SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2023
ORGANIZATION AND MANAGEMENT SERIES NO. 186

CLUSTER ANALYSIS OF EPIDEMIC CURVES OF THE THIRD COVID-19 EPIDEMICS WAVE IN DIFFERENT REGIONS OF POLAND

Artur MARKIEL[1], Wojciech M. KEMPA[2*]

[1] Silesian University of Technology, Faculty of Applied Mathematics; artumar972@student.polsl.pl
[2] Silesian University of Technology, Faculty of Applied Mathematics; wojciech.kempa@polsl.pl,
ORCID: 0000-0001-9476-2070
* Correspondence author

Purpose: The purpose of the analysis was to divide Polish regions (as a single region we mean "powiat") into categories according to the way in which their population passed the third wave of COVID-19 epidemics and the attempt of linking the resultant classifications with other factors.
Design/methodology/approach On the basis of data about daily COVID-19 cases per 10,000 inhabitants in regions, weekly averages in every day were calculated. The curves of these averages for each region were approximated by a polynomial function. Then, for the majority of created functions, five characteristic "points" were determined: a maximum of the function together with a maximum and a minimum of its first derivative and two maxima of its second derivative, which are all located the closest to the function maximum. On the basis of coordinates of these points, regions were grouped by using of K-means method. Finally, the mean levels of various factors in obtained categories were analyzed as well as different classification models determined on this basis.
Findings: The performed analysis allowed to construct predictive models of the approximate shape of the epidemic curve in a given region. These models can be used in more-depth analysis of epidemics evolution.
Research limitations/implications: These models can be used in more-depth analysis of epidemics evolution.
Originality/value: By using cross-validation test, the number of clusters equal to five was determined. Mean values of the mentioned coordinates in each cluster allowed to determine an approximate shape of characteristic epidemic curve for a given group of regions. Only among some clusters there was a significant difference in population density, the percentage of population living in cities and the approximate percentage of inhabitants vaccinated after the third wave of COVID-19 epidemics. Nevertheless, on the basis of these factors and the age structure of the population, decision trees which classify most of the wave categories with a satisfactory accuracy were determined.

Keywords: Cluster analysis, decision tree, epidemic curve, method of K-means, region.

Category of the paper: Research paper.

1. Introduction

The COVID-19 pandemic was one of the greatest challenges for the entire world, individual regions, countries and societies in the 21st century. The powerful crisis that it caused both in the area of health care as well as in the sphere of economy and economy is and will certainly be the subject of numerous analyzes, studies, publications and discussion forums for a long time. It can be safely said that both on a global scale, i.e. worldwide, and locally - at the level of countries and regions - there is virtually no area where the pandemic would not leave its mark. Even a cursory bibliographic query allows to quickly realize that the number of studies on the impact of the COVID-19 epidemic on various aspects of life, business or social relations is already huge. At this point, it is difficult to try to distinguish any specific sphere or area of human activity affected by the pandemic without omitting another. The following brief bibliographic study therefore aims to provide an overview of those areas where the impact of the pandemic is currently being intensively studied. One can find the analysis of the impact of COVID-19 disease on environment and health in (Sneha et al., 2020; Verma et al., 2020), on agriculture in (Siche, 2020), on firm performance in (Shen et al, 2020), on globalization in (Shad et al., 2020), on mental health in (Banerje, 2020), on cancer care in (Richards et al., 2020), on business expectations in (Meyer et al., 2022), on waste management in (Sarkodie et al., 2021) and on education system in (Tarkar, 2020).

2. Short epidemic data description

The data come from Raport zakażeń koronawirusem (SARS-CoV-2), Archiwalne dane dla powiatów (eng. Coronavirus infection report (SARS-CoV-2), Archival data for powiats) in Poland's Data Portal (https://dane.gov.pl/), available in the Internet: https://dane.gov.pl/pl/dataset/2477/resource/33194/table, date of access: 11.04.2022. Data was written in csv files, which contain among others the following information about Polish powiats: name of voivodeship, the name of powiat, TERYT number, number of new COVID-19 cases per 10,000 inhabitants and the date of record state (but not in every csv file - in case of lack of this column, the date was deducted from the name of file).

Details about data and methodology are available in file "readme.txt" placed with csv files in a zip file downloaded from the mentioned source. It contains among others, the definition of the variable considered in our research, i.e. (translated from Polish) "Number per 10,000 inhabitants - Number of people for whom the day before the EWP system received a positive result for the first time, per 10,000 inhabitants", as well as definitions of other terms and general information about possible inaccuracies (e.g. corrections of previous data, delays in reporting,

cases of missing data on people being tested). However, in the studies described in this article, the data uncertainties indicated in mentioned file were found to be acceptably small. In the analysis data from a period 26.01.2021 - 9.06.2021 was considered. Number of COVID-19 cases per 10,000 inhabitants showed two kinds of variability:

- natural and random – that is characteristic in the case of such a phenomenon - that randomness is an integral part of many epidemic models/studies (e.g. (Bittihn et al., 2020; Britton, 2020); Chen-Charpentier et al., 2010; Fraser et al., 2004);
- unnatural and periodic (weekly) – number of cases often drops in Sundays and Mondays and later increases, that is the effect of specificity of testing/reporting system of COVID-19 cases in Poland. For week average = 1, mean level of cases on subsequent days of the week for all powiats for full weeks in February and March (month without additional work free days in considered period) was for example 0.69, 1.19, 1.27, 1.18, 1.19, 0.98, 0.5.

In the plot below (Fig. 1) the example of the mentioned data can be seen in the case of bolesławiecki powiat.

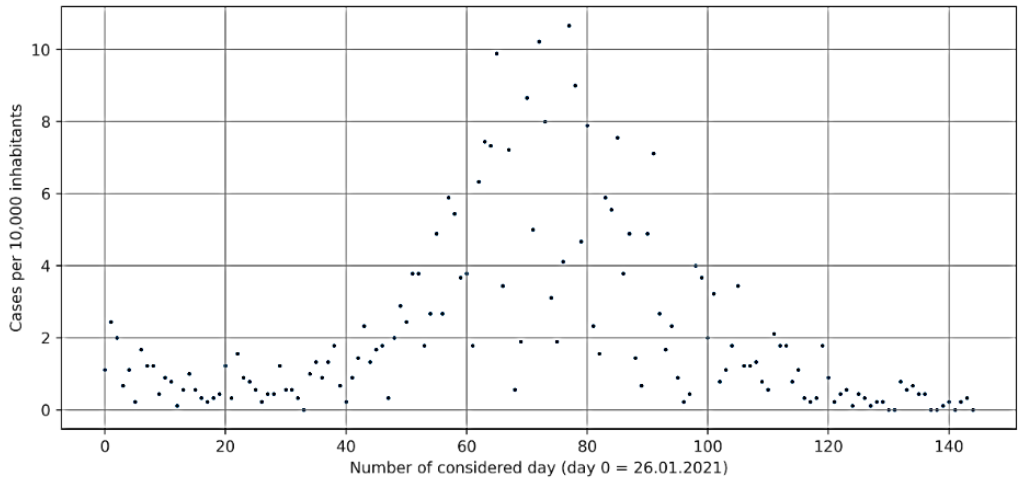

Figure 1. Original data in bolesławiecki powiat. Plot made with Matplotlib.pyplot. Version of Matplotlib: 3.5.0.

3. Operations on the data

In order to reduce the influence of the aforementioned factors on the shape of disease curves, the data from each day was converted into weekly averages, with using a moving average:

$$(new)X_i = \frac{\sum_{j=i-3}^{i+3}(old)\,X_j}{7}, \tag{1}$$

where X_i = number of cases per 10,000 inhabitants in i-th day of the considered period of time before conversion (and weekly average in i-th day after conversion). In the plot below (Fig. 2) the data from bolesławiecki powiat after mentioned transformation can be seen.

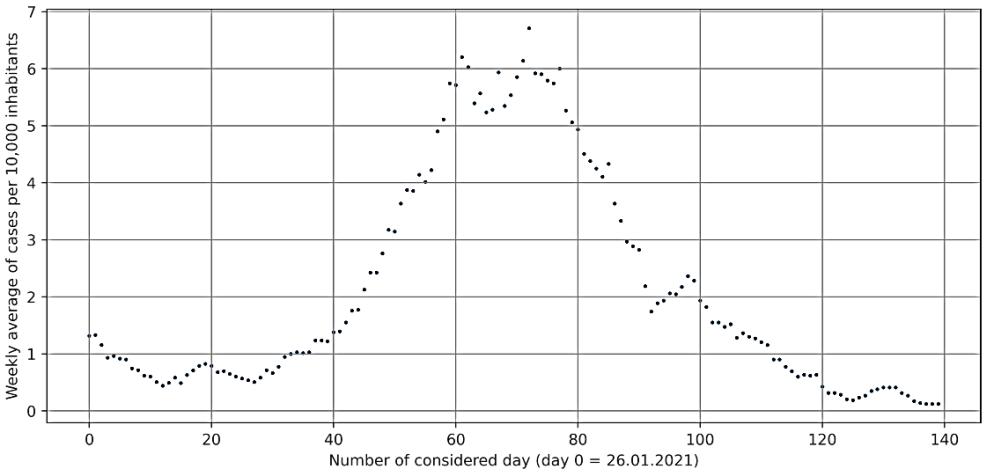

Figure 2. Data from bolesławiecki powiat after applying moving average. Plot made with Matplotlib.pyplot. Version of Matplotlib: 3.5.0.

The converted data was used to determine the approximated polynomial functions of degree 8, which return a value of weekly average of number of cases per 10,000 inhabitants in i-th day. For this purpose function ”polyfit” from the package ”NumPy” of Python programming language was used (version of Python: 3.8.12, version of NumPy: 1.21.2). The degree of polynomial function was chosen by trial and error method based on observation of the plot with function and data. In the plot below integral values of polynomial function created in mentioned way for case of bolesławiecki powiat can be seen (Fig. 3).

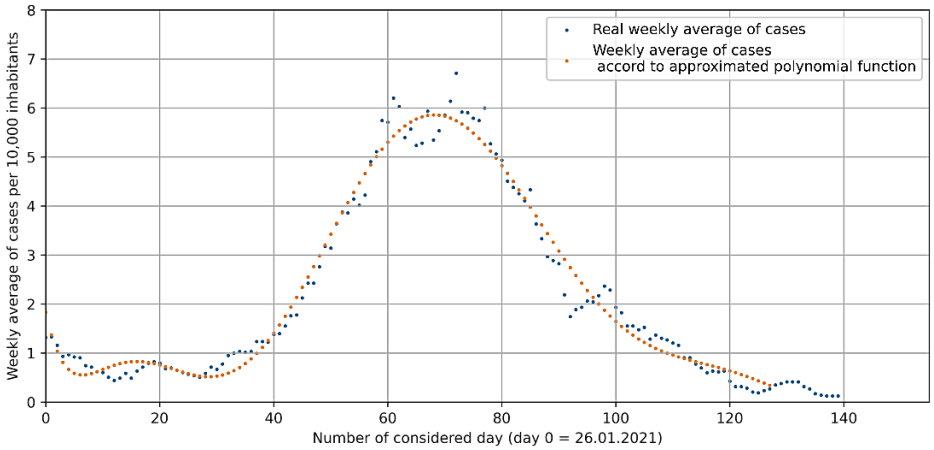

Figure 3. Data from bolesławiecki powiat after applying moving average with created polynomial function. Plot made with Matplotlib.pyplot. Version of Matplotlib: 3.5.0.

In order to describe the functions/epidemic curves obtained in previous step, 5 characteristic points (hereinafter referred to as ch.p.) were determined for each of them:

- first point: maximum value of the function – which indicates moment of highest number of new cases per 10.000 inhabitants;
- second point: local minimum of first derivative of the function – which indicates moment of the fastest decreasing of the number of new cases;
- third point: local maximum of first derivative of the function – which indicates moment of the fastest increasing of the number of new cases;
- fourth point: first local maximum of second derivative of the function - which indicates the beginning of the increase in the number of new cases;
- fifth point: second local maximum of second derivative of the function - which indicates the stop of the decrease in the number of new cases.

For points 2-5, maxima/minima closest to the maximum from point 1 were selected. Because in other places there can occur other local minima/maxima only integer X coordinates of the points were considered (to indicate the day of minima/maxima), so in practice many obtained points are only close to considered maxima/minima. This approach was chosen to not consider exact moments of epidemic (as 16:00, for example). The mentioned points and derivatives for data from bolesławiecki powiat can be seen in the plot below (Fig. 4).

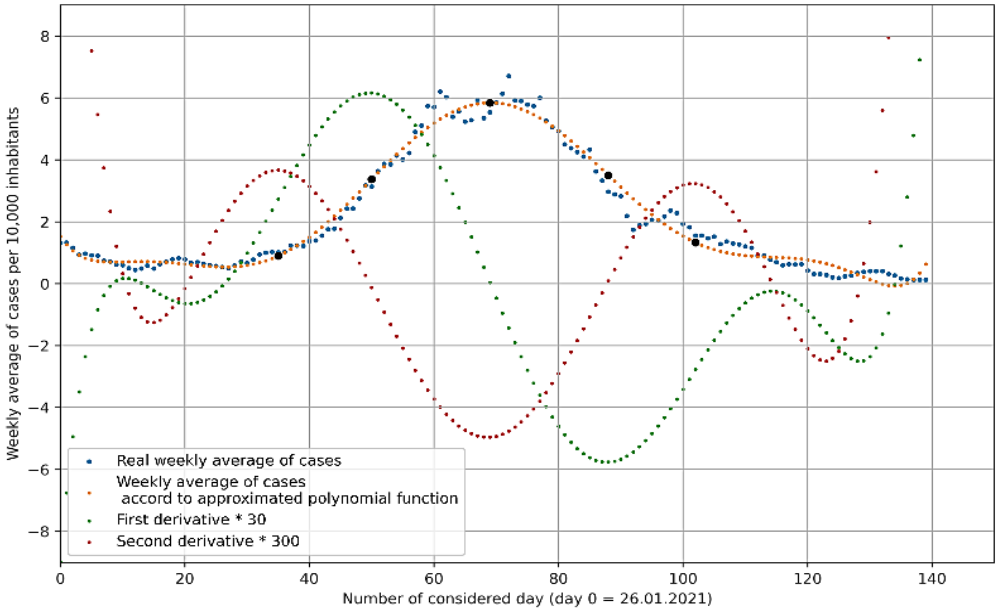

Figure 4. Data from bolesławiecki powiat after applying moving average with created polynomial function, derivatives and characteristic points. Plot made with Matplotlib.pyplot. Version of Matplotlib: 3.5.0.

Of course some of the analyzed curves were not shaped like a single bell curve (”classic wave”), what caused some difficulties in the analysis made in the way mentioned before. For example, it was difficult or even impossible to choose locations of the mentioned 5 characteristic points in the case similar to a monotonic declining function. For this reason,

powiats with an unusual epidemic curve shape were excluded from the further analysis. Their number, however, was acceptably small (only 21 powiats in total 380).

4. Cluster analysis of epidemic curves

Shortly about the used clustering method. In cluster analysis coordinates of 5 characteristic points on every curve were used as variables which described curves. Analysis was performed in program Statistica (TIBCO Software Inc. (2017). Statistica (data analysis software system), version 13. http://statistica.io).

The K-means method was used for the cluster analysis. In this approach firstly K clusters are chosen. Later analyzed objects are moved between them to minimize the variance inside them and maximize the variance between them. In the considered analysis data had been standardized, the cross-validation test (number of tries = 10, generator kernel = 1, number of clusters form 2 to 25, minimal decline=5%) was used to determine the number of clusters (result = 5), Euclidean distance was chosen as a measure of distance and maximum number of iteration of algorithm was 50. Initial cluster centers were chosen in the way to maximize the distance between them.

Numbers of powiats in subsequent clusters were (respectively): 88, 85, 86, 44, 56. To verify the hypothesis that these clusters do not differ, the Kruskal-Wallis test was used. The classic ANOVA analysis was abandoned due to fact that the condition of homogeneity of variance (in majority of cases) on the level of confidence 0.05 was not satisfied. To check this condition Levene's and Brown-Forsythe's tests were used.

Table 1.
Results of checking homogeneity of variance

Variable	P-value of Levene's test	P-value of Brown-Forsythe's test
X coordinate (day) of 1. ch.p.	0,001843	0,007093
Y coordinate (number of new cases per 10.000 inhabitants) of 1. ch.p.	0,000000	0,000000
X coordinate of 3. ch.p.	0,062809	0,135159
Y coordinate of 3. ch.p.	0,000000	0,000002
X coordinate of 2. ch.p.	0,000001	0,000005
Y coordinate of 2. ch.p.	0,000000	0,000000
X coordinate of 4. ch.p.	0,000001	0,000008
Y coordinate of 4. ch.p.	0,000006	0,000018
X coordinate of 5. ch.p.	0,000000	0,000000
Y coordinate of 5. ch.p.	0,034301	0,043779

Source: Authors' own.

P-values of the Kruskal-Wallis test for the clusters were 0 for each considered variable (coordinate). It allows us to reject the hypothesis that there are no differences between clusters. Meanwhile, post-hoc tests showed that pairs of clusters differed (at the significance level of 0.05) in most cases. In the analysis post-hoc comparisons for the average ranks of all pairs of groups was used (see e.g. Sigel et al., 1998; TIBCO, 2017) and p-values for the two-tailed test with Bonferroni's correction for each compared pair were taken.

Table 2.
P-values of post-hoc tests (columns from 2 to 11, titled "a:b", show p-values of the test between a and b clusters)

Variable\ Clusters	1:2	1:3	1:4	1:5	2:3	2:4	2:5	3:4	3:5	4:5
X coordinate of 1. ch.p.	1	0	0	0	0	0	0	0,001080	0,127108	0
Y coordinate of 1. ch.p.	0	0,000278	0,466919	0	0,000001	0	0	0,000001	0	0,002889
X coordinate of 3. ch.p.	1	0	0	0	0	0	0	0,005044	0,062469	0
Y coordinate of 3. ch.p.	0	0,000029	0,009989	0	0,000191	0	0	0	0	0,250558
X coordinate of 2. ch.p.	0,347494	0	0	0	0	0	0	0,000384	0,849416	0,000001
Y coordinate of 2. ch.p.	0	0,001664	0,016016	0	0	0	0	0	0	0,084254
X coordinate of 4. ch.p.	1	0	0	0	0	0	0	0,034809	0,022110	0,000001
Y coordinate of 4. ch.p.	0,889435	1	0	0,725732	1	0	0,010150	0	0,024790	0,000273
X coordinate of 5. ch.p.	0,729320	0	0	0	0	0	0	0,000195	1	0,000002
Y coordinate of 5. ch.p.	0,000744	1	0	0,007986	0,039138	0	0	0	0,000189	0,071558

Source: Authors' own.

The plot below (Fig. 5), made with Matplotlib.pyplot. Version of Matplotlib: 3.5.0, shows means and standard deviations of the considered coordinates of characteristic points of the epidemic curves in every cluster.

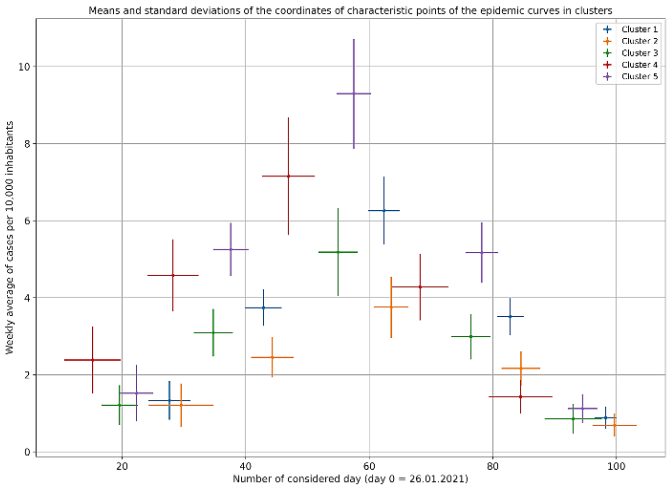

Figure 5. Means and standard deviations of the considered coordinates of characteristic points of the epidemic curves.

On this basis, it is possible to compare the course of the epidemic in every cluster. For example, statistically speaking, powiats from the fifth cluster suffered the most in 3rd wave of COVID-19, and powiats from the second cluster suffered the least (in terms of number of cases per 10,000 inhabitants).

5. Analysis of the relationship with other factors and an attempt to determine predictive models

In this stage of the analysis the following data was considered:

- population density per 1 km sq. in 2020 (in powiats);
- population by age groups in 2020 (in powiats);
- population in cities in% of the total population (in powiats);
- vaccinated population (in powiats).

First three data mentioned above come from Local Data Bank of Statistics Poland, available in the Internet (https://bdl.stat.gov.pl/bdl/start, 13.04.2022). Symbols of the category and the group (in Local Data Bank) from that data comes are respectively K3 and G7, and symbols of subgroups (respectively): P2425, P2463, P2137. URLs of concrete data were unavailable.

The information about the vaccinated population comes from Raport szczepień przeciwko COVID-19, Dane historyczne (eng. COVID-19 vaccination report, historical data), available in the Internet (https://dane.gov.pl/pl/dataset/2538,raport-szczepien-przeciwko-covid-19/resource/34430/table, 13.04.2022).

6. Operations on data and short analysis of their values in clusters

Based on the data about the population in specific age groups, the percentage share of these groups in the total population of a given powiat was determined. Later based on the mentioned population and vaccination data, the approximate percentage of the population vaccinated on August 1, 2021, in given powiats has been calculated (as indicator of society approach to epidemic). These steps were done in Python programming language.

To compare levels of considered factors in clusters, boxplots were made in Statistica environment.

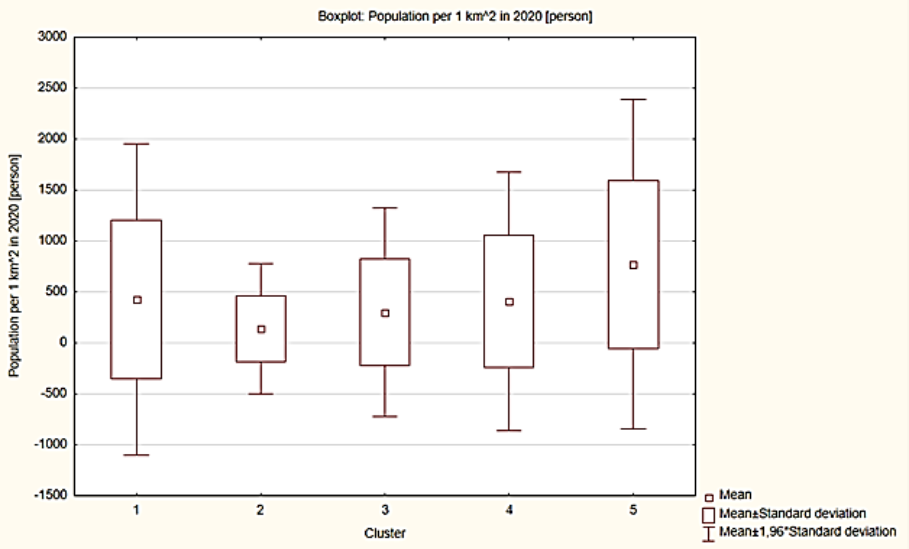

Figure 6. Boxplot 1.

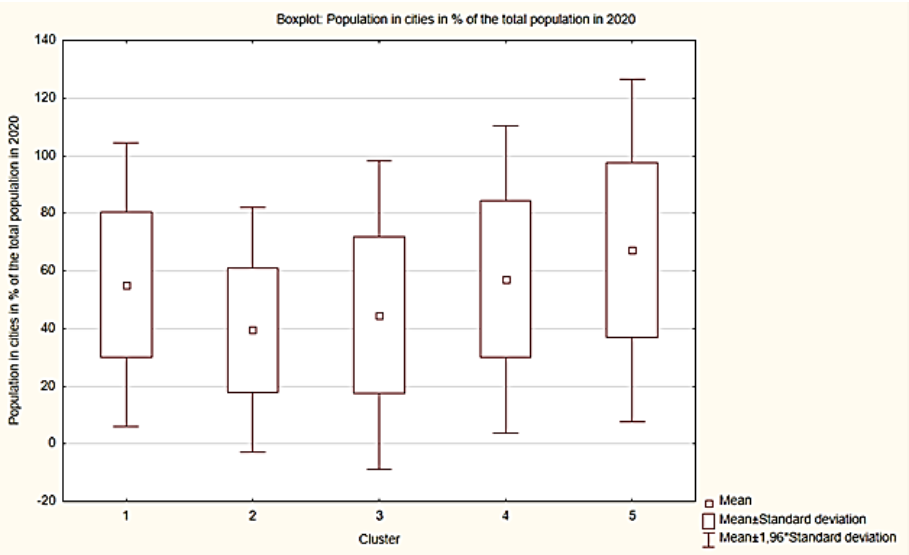

Figure 7. Boxplot 2.

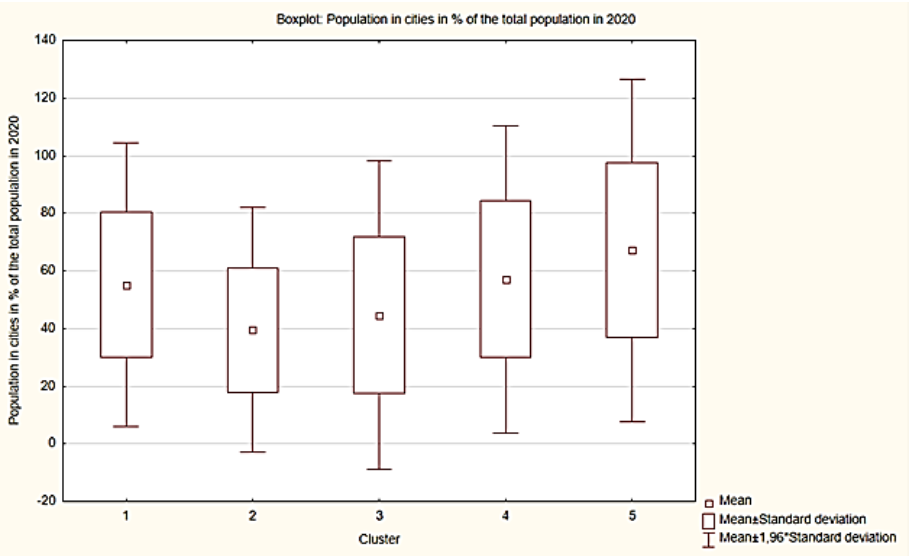

Figure 8. Boxplot 3.

Boxplots shows that the difference between clusters occurs in ranges of standard deviations. To analyze the difference between clusters in considered factors more deeply, Kruskal-Wallis test and post-hoc tests were applied. The classic ANOVA analysis was abandoned because the condition of homogeneity of variance in most cases was not satisfied. To check this condition Levene's and Brown-Forsythe's tests were used.

Table 3.
Results of checking homogeneity of variance, and Kruskal-Wallis test

Variable	P-value of Levene's test	P-value of Brown-Forsythe's test	P-value of Kruskal-Wallis' test
Population per 1km^2	0,000000	0,000058	0
population in cities in% of the total population	0,001332	0,002457	0
% of population in age 0-4	0,055125	0,111768	0,0036
% of population in age 5-9	0,023932	0,025041	0,0129
% of population in age 10-14	0,000130	0,000156	0,2534
% of population in age 15-19	0,001194	0,002659	0
% of population in age 20-24	0,050216	0,103416	0
% of population in age 25-29	0,017470	0,047625	0
% of population in age 30-34	0,006452	0,006751	0,3739
% of population in age 35-39	0,017776	0,164973	0
% of population in age 40-44	0,000002	0,000039	0
% of population in age 45-49	0,000000	0,000076	0,0017
% of population in age 50-54	0,003891	0,009110	0,0098
% of population in age 55-59	0,000005	0,000011	0
% of population in age 60-64	0,006249	0,008411	0,156
% of population in age 65-69	0,050905	0,089036	0,0003
% of population in age 70 and more	0,018269	0,023300	0,1121
% of population in age 70-74	0,003754	0,010453	0
% of population in age 75-79	0,022411	0,043057	0,0421
% of population in age 80-84	0,031218	0,035508	0,6014
% of population in age 85 and more	0,145800	0,160838	0,0597
% of population in age 0-14	0,057461	0,058451	0,0276
Approximated % of population vaccinated against COVID-19 1 August 2021	0,000061	0,000627	0,0002

Source: Authors' own.

Table 4.
P-values of post-hoc tests (columns from 2 to 11, titled "a:b", show p-value of test between a and b clusters)

Variable\ Clusters	1:2	1:3	1:4	1:5	2:3	2:4	2:5	3:4	3:5	4:5
Population per 1km^2	0,000026	0,630083	1,000000	0,000036	0,046110	0,073292	0,000000	1,000000	0,000000	0,000006
population in cities in% of the total population	0,000385	0,011873	1,000000	0,931024	1,000000	0,004056	0,000001	0,049730	0,000058	1,000000
% of population in age 0-4	1	1	1	0,003205	1	1	0,006447	1	0,073277	0,631074
% of population in age 5-9	1	1	1	0,006392	1	1	0,067322	1	0,357379	1
% of population in age 10-14	1	0,417706	0,917887	1	1	1	1	1	1	1
% of population in age 15-19	0,022917	0,000074	0,223467	1	1	1	0,009780	1	0,000051	0,091039
% of population in age 20-24	0	0,000017	0,379794	0,001531	1	0,055360	0	0,638719	0	0,000003
% of population in age 25-29	0	0,002171	1	0,000112	0,066111	0,001179	0	1	0	0,000005
% of population in age 30-34	1	1	0,558777	1	1	1	1	1	1	1
% of population in age 35-39	0,000115	1	1	0,000008	0,047961	0,001868	0	1	0	0,000494
% of population in age 40-44	0	0,018630	1	0,002990	0,080604	0,000180	0	0,348241	0	0,005234
% of population in age 45-49	0,004611	0,045670	0,932833	1	1	1	0,047556	1	0,256082	1
% of population in age 50-54	0,437357	1	1	0,884674	1	0,447016	0,005130	1	0,174725	1
% of population in age 55-59	0,018794	1	1	0,206077	0,385064	0,994277	0,000005	1	0,013059	0,052346
% of population in age 60-64	0,588289	0,274966	1	0,462634	1	1	1	1	1	1
% of population in age 65-69	0,000270	0,004486	0,170404	1	1	1	0,270742	1	1	1
% of population in age 70 and more	1	0,331649	0,269208	1	1	1	1	1	1	1
% of population in age 70-74	0,000014	0,000386	0,012623	0,315325	1	1	0,335752	1	1	1
% of population in age 75-79	1	1	1	0,810500	1	1	0,702522	1	0,119145	0,038024
% of population in age 80-84	1	1	1	1	1	1	1	1	1	1
% of population in age 85 and more	0,481963	1	1	1	0,937631	0,127609	0,156057	1	1	1
% of population in age 0-14	1	1	0,998236	0,017905	1	1	0,166293	1	0,921977	1
Approximated % of population vaccinated against COVID-19 1 August 2021	0,000994	0,154063	1	1	1	0,251513	0,002202	1	0,166014	1

Source: Authors' own.

While searching for the relationship between the type of the epidemic curve (previously determined) and other factors, the following models were tested:

- naive Bayes classifier;
- the method of k-nearest neighbors;
- decision tree(s) in various forms.

The analysis included 356 powiats (in the case of the remaining 3, there was no data on the % of population in cities). Every model was created (and analyzed) in Statistica environment.

7. Results of naive Bayes classifier

Assuming that the predictors have a normal distribution and without divided data into training and test sets - the accuracy of the classification was approximately 40.45% in total. And as below regarding clusters: 1st: 41.75%, 2nd: 39.72%, 3rd: 41.18%, 4th: 8%, 5th: 54.72%. Accuracy should be understand here as a percentage of correctly classified powiats. It is easy to notice, that the accuracy for cluster no. 4 is significantly lower than for other cluster. A similar situation can be observed in the case of trees and random forest model (described later). Probably this is due to a fact, that the cluster no. 4 is the smallest one. Total accuracy equal to circa 40% is not big in the case of the consideration of the prediction model effectiveness but it is enough to conclude that the mentioned factors are connected in some way with the shape of the epidemic curve that was the main goal of analysis.

8. Results of the method of k-nearest neighbors

To determine the k-nearest neighbors model Euclidean distance, data standardization, random test sample size of 25% and distance weighting were used. Number of neighbors (5) was chosen with cross-validation test. Accuracy in classification in test sample was approximately 40.23% And as below regarding clusters: 1st: 32%, 2nd: 37.93%, 3rd: 45%, 4th: 50%, 5th: 55.56. The total accuracy is similar as in the naive Bayes model, however noticeable differences in accuracies in case of clusters can be observed. Especially in the case of the cluster no. 4, that probably follows from the fact that in the case of the k-nearest neighbors model the size of a cluster is not as important/influential as in the naive Bayes model.

9. Results for single decision trees

The accuracy of this model is strongly depended on the parameters adopted when creating the tree (regarding its size, etc.). Trees were made in the module General Classification and Regression Trees (C&RT) in the program Statistica. The following tree creation settings/parameters/conditions were applied:

- equal cost of incorrect classification;
- the Gini measure as an indicator of goodness/suitability of fit;
- estimated a priori probability;
- FACT direct stopping as a stop rule.

The remaining parameters differed between trees: minimal capacity of the leaf, object fraction (parameter of FACT method) and maximal number of nodes. More details about creating process can be found in (TIBCO, 2017) and in sources indicated there.

In the table below (Table 5) there are accuracies of chosen trees in total and in context of concrete clusters, compiled with some parameters and number of nodes. Results regards to the training set (test set was not created in this case).

Table 5.
Accuracies of chosen decision trees

Min. capacity of the leaf	Object fraction	Max. number of nodes	1 acc.	2 acc.	3 acc.	4 acc.	5 acc.	Total acc.	Number of decision nodes	Number of end nodes/ leaves
40	0.1	15	40.91	82.93	17.44	13.64	75	46.91	7	8
30	0.1	20	62.50	43.9	55.81	29.55	75	54.49	10	11
20	0.1	50	56.82	67.07	53.49	29.55	75	57.87	13	14
30	0.05	50	64.77	67.07	56.98	45.45	80.36	63.48	20	21
20	0.05	50	64.77	59.76	68.6	45.45	80.36	64.61	22	23
20	0.01	70	72.73	62.2	69.77	47.73	80.36	67.7	35	36
10	0.01	100	79.55	84.15	79.07	52.27	80.36	77.25	50	51

Source: Authors' own.

Observation indicates that, generally speaking, the bigger tree, the better accuracy. Of course, in the same time the risk of overfitting model is increasing. The results obtained here, as in the previous models, indicate the presence of connections between the type of the epidemic curve (determined before) and other considered factors, but in this case not necessarily every factor (because some factors can be not present in a given tree).

10.Results for random forest

As in the case of a single decision tree, also in that method results were strongly dependent on the parameters adopted during creating the model. Random forests were made in the program Statistica. The following random forest creation settings/parameters/conditions were applied:

- equal costs of incorrect classification;
- number of predictors = 23 (max);
- proportion of test sample = 0.3;
- proportion for subsamples = 0.5;
- the initial value of the random number generator = 1.

The remaining parameters differed between forests: the form of determining a priori probability (equal or estimated) number of trees in forest, minimal capacity of the leaf, minimal capacity of the descendant, maximal number of levels and maximal number of nodes. More details about creating process can be found in (TIBCO, 2017) and in sources indicated there.

In the table below (Table 6) there are accuracies of chosen random forests in total and in context of concrete clusters, compiled with some parameters.

Table 6.
Accuracies of chosen random forests

Number of trees	Min. capacity of the leaf	Min. capacity of descendant	Max. number of levels	Max. number of nodes	1 acc.	2 acc.	3 acc.	4 acc.	5 acc.	Total acc.	Prob. a priori
100	8	5	10	100	56.76	43.48	40.91	0	62.5	43.86	estim.
50	10	5	6	50	56.76	65.22	27.27	0	56.25	44.74	estim.
200	5	3	15	500	43.24	52.17	45.45	0	62.5	42.11	estim.
100	8	5	10	100	48.65	56.52	31.82	6.25	68.75	43.86	equal
50	10	5	6	50	48.65	39.13	40.91	25	68.75	44.74	equal
200	5	3	15	500	51.35	52.17	40.91	6.25	62.5	44.74	equal

acc. - accuracy, min – minimum, max – maximum, estim. - estimated, prob. - probability

Source: Authors' own.

This time accuracies were checked in test set of powiats. The total accuracy was similar in every case, but accuracies in context of clusters changed significantly for other parameters of forests. It is easy to notice that for the cluster no. 4 the accuracy was always low, and not zero only when a priori probability was assumed to be equal, so it indicates that small number of powiats in this cluster can cause such situation.

11.Conclusions

1. The shape of the COVID-19 epidemic curves in most Polish powiats (during the third wave of the epidemic) corresponded to the bell curve (after the mentioned modifications). Only 21 powiats did not have such epidemic curve shape.
2. The curves can be divided into 5 types indicating the course of the epidemic in time.
3. The performed analysis allows to assume that the mentioned curve types depend in significant way on factors such as the age structure of the population, population density, % of population in cities and the approach of the population to vaccination/epidemic; if this were not the case, the accuracy of the models would not be significantly higher than circa 20% what is approximated accuracy of random selection.
4. The determined predictive models may be a tool supporting the prediction of epidemic development, but their effectiveness is moderate.
5. The performed analysis may be the basis for further, more-depth research.

References

1. Banerjee, D. (2020). The impact of Covid-19 pandemic on elderly mental health. *International Journal of Geriatric Psychiatry, vol. 35, no. 12,* 1466-1467.
2. Bittihn, P., Golestanian, R. (2020). Stochastic effects on the dynamics of an epidemic due to population subdivision. *Chaos, vol. 30,* https://doi.org/10.1063/5.0028972
3. Britton, T. (2020). Epidemic models on social networks - with inference. *Statistica Neerlandica, vol. 74,* 222-241.
4. Chen-Charpentier, B.M., Stanescu, D. (2010). Epidemic models with random coefficients. *Mathematical and Computer Modelling, vol. 52, no. 7-8,* 1004-1010.
5. Fraser, Ch., Riley, S., Anderson, R.M., Ferguson, N.M. (2004). Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences, vol. 101, no. 16,* 6146-6151.
6. Meyer, H., Prescott, B., Sheng, X.S. (2022)., The impact of the COVID-19 pandemic on business expectations. *International Journal of Forecasting, vol. 38, no. 2,* 529-544.
7. Richards, M., Anderson, M., Carter, P. et al. (2020). The impact of the COVID-19 pandemic on cancer care. *National Cancer, vol. 1,* 565-567.
8. Sarkodie, S.A., Owusu P.A. (2021). Impact of COVID-19 pandemic on waste management. *Environment, Development and Sustainability, vol. 23,* 7951-7960.
9. Shad, Y., Ulvi, O., Khan, M.H., Karamehic-Muratovic A. et al. (2020). The impact of COVID-19 on globalization. *One Health, vol. 11, art. no. 100180*.
10. Shen, H., Fu, M., Pan, H., Chen, Z.Y.Y. (2020). The impact of the COVID-19 pandemic on firm performance. *Emerging Markets Finance and Trade, vol. 55, no. 10,* 2213-2230.
11. Siche, S. (2020). What is the impact of COVID-19 disease on agriculture? *Scientia Agropecuaria, vol. 11, no. 1,* http://dx.doi.org/10.17268/sci.agropecu.2020.01.00
12. Siegel, S., Castellan, N.J. (1998). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.
13. Sneha, G., Luc, H. (2020). COVID-19: impact by and on the environment, health and economy. *Environment, Development and Sustainability, vol. 22,* 4953-4954.
14. Tarkar, P. (2020). Impact of Covid-19 pandemic on education system. *International Journal of Advanced Science and Technology, vol. 29, no. 9s,* 3812-3814.
15. TIBCO Software Inc. (2017). Statistica (data analysis software system), version 13. http://statistica.io.
16. Verma, A.K., Prakash, S. (2020). Impact of COVID-19 on environment and society. *Journal of Global Biosciences, vol. 9, no. 5,* 7352-7363.