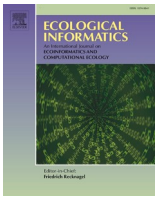
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Ecological Informatics 63 (2021) 101284

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Numerical analysis of factors, pace and intensity of the corona virus (COVID-19) epidemic in Poland

Piotr Andrzej Kowalski [a,b,*], Marcin Szwagrzyk [c,d], Jolanta Kielpinska [e], Aleksander Konior [d], Maciej Kusy [f]

[a] Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Cracow, Poland
[b] Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01–447 Warsaw, Poland
[c] Faculty of Geography and Geology, Institute of Geography and Spatial Management, Jagiellonian University, Gronostajowa 7, 30-387 Cracow, Poland
[d] Airly Inc., ul. Mogilska 43, 31-545 Cracow, Poland
[e] Department of Aquatic Bioengineering and Aquaculture, Faculty of Food Science and Fisheries, West Pomeranian University of Technology in Szczecin, Poland
[f] Faculty of Electrical and Computer Engineering, Rzeszow University of Technology, al. Powstancow Warszawy 12, 35-959 Rzeszow, Poland

ARTICLE INFO

Keywords:
Corona virus
COVID-19
Disease curve
Epidemiological model
Excess mortality
Pearson's correlation
Multivariate linear regression
Factor analysis
Least-squares estimation

ABSTRACT

This article focuses on a statistical analysis of the corona virus disease 2019 (COVID-19) data that appeared until November 31, 2020 in Poland. The studied database, expressed in terms of both population and air pollution (particulate) indicators, is provided mainly by the Airly company, the Central Statistical Office (GUS) and the Rogalski project. The particular measured factors, which underwent standardization, were assessed for mutual dependency by means of a Pearson correlation coefficient and analysed by a linear regression. Based on the presented models, our results indicate that air quality (air pollution level) is the most important factor in the context of enabling COVID-19 case load increase in Poland.

1. Introduction

Coronaviruses are enveloped viruses whose genome is positive-stranded RNA (+ ssRNA). Until the appearance of a new virus called COVID-19, two viruses in this group were known to cause respiratory infections. The first is SARS-HCoV (Severe Acute Respiratory Syndrome), which is responsible for severe lower respiratory tract infections. Transmission takes place by droplets and the mortality rate reaches 10% (Pancer, 2020). The second known virus from this group, which also infects the human respiratory tract, is MERS (Middle East Respiratory Syndrome). This virus spreads by airborne droplets as well. It causes fever, coughing and a shortness of breath that can turn into pneumonia. The MERS virus is characterized by a much higher (up to 50%) mortality rate (Ahasan et al., 2013). Coronavirus disease 2019 (COVID-19) belongs to the group of viruses that infect many species of animals, including humans. Due to limited human contact with the first species found to be COVID-19 susceptible, the bat, there is a scientific hypothesis that the virus was transmitted to humans from domesticated animal species (WHO, 2020). COVID-19 infection in humans causes a number of clinical symptoms, including: nasal congestion, runny nose, changes in smell and taste, fever, cough, fatigue, muscle pain, changes in chest computed tomography, lack of appetite, nausea, vomiting, decrease blood saturation below 94% and rapid breathing (Inglot et al., 2020; Lovato and de Filippis, 2020). In April 2020, children also had a rash on the skin of (mainly) the upper limbs in the form of red, itchy blisters.

In December 2019, the World Health Organization (WHO) issued a message on new cases of human disease in the city of Wuhan (China) linked to a coronavirus infection labelled COVID-19. In Poland, the first official appearance of the patient "0" was dated March 4, 2020, although, as indicated by Mostowy's analyses (Mostowy, 2020), the epidemic in Poland actually first appeared in the second half of January 2020.

Our own observations suggest that the current coronavirus epidemic

* Corresponding author at: Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Cracow, Poland.
E-mail addresses: pkowal@agh.edu.pl, pakowal@ibspan.waw.pl (P.A. Kowalski), m.szwagrzyk@airly.org (M. Szwagrzyk), jolanta.kielpinska@zut.edu.pl (J. Kielpinska), a.konior@airly.org (A. Konior), mkusy@prz.edu.pl (M. Kusy).

https://doi.org/10.1016/j.ecoinf.2021.101284
Received 1 February 2021; Received in revised form 18 March 2021; Accepted 18 March 2021
Available online 29 March 2021

in Poland is somewhat underestimated. As the data show, since the appearance of the patient "0", the number of diagnosed infections and deaths significantly differs from the averaged data from countries such as Belgium, Spain, Great Britain, Italy, France, Sweden and even the Czech Republic (Table in (Wazna, 2020)). Taking into account the diagnostic possibilities in hospitals and the dedicated COVID-19 diagnosis of coronavirus hospitals in the number of 21, special attention should be paid to the verification of people undergoing tests for the presence of viral RNA. In Poland, an assumption has been made that raises reservations that people with typical clinical symptoms (fever, dry choking cough, headache, dyspnoea) alone are referred for laboratory diagnosis. According to the authors, there is no justification for the mechanism of omitting in the diagnostics asymptomatic people sent to quarantine, who, in principle, may potentially act as vectors (carriers) in public space. This situation has come about because, for technical and financial reasons, hospitals in Poland do not have a sufficient number of tests that would not accidentally allow for the definition of infected people. Indeed, own observations indicate that many hospitals redirect samples for diagnostics to specialized diagnostic laboratories.

In the first wave of the epidemic, the bottleneck of diagnostics in Poland was the introduction of obligatory use by laboratories of the so-called certified reagent kits, the quantity of which was heavily limited as they could only come from one source. It was not possible to use rtPCR reaction components commonly produced by biotechnology companies, including individually ordered primers for the above-mentioned reaction. This is a standard mechanism used in diagnostic work by medical laboratories and scientific institutions. Therefore, diagnostics were performed only on the most symptomatically advanced cases, without the possibility of performing mass tests on people suspected of being infected, or regular testing of medical personnel. This method of verifying the scale of COVID-19 infection in Poland therefore did not give a real picture of the epidemic's development, which in turn would have disrupted the pattern and pace of its spread in Poland.

At the time of the second wave of the COVID-19 epidemic in Poland in September–November 2020, however, a strategy was adopted to redirect symptomatic patients to GPs (primary care physician) and to reintroduced social restrictions, such as bans on assembly or on eating inside restaurants and business/commuting only travel restrictions. Yet, according to the scientists of the Polish Academy of Sciences (Duszyński et al., 2020), there is no simple translation between the increase in infections in Poland and the increase in the public's sense of threat. This situation is not conducive to maintaining social restrictions during the second wave of the pandemic, and is manifested by "anti-mask" social groups, questioning of the restrictions imposed and even questioning the existence of COVID-19 disease. Since autumn 2020, due to the drastic increase in the number of cases and deaths in Poland (GUS - Poland's Central Statistical Office - data for four years), (Fig. 1), the population classification system has been changed to incorporate the results of testing for COVID-19.

Because of the daily number of deaths due to COVID-19 in Poland,

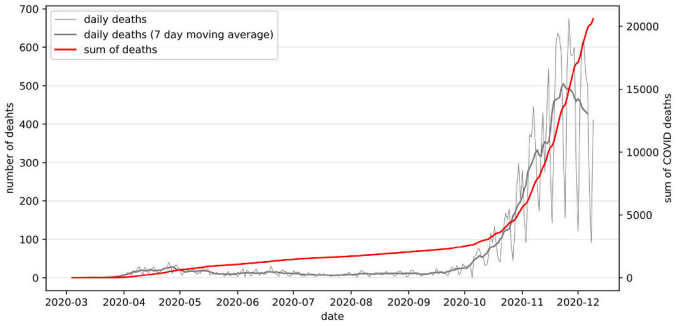


Fig. 2. Daily and total number of COVID-19 deaths in Poland.

which, according to the Ministry of Health (Fig. 2), has reached 650, decisions were also made to change the rules for referring patients to quarantine or isolation. At this point, it should be emphasized that in Fig. 2, which indicates the number of deaths, one observes quite significant fluctuations. Due to the fact that they are difficult to interpret, a line showing the number of deaths in the form of 7 day moving average has been included in the plot. The article (Ricon-Becker et al., 2020) attempts to explain these fluctuations as the effect of delayed reporting of these events. In addition, the increase in incidence recorded from September 2020 related to the second wave of the epidemic in Poland has resulted in first modification and then verification of the testing strategy. Three types of tests have started to be used on a massive scale: the SARS-CoV-2 virus RNA test, the antigen test and the serological tests. The last type of test is used in population monitoring studies in large cities because it gives a picture of the percentage of the population already exposed to COVID-19 as evidenced by antibody production.

Professor Andrzej Fal, head of the Department of Allergology, Lung Diseases and Internal Diseases at the Ministry of Interior and Administration Hospital (oral information), has reported that in the development of the epidemic in November 2020, Poland was among the top five countries with the highest number of cases per 1 million inhabitants. Yet, according to the data of the Ministry of Health (as of November 7, 2020), the daily number of infections has exceeded 27,000 patients - with a slightly decreasing trend over the next 20 days. With such an ineffective method of testing in relation to people with mild symptoms or with the completely asymptomatic taken out of the system, this is an unbelievable number (oral info, Prof. Tomasz Dzieciatkowski, virologist from the Medical University of Warsaw).

In Fig. 3, one can see the plot of the total number of COVID-19 tests performed in Poland and the fraction of positive results. Both of these represent the 7-day moving average. Based on observations and interviews with clinicians working in the covid wards, it can be assumed that the figures released regarding the number of daily infections detected by currently available testing (Fig. 3) are lower than in reality, because these do not include the hospitalizations of patients with moderate symptoms. These people have been treated outside the

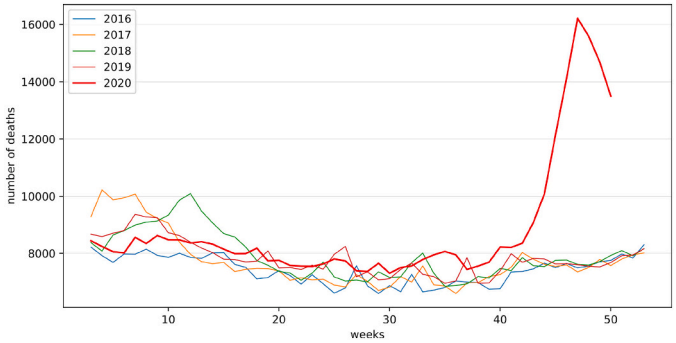


Fig. 1. The number of deaths in Poland by individual years: 2016–2020.

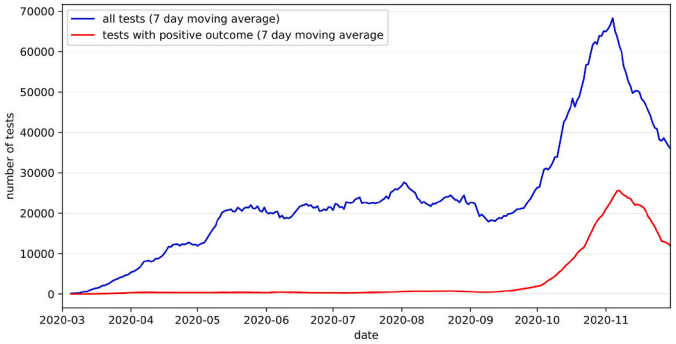


Fig. 3. Total number of COVID-19 tests in Poland.

hospital system and have not been subjected to diagnostic tests.

Many pharmaceutical companies are working on the vaccine, such as: Moderna Therapeutics, Inovio Pharmaceuticals, Novavax, Vir Biotechnology, Stermirna Therapeutics, Johnson & Johnson, VIDO-InterVac, GeoVax-BravoVax, Clover Biopharmaceuticals, CureVac, Codagenix and Pfizer/BioNTech. The appearance of approved COVID-19 vaccines (Pfizer / BioNTech and Moderna Therapeutics) in December 2020 has brought about a real opportunity to create an immune block. This mechanism also arises in the spontaneous and uncontrolled exposure of the population to the virus, through the natural selection of non-immune people and the acquisition of immunity through an increasing number of convalescents. In Poland, it is planned to vaccinate first of all people from the highest risk groups, i.e. medical personnel, people over 65 and people with diseases that increase the risk of a severe clinical course of COVID-19 infection. It should be assumed that the acquisition of herd immunity in the event of the commencement of mass vaccination at the turn of January and February 2021, will be possible around August of this year. The condition is vaccinating 70% of the entire population, because only this level of resistant people guarantees the development of herd immunity.

1.1. Tipping points in the epidemic chain

According to the data of the Ministry of Health in Poland, the main centres of the development of the coronavirus epidemic in Poland were hospital departments, nursing homes, schools and the so-called scattered outbreaks related to holidays and heightened population density in resorts at the turn of July and August. The worst situation in the first weeks of the spread of the epidemic in Poland was recorded in large agglomerations (Śląskie and Mazowieckie voivodships), and may have resulted from the population density and the presence of large communication nodes in these cities (Katowice, Warsaw). The commuting of a large number of people from the surrounding towns and villages to their work places and to places of learning was, hence, conducive to uncontrolled infections, especially in the first phase of the epidemic. This meant that from mid-January until the introduction of obligatory remote work and study by the government, asymptomatic people were spreading COVID-19 through most sectors of the economy and administration (including the schools, kindergartens and universities).

Of great importance in the epidemic chain was the lack of unambiguous provisions related to the treatment of patients with chronic diseases, injuries or strokes and heart attacks, who, as undiagnosed, were referred to "clean" wards, and were not treated as potentially infected. Probably this resulted in medical personnel in Poland being among the huge share of people infected with COVID-19. Indeed, according to the Chief Sanitary Inspector, 17% of the infected are medical personnel (Gabriela, 2020). For comparison, the same figure is 10.3% in Italy, 14.4% in Spain, 5.7% in Great Britain, 20% in Portugal and 4.4% in China. This means that in hospitals / clinics / emergency departments, the protection system for healthcare workers is insufficient. Thus, for medical workers redirected to working in the wards renamed 'covid', the lock system, the possibility of epidemiological path separation and non-inter-communication of people in contaminated and clean zones was ineffective.

1.2. Border traffic as an element of social quarantine

The Polish borders were closed on March 13, 2020, and a 14-day quarantine for people who work on the other side of the border was introduced. At the same time, in order to enable Poles working abroad to return home during the pandemic, in mid-March 2020, the government announced the "LOT to go home" program. As part of this project, 388 flights returning citizens from over 70 destinations on six continents were carried out in three weeks, and about 32,000 people returned to Poland during this time.

Border crossings have become a fairly widely criticized element of "introducing the pathogen" into the country. From an epidemiological point of view, the closure of borders with the simultaneous importation of such a large number of people from almost all continents in a short time contributed to the increase in the incidence of the disease in small towns and villages. This was probably due to the fact that people returning from a stay abroad "entered" the multi-generational homes where their parents or grandparents lived. Being from the greater risk zone (Germany, Italy, France), they introduced the coronavirus to local communities. The Border Sanitary Inspectorate, despite the fact that it is the institution responsible for such activities at the border, did not make decisions regarding the sanitary control of passengers on return flights to Poland. Lessons, were, however, learned. In autumn of 2020, due to the huge increase in infections, a ban on free movement in cross-border traffic was reintroduced (except for people working abroad and living in Poland on a daily basis). Limited to January 10, 2021, the possibility of crossing the border under the so-called 'One-day trips' designation (shopping in border regions and cross-border tourism) has had a very positive effect on reducing the appearance of new, uncontrolled outbreaks of COVID-19.

1.3. Air pollution as a factor influencing the course of the COVID-19 pandemic

The aim of the study is to show both the course of morbidity and mortality from COVID-19 in Poland in spatial terms and, above all, to focus on the factors influencing the above phenomena. The research tried to approach individual issues in an interdisciplinary manner. During the research, numerical analyzes were performed that directly result from the data. Efforts were also made to reach specialists who indicated the factors that should be taken into account in the possible interpretations of individual results. The latter were largely the result of many hours of talks and debates conducted by the co-authors of this publication. To a large extent, these studies were possible thanks to the very strong commitment of Airly (which has a virtually countless volume of data related to air pollution in Poland and around the world), and to the efforts of the space-time data specialists working there.

This article is composed as follows. The upcoming Section 2 introduces the problem of the pandemic from a general point of view and is a review of important contributions in this field. In Section 3, the statistical tools applied to the analysis conducted in the current work are described. Section 4 sets out a description of the investigated data set: its sources, nature and considered geographical models. In Section 5, the dependencies and the factors influencing the course of COVID-19 in Poland are numerically analysed. In the subsequent section, 6, the obtained results are discussed; in this part of the paper, we also provide a wide-ranging description of the background to the COVID-19 pandemic spread in Poland, and refer to the current state of the literature. The final Section 7 concludes the work.

2. Background literature review

Pandemics of respiratory disease in humans vary in nature, spread, and mortality. The most important pandemics of influenza in humans include the Spanish (of which almost 50 million people have died), Asian (avian) flu (1–4 million), Hong Kong (avian) flu (1–4 million) and the American (swine) flu (100–400 thousand) (Gliński and Żmuda, 2020). As the authors point out, the most common cause of an epidemic is a mutation in the genome of the pathogen and changes in the host's organism. In the case of influenza, the pathogenic factors are viruses of the subtypes H1N1, H2N2 and H3N2, which are transmitted through secretions from the respiratory system of an infected person. In the case of viruses from the Coronaviridae family, the most dangerous pandemics concerned the SARS-CoV and MERS-CoV viruses. The first is called Severe Acute Respiratory Syndrome coronavirus, the second is Middle East Respiratory Syndrome. The authors of (Masood et al., 2020) indicate the lack of a defined intermediate host for COVID19 transmission. The

epidemiological analysis shows that in the case of the Sars virus, the role of an intermediate link between bats (Chiroptera spp.) and man was played by a domestic animal, most likely a cat (*Felis catus*), and in the case of MERS, it was a dromedary (*Camelus dromedarius*). The SARS virus is an atypical pneumonia that first appeared in 2003 in the Guandong province of China and has spread to many countries around the world. The mortality from SARS was approximately 11%, with increased risk in patients over 60 years of age. For comparison, the MERS virus was first diagnosed in 2012 in the Middle East, from where it was transmitted to many European countries. Epidemiological data indicate that half of the patients infected with this type of virus developed severe respiratory diseases similar to clinical symptoms in infection with the SARS virus. According to (Hussain, 2014), mortality in the case of MERS infection may reach up to 40% of all infected patients.

As early as spring 2020, attention was drawn to the fact that regions particularly affected by COVID-19 (e.g. Lombardy in Italy) were characterized by high air pollution. Even during the first wave of the pandemic, a number of publications appeared, which, based on the examples of The Peoples Republic of China (PRC), the USA and Europe, showed the dependence of COVID-19 infection and mortality on air pollution.

For example, the authors of (Zheng et al., 2020) quantified the effect of air pollution on COVID-19 risk infection based on the historical data of air quality in the PRC and COVID-19 case reports. For this purpose, the data from air quality stations from January 2015 to March 2020 was used; it included the particulate matter concentrations of PM2.5, PM10, SO2, CO, NO2, and O3 for selected 324 cities. A generalized linear model was applied to discover the association between long-term exposure to air pollutants and the risk of COVID-19 infection. It was shown that there was a significant relationship between reported and severe COVID-19 cases and the concentration of historical air pollutant. In particular, in all cities with air quality measured, an increase of $10\mu g/m^3$ in NO2 and PM2.5 concentration was related to 22.41% and 15.35% increase in COVID-19 reported cases, respectively. In the case of severe COVID-19 cases, such an increase was equal to 19.20% and 9.61%, respectively.

The work (Wu et al., 2020) presented an investigation of the influence of long-term average exposure to fine particulate matter (PM2.5) on an increased risk of COVID-19 death in the United States. The data set was obtained from the Center for Systems Science and the Engineering Coronavirus Resource Center of Johns Hopkins University and comprised cumulative number of COVID-19 deaths in 3087 counties until April 22, 2020. The PM2.5 values were computed based on 2000–2016 period by averaging estimates in a given county; 19 county-level variables and one state-level variable were considered in the analysis. The association between the COVID-19 mortality rate and long-term PM2.5 exposure was explored with the use of a negative binomial mixed model, where COVID-19 deaths were selected as the outcome and PM2.5 as the exposure of interest. It was established that an increase of only $1\mu g/m^3$ in PM2.5 resulted in an 8% growth in the COVID-19 death rate.

In (Cole et al., n.d.), the dependence of a long-term air pollution exposure to COVID-19 in 355 cities in the Netherlands was studied. The data was obtained from the National Institute for Public Health and the Environment and covered (i) infected cases between February and June, and (ii) annual concentrations of PM2.5, NO2, and SO2 measures. The average concentrations of PM2.5, NO2, and SO2 were reported to be: 10.5, 15.8 and 0.8, respectively. A numerical analysis based on a negative binomial model revealed that there was a statistically significant positive association between air pollution (PM2.5 concentrations) and COVID-19 cases, hospital admissions and deaths. Expressed in numbers, an $1\mu g/m^3$ increase in PM2.5 concentrations contributed to a rise in COVID-19 infections over the value of 9.4, 3.0 more hospital admissions - and 2.3 more deaths.

The study conducted in (Travaglio et al., 2020) showed potential links between air pollution and COVID-19 in England and was based on a wide range of data from regional-level, subregional-level and individual-level resources. The data on patients infected with SARS-CoV-2 were obtained from Public Health England and the UK Biobank, while virus-related deaths was retrieved from the National Health Service and the Office for National Statistics. Air pollution data were provided by the European Environmental Agency (nitrogen dioxide, nitrogen oxide and ozone) and UK Air information resources (ozone, nitrogen oxides, PM2.5 and PM10). Here, a binomial regression model applied to UK Biobank data revealed that PM2.5 was a major contributor to COVID19 cases in England, i.e.: the increase of $1\mu g/m^3$ of PM2.5 was related with a 12% rising in COVID-19 cases. Moreover, a single-unit increase of PM10 involved approximately 8% more COVID-19 cases. Thus, after population density, nitrogen dioxide, nitrogen oxide and ozone levels constituted significant predictors of COVID-19-related deaths.

The authors of (Dong et al., 2021) present research that assessed the short-term impact of PM2.5, PM10, NO2, SO2, CO and O3 on respiratory admissions in Lanzhou, PRC (China). For this purpose, daily hospital admissions from the three largest hospitals in Lanzhou and daily air pollution concentrations were compiled during a 4-year period (2014–2017). To estimate the association of air pollutants on respiratory admissions considering the influence of different confounders such as seasons, sex, and age groups, a generalized additive model was applied. The outcome of this action was the observation that a $10\mu g/m^3$ increase in PM2.5, PM10, SO2, CO and O3 concentrations resulted in 0.89%, 0.33%, 3.01%, 3.20% and 0.73% growth in respiratory admission, respectively. No remarkable relationship was established between NO2 and respiratory disease medication.

The author of (Coro, 2020) introduced a means of establishing the infection rate of COVID-19 globally at a 0.5° resolution via a Maximum Entropy based Ecological Niche Model (ME-ENM). This model was designed to identify geographical areas as a potential that is subject to a high infection rate. It indicated the locations that could contribute to the increase of the infection rate by virtue of their particular geophysical (surface air temperature, precipitation, and elevation) and human-related characteristics (carbon dioxide and population density). ME-ENM was trained on high infection rate data from 54 Italian provinces (up to the end of March 2020) and then tested using datasets from World country reports. The application of ME-ENM allowed for the determination of a risk index capable of identifying countries and regions having a high risk of disease increase. The presented results implied that a complex combination of the selected parameters might play a crucial role in understanding the spread of COVID-19 among human populations, particularly in Europe (Coro, 2020).

The above-presented research immediately met with great interest in Poland, which has been for years, struggling with very poor quality air. However, the mild course of the first wave of the epidemic in Poland did not mobilize the scientific community to conduct in-depth research on this problem. It was only during the second wave that attention was drawn to the link between the concentration of disease and deaths in regions with high air pollution (the southern parts of upper Silesia and of Małopolska). At the same time, other factors that may be responsible for this phenomenon were analysed, and they also characterized the above-mentioned areas - such as, for example, high population density or the proximity of national borders.

3. Materials and methods

This section presents the tools and statistical models used to analyse the dependencies in this article. The first tool that was helpful in pre-processing the data was standardization (Williamson and Piattoeva, 2019). This is a kind of normalization that allows changing the scope of individual variables so that they are presented in a similar numerical range. It should be emphasized here that this transformation preserves the distance relations between individual data. Thus, suppose one have a set of records $x_j \in \mathbb{R}^n$, $j = 1, \ldots, m$ where n is the number of features. For each element $j = 1, \ldots, m$, the variables are transformed as follows:

$$x_{j,i}^{new} = \frac{x_{j,i} - m_i}{std_i}, \tag{1}$$

where: $x_{j,i}^{new}$ is the value of the variable $x_{j,i}$ after standardization with the given data, m_i is the arithmetic mean of the i feature and std_i is the standard deviation of this feature. It is worth emphasizing here that the standardization of data is independent for each of the considered features $i = 1, \ldots, n$. The data standardization procedure is carried out in most tasks that use tools from the domain of statistical analysis and machine learning. Standardization is not recommended for data that we know are not normally distributed. The factors analysed in this study were normalized according to the formula (1), therefore, they had values in a similar range and could be compared with each other.

Another statistical tool used in this article is correlation. This is a measure that allows the checking of whether two variables X and Y are related to each other by a linear relationship. The Pearson R (Benesty et al., 2009) coefficient is a measure of the correlation. It is calculated as follows:

$$R = \frac{Cov(X,Y)}{std(X)std(Y)}, \tag{2}$$

where $Cov(X,Y)$ is the value of the covariance function between the X and Y variables, and $std(X)$ is the standard deviation of the X variable. It is important to interpret the R coefficient. This takes real values in the range $[-1,1]$, and so, the extreme values, i.e. -1 and 1, indicate an ideal, full correlation between the variable X and the variable Y, with the first being a match and the second being out-of-phase correspondence. The value of the R coefficient equal to zero, is the worst option because it indicates a complete lack of dependence between the studied variables. Intermediate thresholds of R values can be assumed and the correlation can be defined as weak: $|R| < 0.2$, mean correlation $0.2 < |R| < 0.5$, strong correlation $0.5 < |R| < 0.7$, and $|R| > 0.7$ is a very strong correlation. Of course, the absolute value $|\ |$ is used in the above notation, because negative values of the R coefficient also prove that the variables correlate well, but against the phase. In the case of the analysis of many variables, their mutual correlation is placed in a square table that is characterized by symmetry and values of 1 on the diagonal.

The third statistical tool used in this work is Linear Regression (LR). It is used to analyse the relationship between variables, and, notably, to study the impact of individual variables on a given phenomenon. Formally, LR is used in statistics to model the relationship between one scalar variable and more variables, the so-called explanatory variables. In this case, we are talking about the so-called multivariate linear regression. Suppose we have a set of pairs $X_j = (x_{j,1}, \ldots, x_{j,n}; y_j)$ for $j = 1, \ldots, m$. The purpose of linear regression is to construct a linear transform that can be expressed as follows:

$$Y_j = w_0 + w_1 x_{j,1} + \ldots + w_n x_{j,n} + E_j \tag{3}$$

for each of the m examples of variables. In formula (3), the values of individual coefficients w_i determine their influence on the variable Y_j, and E_j is the error variable, which we interpret as an unobserved random variable that adds "noise" to the linear relationship. In this method, the main task is to determine the set of coefficients $W = [w_0, \ldots, w_n]$. In this study, the Least Squares Estimation (LSE) (Marquardt, 1963) method was used to achieve this goal. The LSE algorithm is an optimization procedure, which in this case implements the postulate of finding such a set W^* that the sum of squared distances between the value of the estimate $\widehat{Y}$ and the value of Y for all data is as small as possible, therefore:

$$W^* = \sum_{j=1}^{m} \left(WX_j - Y_j\right)^2. \tag{4}$$

In this investigation, in order to find (4), Ordinary Least Squares (OLS) (de Souza and Junqueira, 2005), an optimisation method that belongs to the set of LSE optimisers, is applied. In this method, the difference between the measured data and the estimated vector of W* coefficients can be presented in the following form:

$$eq = Y = XW^*. \tag{5}$$

For this vector e, the following cost function is determined:

$$cost(W) = e^T e = (Y - XW^*)^T (Y - XW^*) \tag{6}$$

Thus, it can be shown that the derivative of the function (6) with respect to the weight vector W is:

$$\frac{dcost(W)}{dW} = -2X^T Y + 2X^T XW. \tag{7}$$

Now, by equating formula (7) to zero, it is possible to determine the coefficients W satisfying eq. (4):

$$W^* = \left(X^T X\right)^{-1} X^T Y. \tag{8}$$

The above method, for the purposes of this contribution, has been described very briefly. More about it can be found in (Uyank and Güler, 2013).

In the case of modelling or prediction of certain issues related to the COVID-19 pandemic, the use of modern and complex algorithmic solutions such as neural networks, fuzzy logic or maximum entropy would be advisable. However, in this study, our primary goal was to investigate the impact of individual factors on the phenomena associated with the pandemic, namely, the incidence of new infections and the surplus of deaths. We chose to apply the LR method. This is well established in literature, and is convergent, fast and, above all, fully interpretable. This interpretability allowed us to conduct a full analysis of the considered factors.

To conclude, it is necessary to stress that the aforementioned techniques are applied in the following strict order. As part of preprocessing, all data are standardized. Subsequently, in order to obtain information on the relationship between variables in the form of correlation, the Pearson coefficient is determined. The last element is the application of LR method – as this allows for the investigation of the influence of individual variables on the analysed quantity. Interpretation of this influence is based on the internal parameters of the LR model.

4. Data characteristics

The data for analysis, depending on the subject matter, was obtained from several sources. The first was information related to air pollution in the form of suspended particulates PM2.5 and PM10. Data related to this type of pollution come from the resources of Airly, which is a leader in Poland in terms of measuring air quality, and uses a dense network of its own devices that enables the possibility of visualization of the current atmospheric condition and prediction of air quality for the next 24 h. Another source of data was information related to deaths in Poland. This can be obtained from the Central Statistical Office. Undoubtedly, the data related to the course of COVID-19 in Poland, voluntarily compiled by Michal Rogalski (Rogalski, 2020) turned out to be very valuable. Unfortunately, the latter data with a detailed breakdown by district were made available by the Polish government only until November 23. After that date, only data related to voivodships has been published. The data from the Central Statistical Office (GUS, 2019) from 2019 shows that in Poland, the total number of residents is equal to approx. 38,411. Table 1 reveals the percentages of women and men diversified into the different age ranges.

Table 1
Percentage division of the population in Poland based on the age range.

Sex	Entire population	Up to 44	44–65	Over 65
Women	51.7: 52.6 (urban), 50.2 (country)	49.2	55.2	62.1
Men	48.3: 47.4 (urban), 49.8 (country)	50.8	44.8	37.9

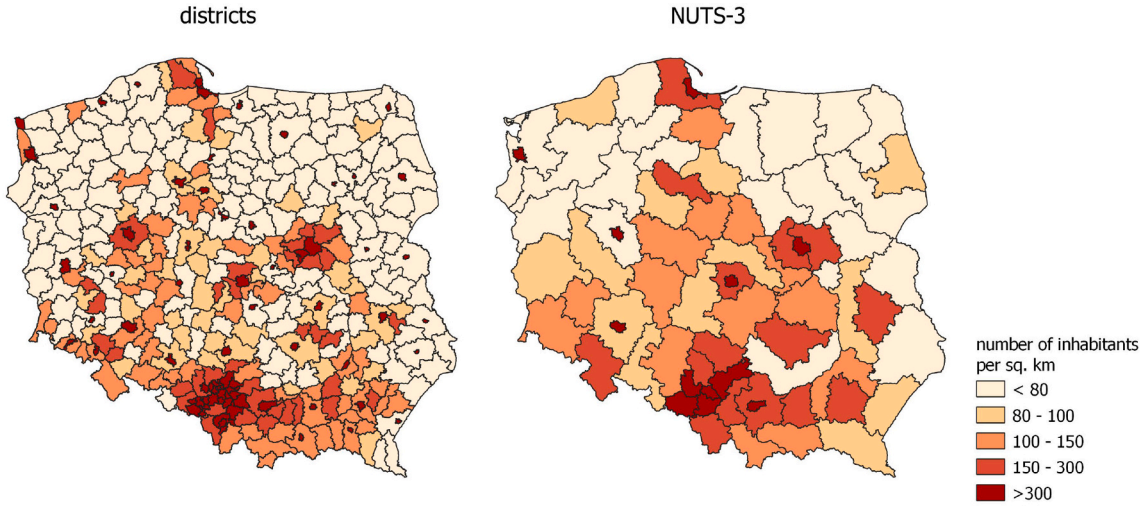


Fig. 4. Population density in the territory of Poland defined by districts (left) and NUTS-3 (right).

Poland is divided into 16 voivodships and 380 districts, and the area of the country is 31,270,000 ha. Due to the quite diverse nature of the population density of Poland's inhabitants (Fig. 4), in this analysis, two geographical models were used independently. The first is divided into 380 districts, and the second one into 73 NUTS-3 units. The latter characterizes different areas in the EU in a uniform way.

The NUTS (fr. *nomenclature des unités territoriales statistiques*) represents spatial units utilized for statistical analysis that are part of the common classification of territorial units for statistics in the European Union. The purpose of establishing the NUTS division was to standardize various administrative divisions between EU Member States and thus ensure comparable statistical data. The division of NUTS units is three-tier and is built upon the administrative divisions of the Member States. However, it has a number of conditions concerning, for example, the minimum and maximum size of units, as well as their changes over time, therefore, they may be different to the current administrative units in a given country.

Fig. 5. The division of the territory of Poland into districts (thin line) and NUTS-3 (thick line).

In Poland, 97 NUTS units have been designated and used statistically since 2018. These are, respectively: 7 NUTS-1 macroregions covering voivodeships (of which there are 16), 17 NUTS-2 regions, which include individual voivodships or their parts, and 73 NUTS-3 subregions grouping districts (in this paper called as 'districts'). In Poland, almost all NUTS-2 units correspond to voivodships, with the exception of the Mazowieckie voivodship, where a separate region is the Warsaw capital macro region, separated from the Mazowieckie voivodship.

NUTS-3 sub-regions will be used in further analyses due to the fact that they are the smallest units for which up-to-date data on deaths are available.

Fig. 5 shows the mutual dependence of the location, i.e. the inclusion of districts within the NUTS3 units.

In summary, it should be emphasized that both the environmental data, i.e. the information related to population and geographic conditions, and the air pollution data, are of the Big Data type. Its volume is in the order of terabytes in size.

5. Data processing results and analysis

5.1. The course of COVID-19 cases

Due to the availability of data, we assumed for further analyses the number of confirmed cases of Sars-CoV-2 coronavirus infections in districts, compared to 100,000 inhabitants (as of November 23, 2020). The above mentioned allows a comparison of data in various regions of Poland, regardless of the adopted type of administrative division. Moreover, this figure tells us about the spread of the epidemic, and it can also be used to infer the course of cases, because acute illness, unlike asymptomatic cases, required more tests and thus has resulted in a greater number of reported cases. Therefore, we assumed that in the case of the testing strategy in Poland, the number of all positive results is strongly correlated with the number of positive results and the symptomatic course.

Fig. 6 shows the number of officially confirmed COVID-19 cases in Poland as of the stated date. As in Fig. 4, the data is presented in two conventions. The first one (Fig. 6 left) shows the division of the country into districts, while the second one (Fig. 6 right) shows the distribution of incidence, taking into account the administrative division of Poland into NUTS-3 units. In both cases, the data covers the duration of the pandemic until 23/11/2020. As you can see, the areas most intensely affected by the disease are naturally associated with large urban centers such as Warsaw, Kraków, Poznań, etc. Moreover, a large number of regions characterized by morbidity are located in the southern part of

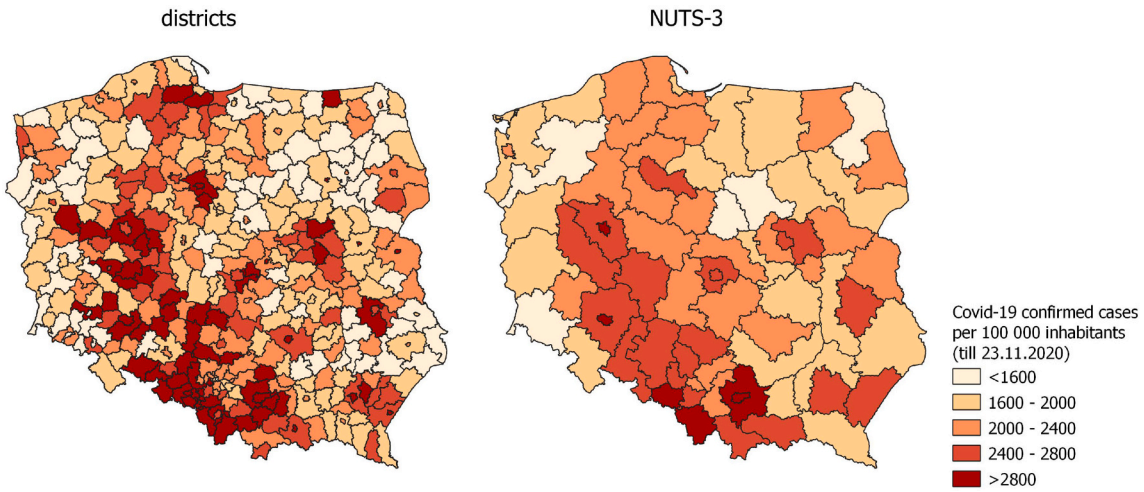


Fig. 6. Number of confirmed COVID-19 cases per 100,000 inhabitants by 31/10/2020 (left for administrative division into districts, right for NUTS-3).

the country, in particular in the Małopolskie, Śląskie, the Lower Silesian and Greater Poland borderlands and in part of the Podkarpackie voivodeship.

5.2. Mortality in 2020 and 2019 in Poland

The second, after the number of cases, explanatory variable, used by us for the analysis, was the surplus of deaths (in total, without distinguishing between causes) in 2020 compared to 2019. This indicator may serve as a descriptor of the course of the epidemic in a given area, including deaths from COVID- 19 - both diagnosed and undiagnosed, as well as deaths resulting from the overload of health care in a given area caused by the pandemic. Current data on deaths are published for NUTS-3 units and such units were adopted for further analysis.

Fig. 7 shows the distribution of the relative number of deaths in each NUTS-3 region. This relativity was presented as the percentage ratio of the number of deaths in 2020 (to November 30, 2020) to the number of deaths in the corresponding period of 2019. The first of the noticeable features of this analysis is that in practically each of the regions, the number of deaths in 2020 for each of the analysed sub-regions is higher than in the previous year by 110%, and in certain regions of Małopolska, Podkarpacie and the Płock subregion this increase covers as much as 20%. The worst situation - a 24% increase in the number of deaths is observed in the region of the Podkarpackie Voivodeship near the border with Ukraine. Interestingly, this area has quite a large number of cases (compare Fig. 6 left) and a relatively small population (Fig. 4 left). However, in the rest of Poland, this increase in deaths is noticeable up to 15%.

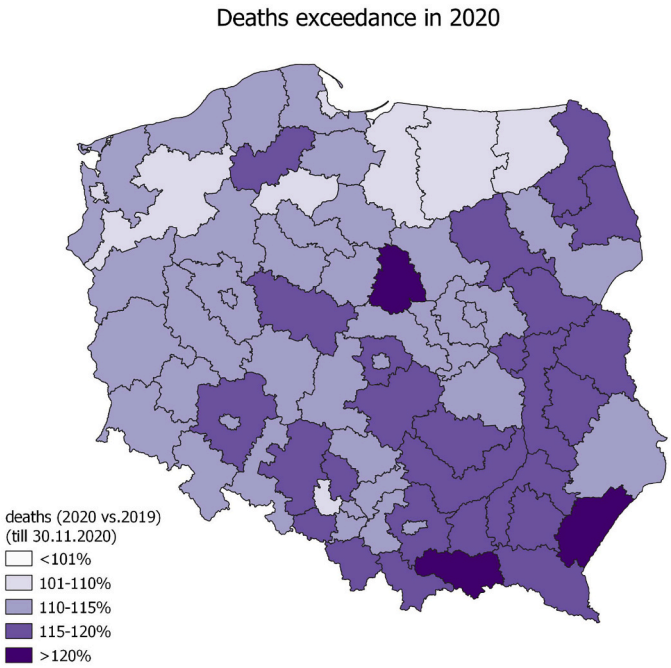


Fig. 7. Ratio of deaths occurring in 2020 to the number of deaths in 2019 expressed as a percentage in individual NUTS-3 territorial units.

Unfortunately, due to the lack of official data on the number of deaths via district, in this case analysis, it was not possible to present the distribution of excess deaths for a smaller territorial unit within Poland.

5.3. Particulate matter air pollution in Poland

One of the factors considered in the analysis is air pollution. As a measure of air pollution, we adopted the annual averaged values of PM2.5 (and PM 10) concentrations from all stations in a given district or NUTS-3 unit. In this study, we focused on averaged values, as such information shows a broader profile of the presence of polluted air in a given area. Moreover, it is precisely this approach to data that is also important when analysing the impact of pollution on respiratory diseases, which include COVID-19 (Wiki, 2020). If there was no Airly sensor in the area of a given district, the data on pollution were supplemented on the basis of the values from neighbouring districts using own interpolation algorithms. Figs. 8 and 9 show the distribution of air pollution in the form of averaged values of dust concentrations PM2.5 and PM10, respectively. Based on the analysis of data from Airly sensors, a very large correlation between these pollutants is easily noticeable. Indeed, for some measuring stations the value is approx. $R = 0.96$ (Kowalski and Warchałowski, 2018). Due to the great correlation between PM10 and PM2.5 assessments and that of these with certain observations related to the analyses already performed in the first phase of the pandemic (Kowalski and Konior, 2020), the inquiry used only the PM2.5 pollutant as a defining unit. On a deeper analysis of Figs. 8 and 9, we can see basically single areas in which we do not observe high correlation. At the same time, it should be clearly emphasized that in both figures we have different scales, which result from separate standards and diverse accepted permissible values of concentrations of these pollutants. Over all, in this assessment, it can be noticed that the air pollution factor is quite strongly independent of the administrative division. On the maps shown in Fig. 8, we can see that the western and northern border regions are areas of lowest air pollution values. In

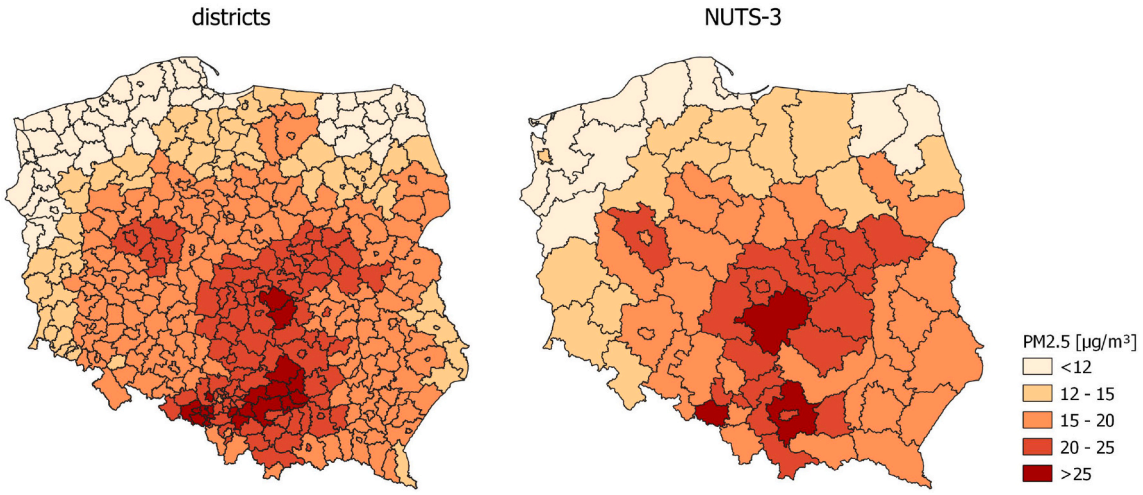


Fig. 8. Average PM2.5 dust pollution in individual regions of Poland (left for administrative division into districts, right for NUTS-3).

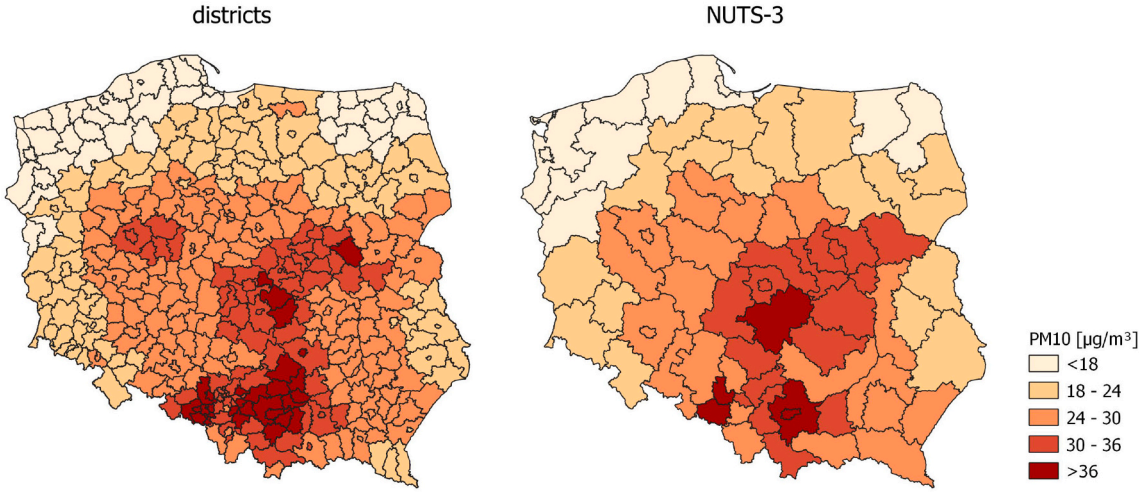


Fig. 9. Average PM10 dust pollution in individual regions of Poland (left for administrative division into districts, right for NUTS-3).

contrast, central regions, such as the districts of Mazowieckie and Wielkopolskie voivodeships, as well as that of Silesia and Małopolska, are very unfavourable in this respect. In particular, the trend is noticeable in the buffer zone of Kraków, Łódź and adjoining districts in Silesia. Interestingly, the above bad situation is very intensively noticeable in the mountainous regions of Poland - in such districts as Nowy Targ, Myślenice, Suski and Wadowicki. To a large extent, such unfavourable analytical results show the influence of geography (hilly locations) and, in most of them, the lack of an organized system of home insulation, as well as the use of coal for heating in wintertime. The poor state of air pollution levels in other regions is largely determined by the strong development of industry.

5.4. Factor analysis

This part of the article presents an analysis of the dependence of factors both on the number of new cases of corona virus and the number of deaths. For the analysis, apart from air pollution, we also took into account the impact of other factors, which will be briefly described. Below, in addition to the name and a short description of the factors, both their designation and the data source are presented. Thus, the following quantities (factors) were considered for the impact analysis:

- population density (POP_DEN, source: GUS), determines the intensity of interpersonal contacts and the probability of infection (Carozzi et al., 2020);
- touristic attractiveness (TUR, source: GUS) expresses the number of hotels per 100,000 inhabitants - this is an indicator of which districts were holiday destinations during times conducive to the spread of the virus;
- % of the population under 14 (YNG_PCT, source: GUS), shows the percentage of children attending (primary and secondary) school; here, size is taken into account because school attendance has contributed to the spread of the epidemic (European Centre for Disease Prevention and Control, 2020);
- % of seniors (SEN_PCT, source: GUS) - defines the percentage of the local population aged 65 and over; testifies to both the peripherality of districts mitigating the course of the epidemic and the extent of the group most vulnerable to the acute course of the disease (Walter and McGregor, 2020);
- feminization rate in the senior population (FEM_SEN, source: GUS) - shows the ratio of women to men in the population over 60 – such figure is included because of the variability in the course of the disease according to the patient's sex;

- % of the population employed in industry (INDST, source: GUS) – this indicates the degree of contribution of industrial or mining workplace to the emergence of disease outbreaks (Money.pl, 2020);
- distance from land border crossings (BOR, source: OpenStreetMap) – this item indicates the 'internationalization' of a given district, and serves as a means of revealing the degree of possible contact with infected people coming from abroad;
- access to medical assistance (DOC, source: GUS) – this factor is representative of the extent of locally available medical services – the number of doctors per 100 thousand residents;
- feminization of society (FEM, source: GUS) - the numeric ratio of women to men;

Additionally, the following determinations were used in the analysis:

- POP (source: GUS) - population of a given region;
- pm25 - concentration of pollution in the form of PM2.5 dust - average value in 2019–2020 based on Airly sensors;
- DEATH_EXC (source: GUS) – the ratio (comparison) of the number of deaths: 2019 versus 2020;
- C19_100k - (source: Rogalski (Rogalski, 2020)) - number of COVID-19 cases per 100,000 residents.

In the first part, the mutual influence of individual factors was examined with the use of correlation. Tables 2 and 3 present the interdependencies between the factors, taking into account the districts and NUTS-3 regions. Table 2 lacks the value of the R factor for the DEATH_EXC variable because, from autumn 2020 onwards, information on the death ratio between 2020 and 2019 has not been published in official government data.

Here, it should be noted that two variables standout: the number of doctors per 100,000 inhabitants (DOC) and the feminization of society (FEM). However, further analysis rejected the first variable due to the very high correlation with the location of large cities. In the second case, the FEM variable was also rejected, this time because of a very strong correlation with POP_DENS and SEN_PCT. Therefore, for the further part of the analysis of the impact of individual factors, a set of data was prepared consisting of 9 quantities (explanatory variables) and 1 (for districts) and 2 for (NUTS-3) dependent variables.

The results presented in Tables 2 and 3 clearly show that there are several examples where the correlation between the selected factors exceeds the value of 0.5. Examples include the following relationships:

- Air pollution and the number of COVID-19 cases,
- Air pollution and ratio of number of deaths, 2019 versus 2020,
- The number of cases of COVID-19 and the ratio of number of deaths, 2019 versus 2020.

The occurrence of the last of these relations is obvious, while the dependence of variables related to the occurrence of COVID-19 and mortality with air pollution is very disturbing.

In the next part of the study, the relationship between individual factors was presented using the multivariate linear regression (MLR) method. Due to the already mentioned problems with data related to deaths, five models were synthesized:

- M1 – a model taking into account the influence of the components on the number of COVID-19 infections per 100,000 (district level);
- M2 - a model that takes into account the influence of the components on the number of COVID-19 infections per 100,000 (NUTS-3 unit populations);
- M3 - a model taking into account the influence of the components on the number of COVID-19 infections per 100,000 inhabitants (NUTS-2 unit population, aka voivodship population);
- M4 – a model taking into account the influence of certain components on the ratio of deaths within NUTS-3 units, taking into account all variables.
- M5 - a model that takes into account the influence of the components on the ratio of deaths within NUTS-3 units - with strongly correlated variables removed. For these territorial division units, the correlation of YNG_PCT with SEN_PCT is −0.8.

Figs. 10, 11 and 12 show the influence of the analysed factors on the number of COVID-19 cases. All factors were normalized, which made it possible to compare the impact of various parameters at the district level. Thus, the graph shows how, for example, an increase in the incidence (in cases per 100,000 inhabitants) will accompany an increase in the value of a given factor from its minimum to its maximum value. It is quite important to evaluate the model, which can be done using the R (2) measure, which for the models from M1 to M5 is 0.27, 0.49, 0.74, 0.43 and 0.36, respectively. On the other hand, it has been observed that the use of larger spatial units (NUTS3, voivodeships) results in a correspondingly better adjustment of statistical models, due to the progressive aggregation (averaging) of data for these units (Fotheringham and Wong, 1991).

Results based on M1, M2 and M3 (Fig. 10) showed that air pollution was the most important factor among the analysed factors, the increase of which was to the highest degree correlated with the increase in the incidence in districts, NUTS-3 and NUTS-2 units. These outcomes match the global-scale result presented in (Coro, 2020). The increase in pollution by 100% was accompanied by an increase in the incidence by as much as 200–250 cases (where the total number of cases throughout Poland is 2170/100 thousand people - data as of November 23, 2020). The population factor was in second place, which is quite a natural and expected result. Another factor is industrialization resulting from the number of people employed in industry or mining per 100,000 residents. This factor is very important at the district level, which is understandable due to the relatively local nature of such jobs. Its impact can be

Table 2
Correlation given as Pearson coefficient between 12 variables, taking into account districts as units of administrative division.

districts	pm25	FEM	SEN PCT	FEM SEN	YNG PCT	TUR	INDST	BOR	DOC	POP DEN	POP	C19 100k
pm25	1.0	0.0	0.3	−0.2	−0.2	−0.3	0.2	0.2	0.1	0.1	0.1	0.4
FEM	0.0	1.0	−0.1	0.0	0.5	0.0	−0.4	0.2	−0.5	−0.5	0.0	−0.2
SEN PCT	0.3	−0.1	1.0	0.1	−0.7	−0.1	−0.1	0.0	0.3	0.3	0.2	−0.1
FEM SEN	−0.2	0.0	0.1	1.0	−0.2	0.0	−0.1	−0.1	−0.1	−0.1	0.0	−0.2
YNG PCT	−0.2	0.5	−0.7	−0.2	1.0	0.1	−0.2	0.2	−0.5	−0.5	−0.2	0.0
TUR	−0.3	0.0	−0.1	0.0	0.1	1.0	−0.2	−0.1	−0.1	−0.1	−0.1	0.0
INDST	0.2	−0.4	−0.1	−0.1	−0.2	−0.2	1.0	−0.1	0.3	0.3	0.1	0.3
BOR	0.2	0.2	0.0	−0.1	0.2	−0.1	−0.1	1.0	0.0	−0.1	0.0	0.0
DOC	0.1	−0.5	0.3	−0.1	−0.5	−0.1	0.3	0.0	1.0	0.8	0.5	0.2
POP DEN	0.1	−0.5	0.3	−0.1	−0.5	−0.1	0.3	−0.1	0.8	1.0	0.5	0.2
POP	0.1	0.0	0.2	0.0	−0.2	−0.1	0.1	0.0	0.5	0.5	1.0	0.2
C19 100k	0.4	−0.2	−0.1	−0.2	0.0	0.0	0.3	0.0	0.2	0.2	0.2	1.0

Table 3
Correlation given as Pearson coefficient between 13 variables, taking into account NUTS-3 as a unit of administrative division.

NUTS-3	pm25	SEN PCT	FEM	FEM SEN	YNG PCT	TUR	INDST	BOR	DOC	POP DEN	POP	C19 100k	DEATH EXC
pm25	1.0	0.3	0.2	−0.1	−0.2	−0.4	0.3	−0.1	0.2	0.2	0.2	0.6	0.5
SEN PCT	0.3	1.0	0.7	0.2	−0.8	−0.3	0.0	0.0	0.6	0.5	0.4	0.1	0.2
FEM	0.2	0.7	1.0	0.4	−0.8	−0.2	0.3	−0.1	0.9	0.8	0.6	0.2	0.0
FEM SEN	−0.1	0.2	0.4	1.0	−0.4	0.0	−0.1	−0.1	0.1	0.2	0.1	−0.2	−0.2
YNG PCT	−0.2	−0.8	−0.8	−0.4	1.0	0.3	−0.4	−0.1	−0.6	−0.5	−0.4	0.1	0.1
TUR	−0.4	−0.3	−0.2	0.0	0.3	1.0	−0.3	0.2	−0.1	−0.1	−0.1	0.0	0.0
INDST	0.3	0.0	0.3	−0.1	−0.4	−0.3	1.0	0.0	0.3	0.3	0.2	0.1	0.0
BOR	−0.1	0.0	−0.1	−0.1	−0.1	0.2	0.0	1.0	0.0	0.0	0.0	0.0	0.1
DOC	0.2	0.6	0.9	0.1	−0.6	−0.1	0.3	0.0	1.0	0.8	0.6	0.2	0.1
POP DEN	0.2	0.5	0.8	0.2	−0.5	−0.1	0.3	0.0	0.8	1.0	0.6	0.2	0.0
POP	0.2	0.4	0.6	0.1	−0.4	−0.1	0.2	0.0	0.6	0.6	1.0	0.3	0.2
C19 100k	0.6	0.1	0.2	−0.2	0.1	0.0	0.1	0.0	0.2	0.2	0.3	1.0	0.5
DEATH EXC	0.5	0.2	0.0	−0.2	0.1	0.0	0.0	0.1	0.1	0.0	0.2	0.5	1.0

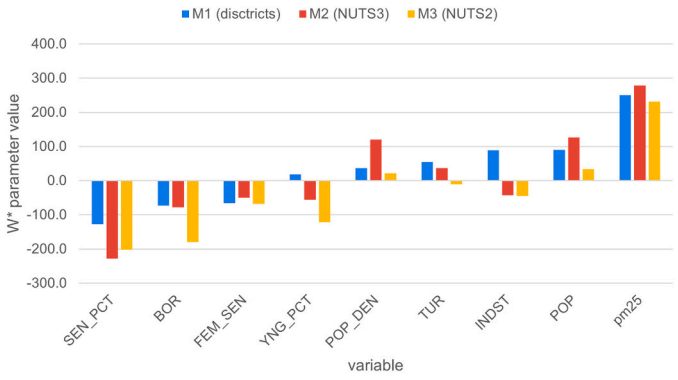


Fig. 10. *W** parameters values in MLR M1, M2 and M3 models.

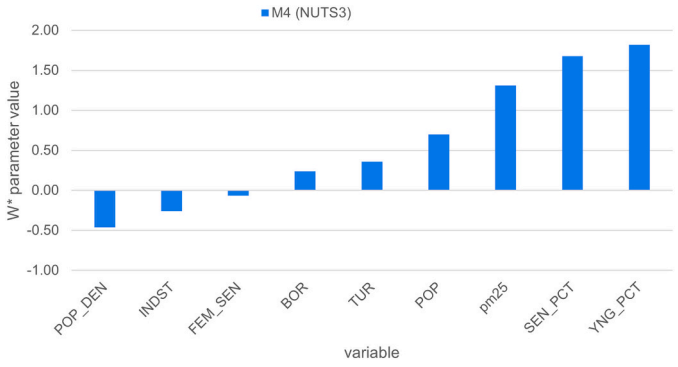


Fig. 11. *W** parameters values in MLR M4 model.

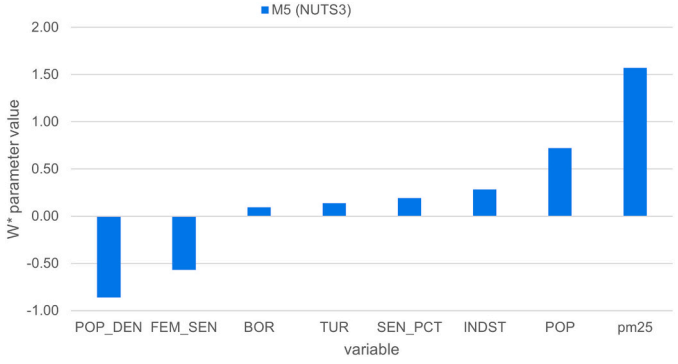


Fig. 12. *W** parameters values in MLR M5 model.

determined on a 1:1 scale, so an increase by 100% determines the average increase in the incidence of about 90 people per day per 100,000 residents. The fourth most negatively influencing factor at the level of districts and smaller NUTS-3 units is touristic attractiveness. The impact of this premise, with its 100% increase, is estimated at 55 and 37 new disease cases per 100,000 residents in relevant territorial units. The last variable with a positive impact is the share of the number of children under 14 (the number people who are attending primary or secondary school per district), whose impact on the number of cases was limited to 19. This factor is greatly underestimated as the data also covers the period in which children were not allowed to participate in live school activities. In addition, in this period in Poland, school children were forbidden from Monday to Friday to be moving about publicly between 8 am and 4 pm without adult supervision. This could significantly reduce the impact of the YNG_PCT variable on the emergence of new chains of infections. Interestingly, this factor did not have a negative effect on the remaining M2 and M3 models. The remaining factors, i.e. feminization of senior population, distance from border crossings, the percentage of seniors in society, are not factors that accompanied the increase in the disease. The reason for such an impact of these factors may be, on the one hand, the quite strong self-preservation instinct of this group, which manifests itself in a fairly strong adherence to the recommendations, and on the other hand, the introduction of the so-called 'hours exclusively for seniors'.

The second round of analyses is related to the results presented in Figs. 11 and 12, which show the numerical impact of individual factors on excess deaths for the M4 and M5 models. In the first model, the first two positions are occupied by the variable share of seniors and young people in local society. While the first factor does not come as a surprise, because this group is very much at risk of dying from COVID-19 or having complications after suffering from this disease, the group of people under the age of 14 is quite surprising. Thus, the obtained dependencies were analysed once more. As a result, it turned out that both variables are very strongly correlated ($R = -0.8$), so the replacement YNG_PCT was rejected from the M5 model. Air pollution is another factor that has a significant impact on the excess deaths. In this case, an increase of 100% of this factor induces an increase of 1.3% in the excess of deaths compared to the previous year (where the average increase in the ratio of deaths from 2020 to 2019 throughout Poland is 14%). Other factors positively influencing the excess deaths are the population and touristic attractiveness of the region. However, these variables have an impact of less than 1%. The remaining variables, i.e. industrialization and population density, have a negative impact on the number of deaths. The first of these, as shown in the previous factor analysis, is important in smaller units of territorial division such as districts. In contrast, the variable of population density is correlated with the presence of large cities, where there is much better access to professional medical care through a significant number of health centres and hospitals.

As described above, only 8 of the describing variables were adopted for the next tested model. In the M5 model, the alternative YNG_PCT was

rejected, thanks to which the influence of the percentage share of seniors significantly decreased (approx. 0.5%). On the other hand, the prerequisite for a very strong impact of air pollution on the excess deaths has come to the fore. In this model, the doubling of air pollution in the form of PM2.5 dust condensation implies a 1.6% increase in the excess of deaths compared to the previous year. Two other variables having a positive impact on this phenomenon are the population and the % of employed in industry (0.7% and 0.3%). On the other hand, such variables as feminization of the senior population and population density have a negative impact on the excess mortality of approximately 0.55% and 0.8%, respectively. The first of these facts reports the different effects of COVID-19 on men and women. As shown by this model, women are less likely to have a fatal outcome when afflicted. On the other hand, the reverse impact of population density - as in the M4 model - was determined by better access to medical services in large agglomeration clusters.

6. Discussion

There are many studies in the literature related to both the COVID-19 disease itself and attempts to predict the direction of infection development and its inhibition through the use of prevention methods (Dansana et al., 2020; Melin et al., 2020; Rosario et al., 2020). Evaluation of the effectiveness of disinfection and prevention undertaken in Wuhan and assessment of the COVID-19 pandemic control methodology became the subject of the work of (Zhu et al., 2020). In turn, (Gatto et al., 2020) analysed the effects of the introduction of drastic preventive measures in Italy in 2019 – as based on modelling the developing epidemic. The analysis of factors influencing the recovery or death of the afflicted COVID-19 patient was the subject of a study by (Kang et al., n. d.) and Kang et al. (2020). Other reference sources are publications similar to this one and dealing with the course and development of the pandemic in a given country, e.g. in Italy (Lolli et al., 2020). The publications describing various models of infectious diseases (Funk et al., 2015; Melin et al., 2020) are very interesting, and aid in establishing viable predictivity of at least some phenomena.

The problems and doubts related to the study of the impact of air pollution on COVID-19 disease were systematically analysed in the article (Heederik et al., 2020). The authors of this publication point to the rather cursory nature of the analyses and deliberations with regard to this quite important topic. The tendency to judge the causality of the disease on the basis of the correlation between airborne particulate matter and disease was also subjected to criticism, especially without the analysis of other factors that are obviously also correlated with the number of cases and air pollution - such as population density. That is why the authors of this study conducted a multivariate analysis that took into account factors other than air pollution. It is worth recalling here that for the purposes of the analysis, it was air pollution that turned out to be the most highly correlated variable among the analysed factors with the number of cases per 100,000 residents.

Moreover, the authors of the discussed critical article (Heederik et al., 2020) point to the problem of research quality and publication reliability, which is particularly important with regard to the social significance of the studied problem. The authors of this publication can assure the research population that it was written on the basis of a multivariate analysis of a very large volume of data. Furthermore, the information it contains is the result of in-depth conclusions of an interdisciplinary team of scientists from various research establishments with input from external experts.

Another interesting conclusion from the analysis carried out was that the use of larger spatial units resulted in a better fit of the statistical models - as measured by the correlation coefficient *R*. This results in the risk of hastily establishing a relationship when analysing large units, such as provincial or state entities (voivodships) and countries. On the other hand, analyses of smaller territorial units such as districts, require the availability of spatially accurate data. In the case of data on air quality in Poland (and in other EU countries), the number of state stations measuring air pollution (and particularly particulate matter) is insufficient to conduct such analysis; it is necessary to use a network of low-cost devices in order to accurately map pollution.

Since the archiving of data on the COVID-19 pandemic by individual countries of the world, many studies have been undertaken to estimate the increased risk of death among general or specific population categories. In general, clinical observations indicate which are high-risk groups. These include people over 65 years of age, patients with chronic respiratory diseases (including asthmatics), people undergoing oncology treatment, as well as patients who are administered immunosuppressants or who are long-term immunosuppressed (afflicted with Lyme disease, hepatitis C and many other chronic diseases). As the study by (Goldstein and Lee, 2020) indicates, the age patterns of COVID-19-related mortality in different countries are remarkably similar and underline a clear correlation between the age of patients and the death rate. The authors established that, unlike HIV / AIDS and the drug epidemic, COVID-19 deaths will be concentrated over a period of months, not decades. Furthermore, according to the report 20–16 (2020) by the International Center for Public Policy (Austin et al., 2020), there is a link between exposure to fine particulate matter and the morbidity and mortality from COVID-19. This research used an instrumental approach to variables based on the wind direction.

According to the age of hypotheses emerging among clinicians and virologists, it is difficult to determine which genetic factors (cancer burden, chronic diseases) or non-genetic factors (the state of the environment, including air and water quality, diet and lifestyle, as well as the general physiological state of the body) have a decisive influence on the course of the disease. An additional element is the exposure of the organism to the pathogenic factor (in this case it is the amount of virus particles that enters the blood through the mucosa). Due to the tropism of the COVID-19 virus to the tissues of the lower respiratory tract, an important element is the initial presence of such an amount of virus particles that will not give the body time for an immune response and production of antibodies. For this reason, the condition of the infected patient and the environment in which he has to combat the disease are of great importance.

7. Conclusion

This publication describes the course of the corona virus pandemic in Poland that occurred until November 31, 2020. For the purposes of statistical analysis, many data sources were used, such as the population-related statistics that were provided by the Central Statistical Office. Another important source of data was Airly, which has been, for many years, generating air quality data through the use of a dense mesh of air pollution measurement devices established throughout Poland and much of the European Union. The last of the data sources is the Rogalski project (Rogalski, 2020), operating as part of a volunteer project that has contributed comprehensive data on the course of COVID-19 in Poland in a consistent form.

The research used known statistical tools, on the basis of which it was possible to generate numerical conclusions on the relationship between both the number of COVID-19 cases per 100,000 local area inhabitants and the surplus of deaths occurring in 2020 and selected social, geographic and environmental factors.

The applied statistical models made it possible to extract some dependencies flowing from the source data, which made it possible to explain, using the analysed factors, approx. 30% of the differentiation in the number of cases between individual districts. The remaining differentiation, which we have not been able to explain by means of the models used, may be related to the role of other factors (not yet identified), as well as to the randomness of outbreaks in the initial stage of the epidemic.

Our analysis showed that the correlation between air-borne particulate matter (smog) and COVID-19 incidence was very clear and, what

must be underlined, was much higher than the correlation for any other factor that, according to the literature, may have a positive effect on the incidence and for which we were able to obtain data.

Similar analyses are certainly worth carrying out for other countries - especially those harder hit by the coronavirus pandemic. It also remains to be analysed, in the light of data from Poland, the role of high pollution episodes and their impact on the course of the disease at a given moment, as well as the role of smog in virus transmission - which may be related to the observed seasonality of the disease. Therefore, the fight for clean air, as well as the monitoring of pollution and the disclosure of information about it to the public, may turn out to be crucial for finally defeating the COVID-19 pandemic.

The researchers' further plans with regard to COVID19 data are primarily related to the next phase of the pandemic. In future research, a study associated with vaccination effectiveness will be carried out in order to assess the potential of pandemic containment.

Declaration of Competing Interest

None.

Acknowledgment

The authors of the publication would like to thank all institutions (Airly Inc. and GUS) and Michal Rogalski for making the data available for analysis.

This work was partially supported by Grants for Statutory Activity from Faculty of Physics and Applied Computer Science of the AGH University of Science and Technology in Cracow, Department of Aquatic Bioengineering and Aquaculture, Faculty of Food Science and Fisheries, West Pomeranian University of Technology in Szczecin and Rzeszow University of Technology, within the subsidy for the maintenance and development of research potential (UPB).

The authors would also like to thank the anonymous referees for their careful reading of the paper and their contribution of useful suggestions which helped to improve this article.

References

Ahasan, H.N., Das, A., Chowdhury, M.K., Minnat, B., 2013. Middle east respiratory syndrome coronavirus (mers cov): an emerging pathogen. J. Med. 14 (2), 156–163.

Austin, W., Carattini, S., Mahecha, J.G., Pesko, M., 2020. Covid-19 Mortality and Contemporaneous Air Pollution, Tech. Rep., CESifo Working Paper.

Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Pearson correlation coefficient. In: Noise Reduction in Speech Processing. Springer, pp. 1–4.

S. Carozzi, Felipe nad Provenzano, S. Roth, 2020. Urban density and Covid-19. https://www.iza.org/publications/dp/13440/urban-density-and-covid-19.

Cole, M., Ozgen, C., Strobl, E., 2020. Air pollution exposure and covid-19.

Coro, G., 2020. A global-scale ecological niche model to predict sars-cov-2 coronavirus infection rate. Ecol. Model. 431, 109187. URL. https://www.sciencedirect.com/science/article/pii/S0304380020302581.

Dansana, D., Kumar, R., Bhattacharjee, A., Hemanth, D.J., Gupta, D., Khanna, A., Castillo, O., 2020. Early diagnosis of covid-19-affected patients based on x-ray and computed tomography images using deep learning algorithm. Soft. Comput. 1–9.

de Souza, S.V., Junqueira, R.G., 2005. A procedure to assess linearity by ordinary least squares method. Anal. Chim. Acta 552 (1–2), 25–35.

Dong, J., Liu, Y., Bao, H., 2021. Revalue associations of short-term exposure to air pollution with respiratory hospital admissions in lanzhou, china after the control and treatment of current pollution. Int. J. Hyg. Environ. Health 231, 113658. URL. http://www.sciencedirect.com/science/article/pii/S1438463920306040.

Duszyński, J., Afelt, A., Ochab-Marcinek, A., Owczuk, R., Pyrć, K., Rosinska, A., Rychard, T., 2020. Smiatacz, Zrozumieć covid-19. opracowanie zespołu ds. covid-19 przy prezesie polskiej akademii nauk. Tech. Rep. 1–70. URL. https://informacje.pan.pl/images/2020/opracowanie-covid19-14-09-2020/ZrozumiecCovid19_opracowanie_PAN.pdf.

European Centre for Disease Prevention and Control, 2020. Covid-19 in children and the role of school settings in covid-19 transmission. Tech. Rep. 1–59. URL. https://www.ecdc.europa.eu/en/publications-data/children-and-school-settings-covid-19-transmission#no-link.

Fotheringham, A.S., Wong, D.W., 1991. The modifiable areal unit problem in multivariate statistical analysis. Environ. Plan. A 23 (7), 1025–1044.

Funk, S., Bansal, S., Bauch, C.T., Eames, K.T., Edmunds, W.J., Galvani, A.P., Klepac, P., 2015. Nine challenges in incorporating the dynamics of behaviour in infectious diseases models. Epidemics 10, 21–25.

Gabriela, S., 2020. Odsetek zakażonych w służbie zdrowia: Polska a inne kraje. Tech. Rep. https://konkret24.tvn24.pl/zdrowie,110/odsetek-zakazonych-w-sluzbie-zdrowia-polska-a-inne-kraje,1011959.html.

Gatto, M., Bertuzzo, E., Mari, L., Miccoli, S., Carraro, L., Casagrandi, R., Rinaldo, A., 2020. Spread and dynamics of the covid-19 epidemic in Italy: effects of emergency containment measures. Proc. Natl. Acad. Sci. 117 (19), 10484–10491.

Gliński, Z., Żmuda, A., 2020. Epidemie i pandemie chorób zakaźnych.

Goldstein, J.R., Lee, R.D., 2020. Demographic Perspectives on Mortality of Covid-19 and Other Epidemics, Tech. Rep. National Bureau of Economic Research.

GUS, 2019. Polska w liczbach 2019. Tech. Rep. 1–43. https://stat.gov.pl/obszary-tematyczne/inne-opracowania/inne-opracowania-zbiorcze/polska-w-liczbach-2019,14,12.html.

Heederik, D.J., Smit, L.A., Vermeulen, R.C., 2020. Go Slow to Go Fast: A Plea for Sustained Scientific Rigour in Air Pollution Research during the Covid-19 Pandemic.

Hussain, H.Y., 2014. Incidence and mortality rate of middle east respiratory syndrome-corona virus (mers-cov), threatens and opportunities. J. Mycobac. Dis. 4, 162.

Inglot, M., Bereza, D., Biały, M., Bieniasz, J., 2020. Covid-19-opracowanie zgodne ze stanem wiedzy na 26.03.2020r. Tech. Rep. 1–45. https://www.umed.wroc.pl/sites/default/files/files/aktualnosci/2020/03/COVID_19_1_0_poprawiony_26_03_2020__wersja_po_formatowaniu_(2).pdf.

D. Kang, H. Choi, J.-H. Kim, J. Choi, Spatial epidemic dynamics of the covid-19 outbreak in China, Int. J. Infect. Dis.

Kowalski, P.A., Konior, A., 2020. Why Air Pollution is Linked to a Faster Spread of Coronavirus. Portal Air Quality News. URL. https://airqualitynews.com/2020/04/09/why-air-pollution-is-linked-to-a-faster-spread-of-coronavirus/.

Kowalski, P.A., Warchałowski, W., 2018. The comparison of linear models for pm10 and pm2. 5 forecasting. WIT Trans. Ecol. Environ. 230, 177–188.

Lolli, S., Chen, Y.-C., Wang, S.-H., Vivone, G., 2020. Impact of meteorological conditions and air pollution on covid-19 pandemic transmission in Italy. Sci. Rep. 10 (1), 1–15.

Lovato, A., de Filippis, C., 2020. Clinical presentation of covid-19: a systematic review focusing on upper airway symptoms. Ear Nose Throat J. 99 (9), 569–576 (0145561320920762).

Marquardt, D.W., 1963. An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Ind. Appl. Math. 11 (2), 431–441.

Masood, N., Malik, S.S., Raja, M.N., Mubarik, S., Yu, C., 2020. Unraveling the epidemiology, geographical distribution, and genomic evolution of potentially lethal coronaviruses (sars, mers, and sars cov-2). Front. Cell. Infect. Microbiol. 10, 499.

Melin, P., Monica, J.C., Sanchez, D., Castillo, O., 2020. Multiple ensemble neural network models with fuzzy response aggregation for predicting covid-19 time series: the case of mexico. In: Healthcare, vol. 8. Multidisciplinary Digital Publishing Institute, p. 181.

Money.pl, 2020. Kopalnie wylęgarnią koronawirusa. ponad 6,6 tys. zakażonych w 3 spółkach. Tech. Rep.. URL. https://www.money.pl/gospodarka/kopalnie-wylegarnia-koronawirusa-ponad-66-tys-zakazonych-w-3-spolkach-6533789016266881a.html

Mostowy, R., 2020. Pomiar i prognoza pandemii covid-19 w polsce w czasie rzeczywistym. Tech. Rep. URL https://rmostowy.github.io/covid-19/prognoza-polska/.

Pancer, K., 2020. Pandemiczne koronawirusy człowieka-charakterystyka oraz porównanie wybranych właściwości hco-sars i hcov-mers. Postepy Mikrobiologii 57 (1).

Ricon-Becker, I., Tarrasch, R., Blinder, P., Ben-Eliyahu, S., 2020. A seven-day cycle in covid-19 infection and mortality rates: are inter-generational social interactions on the weekends killing susceptible people? Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/2020.05.03.20089508 medRxiv. 2020.05.03.20089508.

Rogalski, Michał, 2020. Covid data set. Tech. Rep. URL http://bit.ly/covid19_powiaty.

Rosario, D.K., Mutz, Y.S., Bernardes, P.C., Conte-Junior, C.A., 2020. Relationship between covid-19 and weather: Case study in a tropical country. Int. J. Hyg. Environ. Health 229, 113587. URL. http://www.sciencedirect.com/science/article/pii/S1438463920305332.

Travaglio, M., Yu, Y., Popovic, R., Leal, N.S., Martins, L.M., 2020. Links between air pollution and covid-19 in england. medRxiv 229, 113587. https://doi.org/10.1016/j.ijheh.2020.113587 issn 1438-4639.

Uyank, G.K., Güler, N., 2013. A study on multiple linear regression analysis. Procedia Soc. Behav. Sci. 106, 234–240.

Walter, L.A., McGregor, A.J., 2020. Sex-and gender-specific observations and implications for covid-19. West. J. Emerg. Med. 21 (3), 507.

M. Wazna, Zgony z powodu covid-19 w polsce i na swiecie. jestesmy wysoko w zestawieniu, Tech. Rep. (2020). URL https://www.medonet.pl/koronawirus/koronawirus-na-swiecie,zgony-z-powodu-covid-19-w-polsce-i-na-swiecie–jestesmy-wysoko-w-zestawieniu,artykul,92538991.html.

WHO, 2020. Coronavirus Disease 2019 Situation Report, 94. URL. https://apps.who.int/iris/handle/10665/331865.

Wiki, 2020. Covid-19. Tech. Rep. URL https://pl.wikipedia.org/wiki/COVID-19.

Williamson, B., Piattoeva, N., 2019. Objectivity as standardization in data-scientific education policy, technology and governance. Learn. Media Technol. 44 (1), 64–76.

Wu, X., Nethery, R.C., Sabath, B.M., Braun, D., Dominici, F., 2020. Exposure to air pollution and covid-19 mortality in the United States. medRxiv. Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/2020.04.05.20054502, 2020.04.05.20054502.

Zheng, P., Liu, Y., Song, H., Wu, C.-H., Li, B., Ug, M., Jia, G., 2020. Risk of covid-19 and long-term exposure to air pollution: evidence from the first wave in 1 China 2. People 32, 33.

Zhu, H., Li, Y., Jin, X., Huang, J., Liu, X., Qian, Y., Tan, J., 2020. Transmission dynamics and control methodology of covid-19: a modeling study. Appl. Math. Model. 89, 1983–1998.