in Medicare Enrollees in the Eastern United States*. Environmental Health Perspectives, 123(5), pp.467-474.
2. William M. Hodan, and William R. Barnard, (n.d.). *Evaluating the Contribution of PM2.5 Precursor Gases and Re-entrained Road Emissions to Mobile Source PM2.5 Particulate Matter Emissions*.
3. Phipps, J., Aronoff, D., Curtis, J., Goel, D., O'Brien, E. and Mancuso, P. (2010). *Cigarette Smoke Exposure Impairs Pulmonary Bacterial Clearance and Alveolar Macrophage Complement-Mediated Phagocytosis of Streptococcus pneumoniae. Infection and Immunity,* 78(3), pp.1214-1220.
4. Orru, H., Maasikmets, M., Lai, T., Tamm, T., Kaasik, M., Kimmel, V., Orru, K., Merisalu, E. and Forsberg, B. (2011). *Health impacts of particulate matter in five major Estonian towns: main sources of exposure and local differences.* Air Quality, Atmosphere & Health, 4(3-4), pp.247-258.
5. Grigorieva I.A., *Subsystem of analysis of data and machine training for information and analytical system ecohealth [Podsistema analiza dannikh I mashinnogo obucheniya dlya informatsionno-analiticheskoy sistemy ecohealth].* (2017). Student, [online] 5(5), pp.40-46. Available at: https://sibac.info/journal/student/5/75629.
6. AirPaca. (2019). *Association de surveillance de la qualité de l'air agréée par le ministère de l'environnement.*. [online] Available at: https://www.airpaca.org/ [Accessed 23 Apr. 2019].
7. En.wikipedia.org. (2019). *Johnson's SU-distribution*. [online] Available at: https://en.wikipedia.org/wiki/Johnson%27s_SU-distribution [Accessed 23 Apr. 2019].