Statistical analysis of the impact of economic activity on the environment and the population in the regions of the Russian Federation

Lyudmila Borisova *, *Irina* Sedykh, and *Marina* Khripunova
Financial University under the Government of the Russian Federation, 125993 Moscow, Russia

Abstract. Rosstat data on the dynamics of major socially significant diseases have been studied. The relationship between such diseases has been investigated. The influence of the main socially significant diseases and the main indicators of socio-economic development of the regions of the Russian Federation on the number of abortions has been studied. The most informative signs related to socially significant diseases and socio-economic development in all countries of the world have been selected. The influence of these signs on infant mortality has been studied. Machine learning methods collected in the Data Master Azforus (DMA) program were applied. The conducted research has demonstrated the effectiveness of using machine learning methods to identify patterns linking the frequency of socially significant diseases and indicators of socio-economic development.

1 Introduction

Environmental protection is inextricably linked with reducing damage to the environment, as in industry and in everyday life, materials have recently been increasingly used, which require considerable time to completely decompose. Soil, water, and air are polluted, which negatively affects the life and health of the population [1]. It is worth noting that the environmental problem worsened in the fifties of the last century, primarily due to the industrial revolution in the developed countries of the Old World. The emergence of new materials, unwillingness to invest money in sewage treatment plants and similar examples of predatory economic activity aimed primarily at maximizing profit, together with ineffective government policies related to public health protection, lead to an increase in morbidity, both general and infectious.

Lifestyle-related diseases pose a threat to the socio-economic aspects of life in countries around the world, and appropriate action to treat them is an urgent need [2]. The treatment of lifestyle-related diseases includes proper diagnosis, screening and treatment of these diseases in addition to providing palliative care to people who need it. High-quality treatment of lifestyle-related diseases should be based on a primary health care approach, in which early detection and appropriate treatment are prioritized [3]. It was found that in the long term, an increase in drinking water pollution from the distribution network and atmospheric air has a

* Corresponding author: lrborisova@

statistically significant effect on the growth of congenital anomalies (malformations), deformities and chromosomal abnormalities in children 0-14 years old [4]. Based on the conducted research, the authors conclude that the incidence of malignant neoplasms is associated with a complex of various factors (from anthropogenic impact to climate change).

The review article [5] lists all 9 diseases that are socially significant in accordance with the Decree of the Government of the Russian Federation dated December I, 2004 No. 715. The list includes the following diseases: diseases caused by HIV, tuberculosis, viral hepatitis B and C, sexually transmitted diseases, diabetes mellitus, malignant neoplasms, mental disorders and behavioural disorders, hypertension — a total of 9 diseases. In 2007, the Federal Target Program "Prevention and control of socially significant diseases (2007-2012)" was adopted. In the following years, similar programs were adopted and implemented by the regional authorities. As a result, the primary incidence in Russia decreased in 2018 compared to 2005: active tuberculosis — by 1.9 times, syphilis - by 4 times. The number of patients taken under observation with a diagnosis established for the first time in their lives also decreased: for mental disorders - by 1.4 times, for alcoholism - by 2.8 times, and for drug addiction - by 1.7 times. And over these 13 years, an increase in the primary incidence of non-communicable diseases was recorded: hypertension by 1.9 times, diabetes mellitus by 1.4 times, malignant neoplasms by 1.2 times and infectious diseases caused by HIV by 2.5 times. The reasons for this multidirectional dynamics have been identified. Regional differences in the prevalence of socially significant diseases were studied by the number of patients registered in medical institutions. In 2018, 10 subjects of the Russian Federation with the largest and 10 subjects with the least number of patients were identified.

The timely and high-quality treatment of various diseases, including socially significant ones, is influenced by the level of professionalism of doctors, not least. It is important to have information about this. In [6], after the conducted research (online questionnaire), it is emphasized that more than 40% of patients prefer to solve their problems by receiving information on the Internet, and 31% of respondents did not understand their diagnosis after contacting doctors in private clinics. Thus, people's distrust of the healthcare system plays an important role in the prevalence of infectious diseases.

The role of social networks in obtaining information about the spread of epidemics was studied in [7]. This work highlights that recent outbreaks of avian influenza across Europe have highlighted the potential of syndrome surveillance systems that take into account other data collection methods, namely social media. This study explores the possibility of using social media, primarily Twitter, to monitor outbreaks of diseases such as avian flu. The method of time series analysis was used in the work. The results show a weak, statistically insignificant relationship between the number of tweets and confirmed cases in a temporal context, implying that relying simply on social media data for observation may not be enough. Spatial analysis provided insight into the coincidences between confirmed cases and tweet locations, shedding light on regional events during breaks. While social media can be useful for understanding public sentiment and concerns during outbreaks, this should be combined with traditional surveillance methods and official data sources for a more accurate and comprehensive approach. However, the authors conclude that improved data mining and real-time analysis techniques can further improve flare detection and response.

A significant role in the spread of infections is played by the movement of people by public transport and crowds during mass events. Special attention should be paid to the spread of infections among children. It is noted in [8] that children traveling with chronic and complex diseases represent a diverse group of vulnerable travellers. Addressing knowledge gaps about how best to help these travellers requires a comprehensive approach. Research is urgently needed to identify the best treatments for the five most common chronic childhood diseases: asthma, depression, attention deficit hyperactivity disorder (ADHD), food allergies and autism.

The authors [9] pay attention to the use of statistical analysis methods in assessing the health risk caused by air pollution. Based on the principles of evidence-based medicine, all risks should be comprehensively analyzed and minimized using modern methodological approaches, taking into account their capabilities and limitations.

To model infections, the apparatus of differential equations, both ordinary and partial derivatives, is usually used. In particular, the classical use of differential equations was demonstrated in [10] when modeling Covid-19.

Currently, with the rapid development of statistical methods of analysis and forecasting, machine learning methods are increasingly used for data processing [11]. This work highlights that by using state data and on-site feedback, data models can be trained using machine learning and statistical concepts.

Machine learning methods have previously been successfully used by us [12] to study the relationship between road deaths and the state of the transport system in the regions of the Russian Federation, between the incidence of Covid-19 and the socio-economic development of the regions of the Russian Federation [13,14].

2 Materials and methods

Materials. The open data of Rosstat on indicators of the socio-economic development in 88 regions of the Russian Federation are used for the study. To study the relationship between the indicators of the socio-economic development data for 2022 were selected. The World Bank's open data was also used.

Methods. The calculations used an algorithm implemented in Microsoft Excel (data Analysis \ Regression...).

For statistical data processing the original method by Senko and Kuznetsova is statistically weighted syndromes (SHS) was used [15,16].

The SHS method is based on the procedure of weighted voting on systems of so-called syndromes – areas of the feature space containing mainly objects of one of the classes.

The ROC AUC value was used as an integral measure of the differences between classes according to indicators of the socio-economic development and socially significant diseases. Sliding control (Leave-One-Out) was used as a validation method.

3 Results. Using the methods of basic statistics and machine learning for comparative analysis.

First for the study, we selected indicators that are common to all countries. These include: effective treatment of tuberculosis (%), the percentage of households in which health care costs exceed 10%, the rate of prenatal care as the percentage of pregnant women who visited a doctor at least four times before giving birth, immunization against measles, pneumonia, tetanus and rotaviruses after birth (in%), the proportion of people using basic and safe sanitation. The data for the calculations are taken from the World Bank website.

First, the Mann-Whitney-Wilcoxon statistical test was performed. When dividing all countries into two groups (above and below the median mortality rate, the median is 8 deaths per 1,000 births), the following indicators turned out to be statistically insignificant: the percentage of households in which health care costs exceed 10%; immunization against rotaviruses after birth (in %). For the other p-value indicators mentioned above, the value is almost equal to 0 (table 1).

The use of sections to divide the text of the paper is optional and left as a decision for the author. Where the author wishes to divide the paper into sections the formatting shown in Table 1 should be used.

Table 1. P-value values calculated using the Mann-Whitney-Wilcoxon test when dividing all countries into 2 groups using indicators critical for socially significant diseases.

Indicator	**p-value**
The level of poverty	0,0000
Immunization at birth against tetanus	0,000365
Effective treatment of tuberculosis	0,000034
Prenatal care	0,000
The second measles vaccine	0,000
Protection against pneumococcal infection for up to a year	0,0067
Percentage of the population using basic sanitation	0,00
Percentage of the population using safe sanitation	0,00
Mortality among newborns	0,00

When using the method of statistically weighted syndromes, pure control was used as a recognition quality check, when 20% of the total population was recruited into the sample. At the same time, countries with a mortality rate among new-borns above the median, that is, African countries, first of all, were correctly recognized with a 100% result. The recognition quality is characterized by the area under the ROC curve. In Fig.1. The ROC curve is presented in the case of using methods of statistically weighted syndromes. The most significant was the correlation coefficient between mortality rates and the percentage of people using basic sanitation services. It turned out to be equal to -0.81. The less access there is to basic sanitation, the higher the mortality rate. Note that 155 countries were used in the processing.

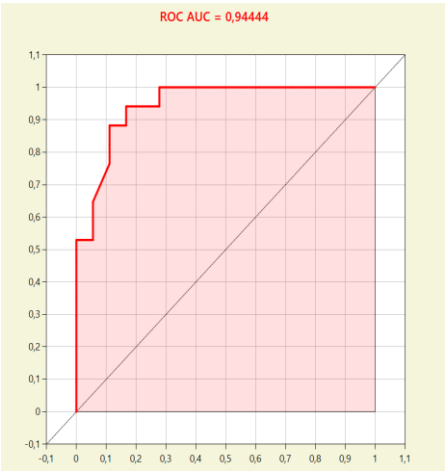

Fig. 1. ROC AUC for recognizing countries by the main indicators critical for the prevention and treatment of socially significant diseases.

The performed data analysis once again convinces that it is not easy to achieve results in a complex dynamic world. Without progress monitoring systems that allow timely course adjustments, testing and evaluating necessary innovations, as well as evaluating results, it will be difficult to achieve the Sustainable Development Goals. To choose the best strategy in choosing forms of support for the most vulnerable segments of the population, it is necessary to analyze not only the data presented on the websites of major international organizations such as the World Bank and the World Health Organization, but, above all, data from regional banks aimed at financing specific programs that not only improve the ecology of regions, but also contribute to earnings the population, so that there is an opportunity not only to meet basic needs, for example, in basic sanitation, but also for confidence in the future.

As the analysis presented in the paper showed, the indicators of prenatal care and the ability to spend no more than 10% of the family budget on medical services are not critical in influencing newborn mortality, which can be considered a basic indicator of poverty (in addition to the generally accepted UN indicator "the proportion of people living per day is less than $2.15).

Based on the statistical study conducted above using machine learning methods, we are convinced that the study of the turnover and use of fresh water, as well as the discharge of contaminated wastewater into surface water bodies, is necessary from the point of view of studying the possible relationship between the morbidity of the population and the environmental indicators listed above.

Let's analyze the latest data from Rosstat using machine learning methods in order to find out whether there is a stable statistical relationship between morbidity and the state of the environment in all regions of the Russian Federation for which there are published open data.

First, consider the following indicators as environmental indicators: discharge of surface wastewater into surface water bodies both per unit area and per capita, the volume of recycled and consistently used water in the same relative units, so that regions can be compared by these indicators. In addition, we will consider data on the use of fresh water in m^3/ha and m^3 per capita, as well as data on atmospheric emissions from stationary sources also per unit area and per capita. Let's present the results of regression analysis for data on 88 regions (Arkhangelsk and Tyumen regions are considered together and without autonomous districts included in them) (Fig.2).

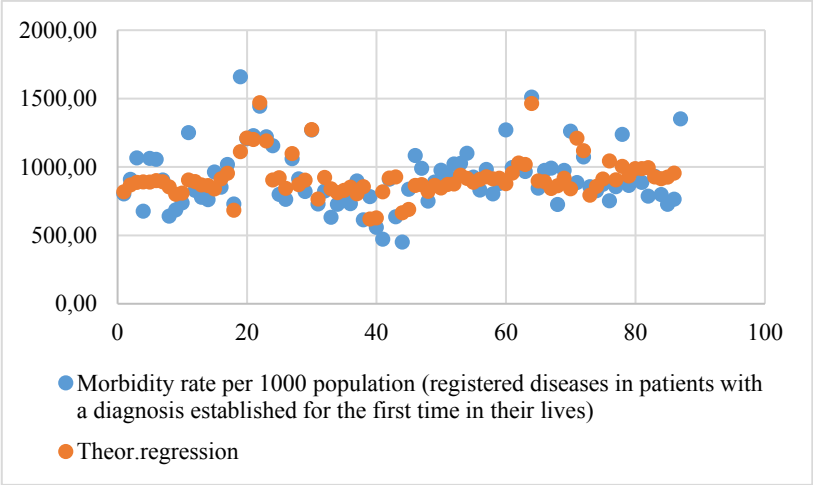

Fig. 2. Dependence of morbidity per 1000 population on environmental indicators.ROC AUC calculated using the method of statistically weighted syndromes to recognize regions with an increased number of abortions.

It should be noted that the statistically significant multiple correlation coefficient is 0.66, which indicates of a moderate relationship between the state of public health and the state of the environment.

It is interesting to look at the results of using machine learning methods, since the number of indicators and the number of regions are sufficient for recognition.

There is no doubt that environmental protection costs are closely related to the environmental situation in the regions of the Russian Federation, therefore, to begin with, we will divide the 88 studied regions into two groups: with environmental spending below and above the median level of 4.25 thousand rubles per person per year (population, as well as expenses environmental impact – according to Rosstat data for the end of 2022).

Fig. 3 shows the quality of recognition using statistically significant syndromes. Since the ROC AUC value turned out to be 0.77, the recognition quality is good. Note that the remaining recognition quality indicators are as follows: accuracy – 0.72; accuracy – 1.00; sensitivity – 0.44 specificity – 1. Regions belonging to the second group with an indicator of environmental protection costs above the median level were unmistakably recognized.

Note that when using the method of statistically weighted syndromes, 70 regions were selected for the training sample (randomly), 20% of the regions were selected for recognition, namely 18, and 9 regions were selected for each class.

Let's look at how environmental indicators affect the morbidity of the population.

Fig. 4 shows the recognition quality in this case. The incidence rate per 1000 population was chosen as the grouping one. The 8 environmental indicators listed above, which were used to build a regression dependence, were considered as dependent features.

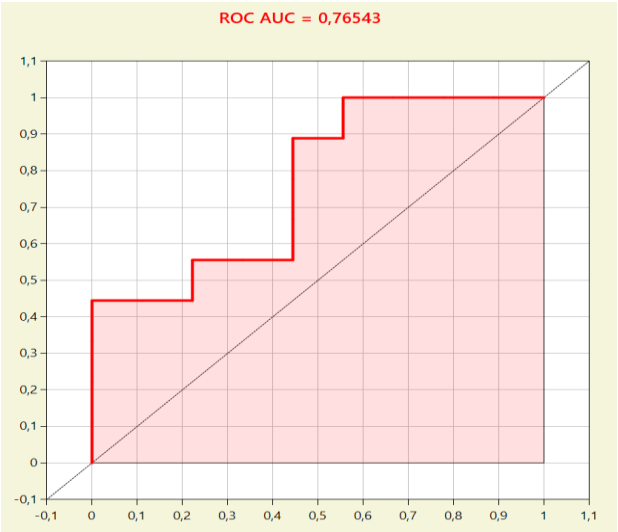

Fig. 3. A ROC curve calculated using the method of statistically weighted syndromes for recognizing regions by environmental protection costs.

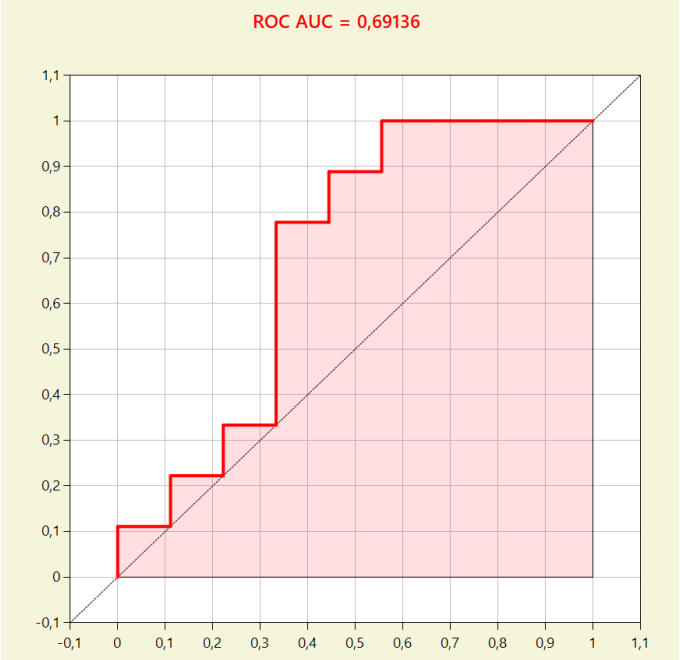

Fig. 4. A ROC curve calculated using the method of statistically weighted syndromes for recognizing regions based on the influence of ecology on the morbidity of the population.

4 Conclusion

The interest in the relationship between growth and equity has a long tradition in economics. The analysis carried out in this paper suggests that data on immunization of the population of poor countries in Africa and Asia, presented much more fully than similar data for developed countries, are redundant and are not informative in identifying the most critical factors determining living conditions in different countries.

It is almost impossible to put together environmental indicators to analyze their impact on living standards in all countries at the same time. In conclusion, we note that after the statistical analysis performed using machine learning methods, the most critical indicator affecting infant mortality was determined (among all indicators presented on the website of the World Health Organization). This is an indicator of the availability of basic sanitation.

There is enough fresh water in the Russian Federation, as evidenced by Rosstat data, but a small number of funds are still spent on environmental protection, modernization of wastewater treatment plants, and prevention of wastewater entering open reservoirs, as evidenced by the low median level of this indicator, calculated according to Rosstat data for 2022. The need to review environmental policy is also evidenced by the results of regression analysis and machine learning. A good recognition of regions was obtained by the incidence rate associated with the diagnosis registered for the first time), depending on the state of ecology in the region.

The results obtained using statistical analysis methods, including machine learning methods, allow us to conclude that it is socio-economic factors, primarily related to environmental protection, that play an important role in the prevention of socially significant diseases.

References

1. L. R. Borisova, Yu. V. Podrezov, Emergency safety issues **2**, 67-72 (2020)
2. S. A Tabish, Journal of Cardiology & Current Research **9**, 3 (2017)
3. Z. Zagdin, Y. Zhao, V. Tsvetkov, S. Sleptsova, M. Vinokurova, E. Sokolovich, P. Yablonskiy, International Journal of Circumpolarr Health **80**, 1 (2021)
4. P. A. Korotkov, A. B. Trubyanov, A. I. Gismieva, A. A. Avdeeva, E. V. Zagainova, Human ecology **29**, 8 (2022)
5. E. V. Budilova, L. A. Migranova, Population **23**, 2 (2020)
6. S. Jerotic, E. Ivancajic, Porto Biomedical Journal **2**, 5 (2017)
7. S. Munaf, K. Swinger, F. Brulisauer, A. O'Hare, G. Guhn, A. Reeves, One Health **17**, 100657 (2023)
8. S. E. Kohl, E. D. Barnett, Travel Medicine and Infectiones Disease **34**, 101438 (2020)
9. A. O. Karelin, A. Lomtev, M. Volkodaeva, G. B. Yeremin, Hygiene and Sanitation **19**, 1, (2019)
10. A. Menon, N. K Rajendran, A. Chandrachud, G.Setlu, *Modelling and simulation of COVID-19 propagation in a large population with specific reference to India*, MedRxiv preprint (2020)
11. S. Khan, T. Yairi, Mechanical Systems and Signal Processing **107** (2018)
12. L. Borisova, G. Zhukova, E3S Web Conf. **371**, 05003 (2023)
13. L. Borisova, G. Zhukova, A. Kuznetsova, Y. Kuznetsova, Lecture Notes in Networks and System **574**, 2648 (2023)
14. A. Kuznetsova, O. Senko, E. Voronin, O. Kravtsova, Yu. Kuznetsova, L. Borisova, G. Zhukova, I. Khrapunova, V. Akimkin, E3S Web Conf. **371**, 05003 (2023)
15. O. A. Senko, A. V. Kuznetsova, Pattern Recogn. Image Anal **20**, 2, (2010)
16. O. V. Senko, D. S. Dzyba, E. A. Pigarova, L. Y. Rozhinskaya, A. V. Kuznetsova, Short paper in Proceedings of KDIR **437** (2014)