

The Second Wave of the COVID-19 Pandemic in Poland – Characterised Using FDA Methods

Patrycja Hęcka

Wrocław University of Science and Technology, Wrocław, Poland

e-mail: patrycja.hecka@gmail.com

ORCID: 0009-0002-3463-1882

© Patrycja Hęcka

This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/4.0/>

Quote as: Hęcka, P. (2023). The Second Wave of the COVID-19 Pandemic in Poland – Characterised Using FDA Methods. *Econometrics. Ekonometria. Advances in Applied Data Analysis*, 27(3).

DOI 10.15611/eada.2023.3.02

JEL Classification: C38, C14

Abstract: The aim of this article was to analyse functional data of the number of hospitalised individuals, intensive care patients, positive COVID-19 tests, deaths and convalescents during the second wave of the COVID-19 pandemic in Poland. For this purpose, firstly the author convert data of sixteen voivodeships to smooth functions, and then used the principal component analysis and multiple function-on-function linear regression model to predict the number of hospitalised and intensive care patients due to the COVID-19 infection during the second wave of the pandemic. Finally, the results were compared with those previously obtained for the combined data of the second and third wave of the COVID-19 pandemic in Poland (Hęcka, 2023).

Keywords: function-on-function regression, functional data analysis (FDA), COVID-19, functional principal component analysis, smooth functions

1. Introduction

The SARS-CoV-2 virus emerged at the end of 2019 in China, and quickly became a global problem. The first case of COVID-19 in Poland was identified on 4 March 2020, with a 66-year-old man hospitalised in Zielona Góra. This article examined the second wave of the pandemic in sixteen Polish voivodeships, which lasted five months – from September 2020 to January 2021. The peak occurred on November 7 with a record increase (from the beginning of the pandemic) – 27,875 infections. Due to the unavailability of the published data to the public, in this article the author assumed that the second wave lasted from 23 October 2020 to 15 February 2021.

An epidemic of this scale is a relatively a new phenomenon, thus it has attracted the attention of many analysts. Numerous articles have been written on the subject of the pandemic in Europe – for example in France (Oshinubi et al., 2022), Italy (Boschi et al. 2021) and Spain (Acal et al., 2021). Multiple methods were used to study the COVID-19 pandemic, for example time-series modelling (Tandon et al., 2022), probabilistic models (Zhou and Ji, 2020), machine learning (Khanday et al., 2020) or deep learning (Shuller et al., 2021). Many of the methods used to analyse the influence of the COVID-19 pandemic were collected and described by Cao and Liu (2022). Hęćka (2023) presented a comparison of the second and third wave of the pandemic in Poland, and that article provided the starting point for this paper, the aim of which was to focus more on the second wave of the pandemic.

The purpose of the study to predict the number of hospitalised individuals and intensive care patients during the second wave of the pandemic in selected Polish voivodeships. In addition, the author provided analysis of the functional principal components and compared the new results with the predictions obtained earlier (Hęćka, 2023).

The results in this work were obtained using software R-packages ‘ggplot2’ (Wickham et al., 2016) and ‘fda’ (Ramsay et al., 2020).

2. Data description and conversion

The data were collected and published by M. Rogalski (2022).

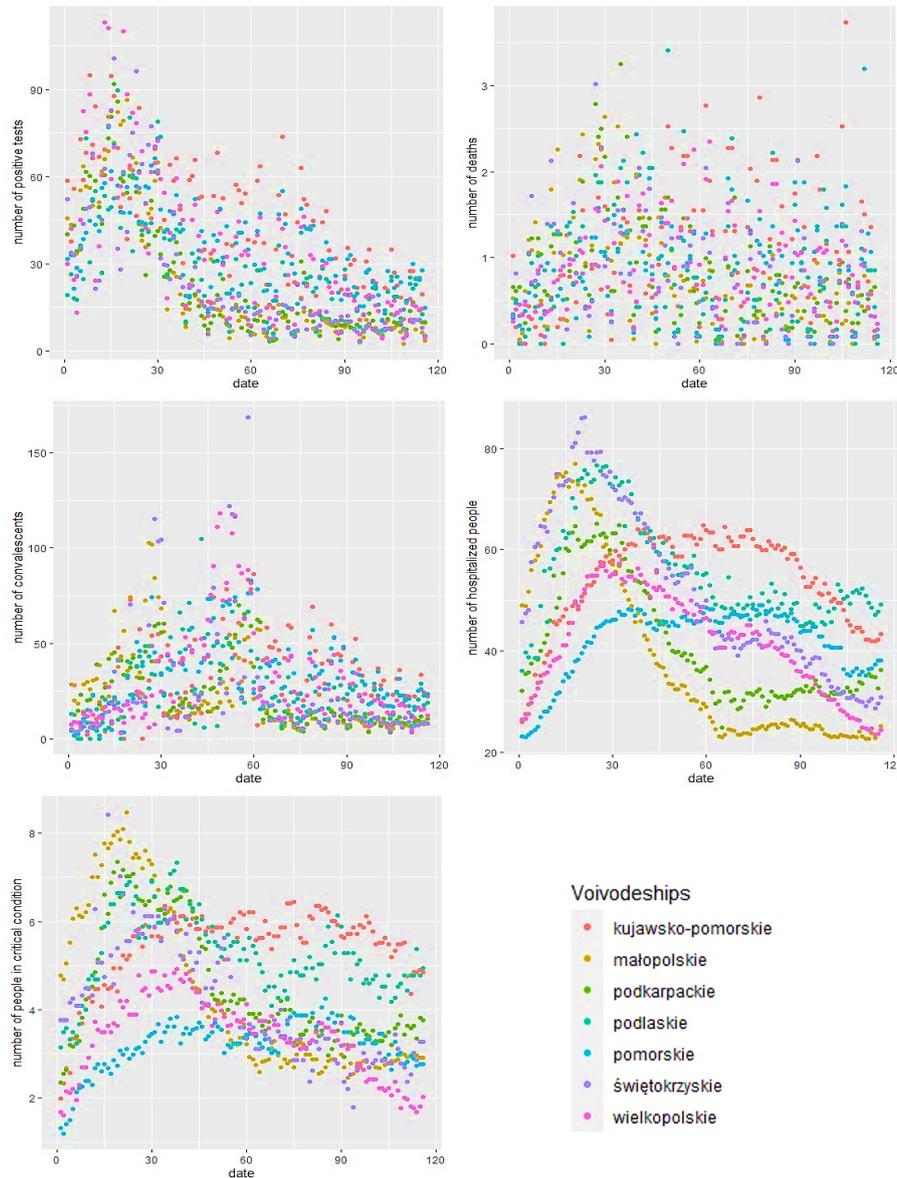
All the curves were transferred to the same interval $[0, 1]$, denoting the original time as $T \in [0, \max(T)]$, any response variable as z and the transformed response variable as z^* ; $\max(T)$ defined the maximum value of T among all the curves for which that value is the same. The registration equation is then

$$z^*(u) = z(u \cdot \max(T)), \forall u \in [0,1].$$

This transformation does not affect the analysis presented in this article, as transforming the data did not impact on the patterns observed in the analysis. The transformation simplifies the analysis, by making it easier to apply traditional, functional data analysis techniques, and makes the figures more readable.

For comparison with the results obtained for the combined data of the second and third wave (Hęćka, 2023), to the test sample of the model analysed in the next sections for the following voivodeships: Świętokrzyskie, Wielkopolskie, Podkarpackie and Małopolskie. These voivodeships were chosen by the author (2023), because the analysis showed that the largest number of hospitalised individuals and patients in a serious condition per 100,000 inhabitants were noted in Świętokrzyskie voivodeship, the lowest number of people were hospitalised in Wielkopolskie voivodeship, while the fewest positive COVID-19 tests were noted in Podkarpackie and the least deaths in Małopolskie. The rest of 12 Polish voivodeships (12) form the reference group.

A number of methods were used to convert data to smooth functions (de Boor, 1978; Ramsay and Silverman, 2010; Ullah and Finch, 2013). Figure 2 shows the fitting of 15 basis functions with equally spaced knots to approximate seven curves showing the number of positive tests, deaths, recoveries, hospitalised individuals and people in a critical condition; the coefficients were fitted by the least squares method (Ramsay and Silverman, 2010, Section 4.2). The author chose the number of basis functions to make the value of the mean squared error the smallest (achieving “bias-variance trade-off”) and to avoid overfitting the model. In cases of converting the combined data of the two waves, 20 basis functions were chosen (Hęćka, 2023).



The number of people was divided by the number of inhabitants of a given voivodeship and multiplied by 100,000

Fig. 1. Daily observations of the number of positive COVID-19 tests, deaths, convalescents, hospitalised people and people in a critical condition during the second wave of the pandemic (23 October 2020 to 15 February 2021) in selected Polish voivodeships

Source: own elaboration using software R and data collected by M. Rogalski (2022).

3. Principal component analysis for functional data

In this section the author used principal component analysis (PCA) for functional data to analyse the COVID-19 data; PCA in the functional context was described in detail by Ramsay and Silverman (2010, Section 8). Let us denote by $X_1(t), X_2(t), X_3(t), Y_1(t), Y_2(t)$ the variability over time, respectively: number of positive COVID-19 tests, deaths, recoveries, hospitalised and intensive care patients. The first principal components explain respectively, approximately 68%, 54%, 57%, 58%, 76% of the variability of $X_1(t), X_2(t), X_3(t), Y_1(t), Y_2(t)$ for voivodeships from the reference sample. The second principal components describe 15%, 20%, 24%, 32% and 16% of the variability, and the third principal components – 10%, 8%, 7%, 6% and 3%.

The first and second principal components for the number of hospitalised people taken together cover 89.79% of variance. In turn, for the number of people in a serious condition, already the single first component refers to 76.06%, and together with the second, they cover 92.43%. These are significantly higher values than in the case of the joint analysis for the second and third wave (Hęćka, 2023, Section 4.1.1), the reason being that the model proposed by Hęćka (2023) was more complex because it combined data from two different waves of the pandemic.

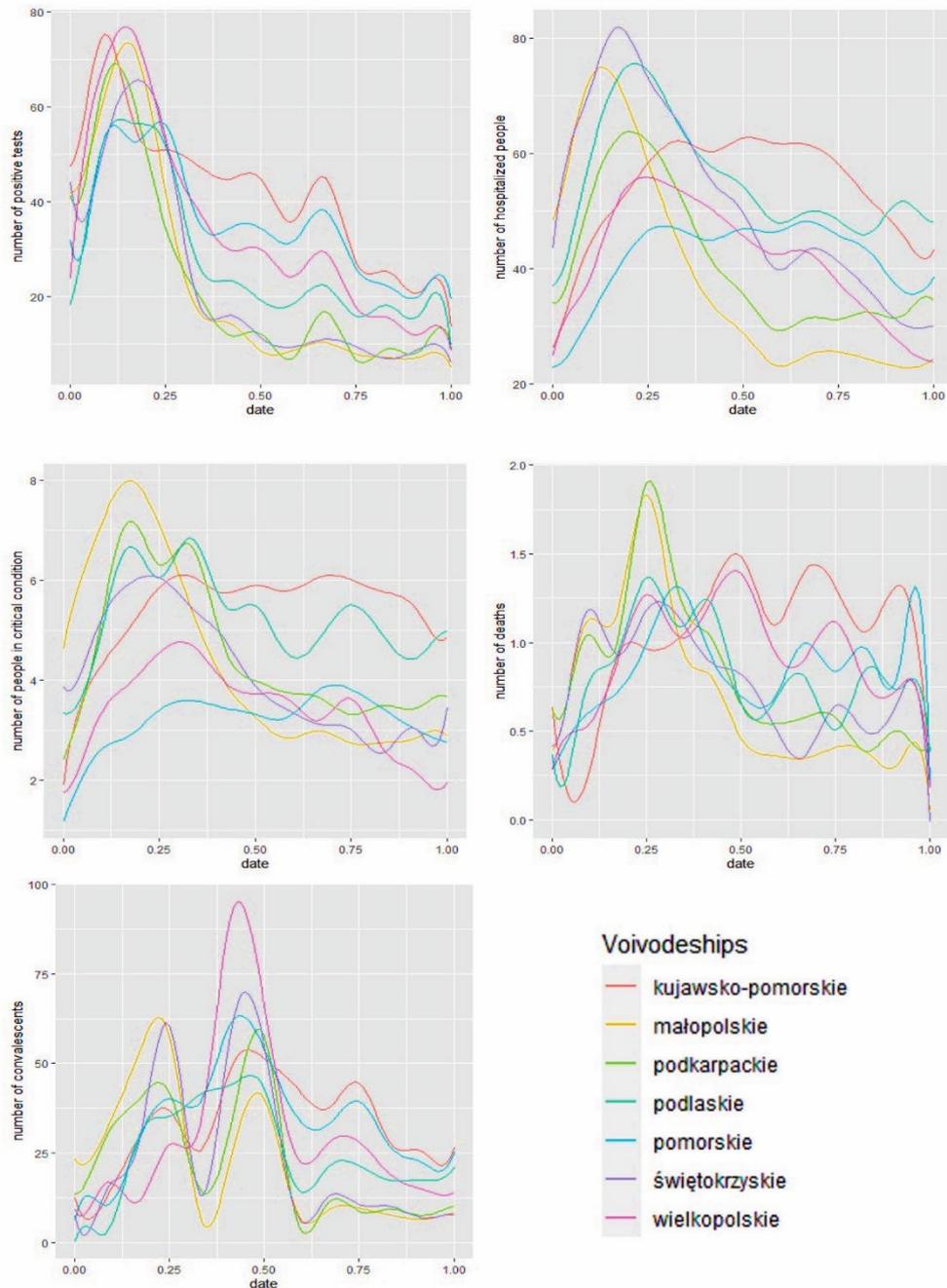


Fig. 2. Fitted curves to daily observations of the number of positive COVID-19 tests, deaths, convalescents, hospitalised individuals and people in a critical condition during the second wave of the pandemic in Poland (23 October 2020 to 15 February 2021) for selected Polish voivodeships (15 basis functions were used)

Source: own elaboration using software R and data collected by M. Rogalski (2022).

3.1. Mean curve

This subsection presents the plots of the mean curve together with the functions obtained by addition and subtraction of the appropriately multiplied weight functions from the mean. Figures 3 and 4 show such plots for the number of hospitalized and intensive care people.

Figures 3 and 4 suggest that the first principal components present the overall trend in the number of hospitalised people and patients in a serious condition. The second principal components show the differences between the beginning and the end of the second wave of the pandemic.

In the case of a joint analysis (Hęćka, 2023), the first principal component for the number of hospitalised people shows the differences between the second and third wave of the pandemic, while the second principal component for the general number of hospitalised people. In turn, for the number of people in a critical condition, the first principal component shows the overall trend and the second – the differences.

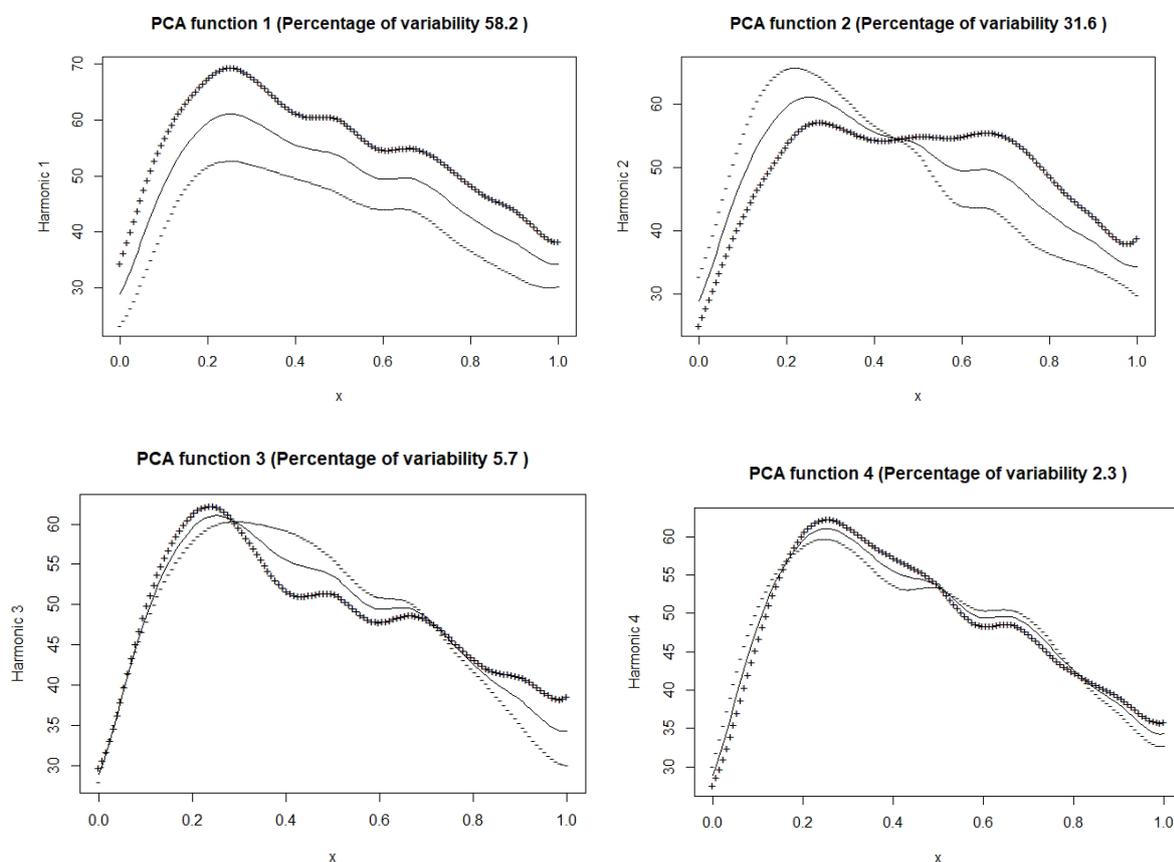


Fig. 3. Mean curve of hospitalized people with curves resulting from adding (+) and subtracting (–) appropriately scaled harmonic coefficients from the mean

Source: own elaboration using software R and data collected by M. Rogalski (2022).

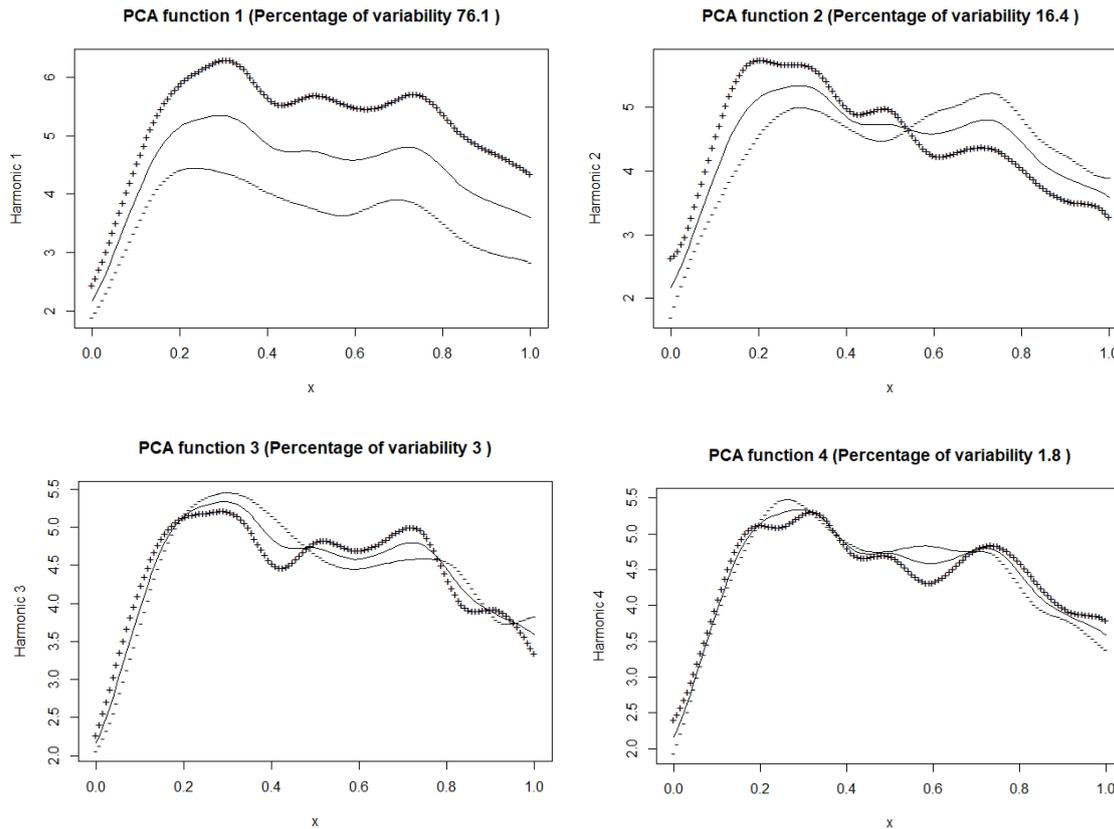


Fig. 4. Mean curve of the number of people in a critical condition with curves resulting from adding (+) and subtracting (–) appropriately scaled harmonic coefficients from the mean

Source: own elaboration using software R and data collected by M. Rogalski (2022).

3.2. Plotting principal component scores

This subsection presents plots with values of the variables achieved on the first and second scores. These are interesting plots for functional principal component analysis (Ramsay and Silverman, 2010, Section 8.3.2).

From subsection 3.1. (Figures 3 and 4), one can conclude that the voivodeships which are placed on the right sides of Figures 5 and 6 should reach the highest total number of hospitalised people and patients in a serious condition. The voivodeships on the left side of these charts should note the lowest during the second wave of the pandemic. Interestingly, Lubelskie voivodeship had both the highest number of hospitalised people and patients in a serious condition during the second wave of the pandemic. The smallest number people were hospitalized in Śląskie and Pomorskie voivodeships, and the lowest number of patients in a serious condition were recorded in Opolskie and Pomorskie.

One can observe that the biggest differences in the number of hospitalised people between the beginning and end of the second wave of the pandemic were in Kujawsko-Pomorskie and Warmińsko-Mazurskie voivodeships, while in Opolskie that difference was the smallest. In turn, in Lubelskie and Śląskie voivodeships the largest differences in the number of people in a critical condition between the beginning and end of the second wave were observed, whereas in Warmińsko-Mazurskie and Pomorskie – the smallest.

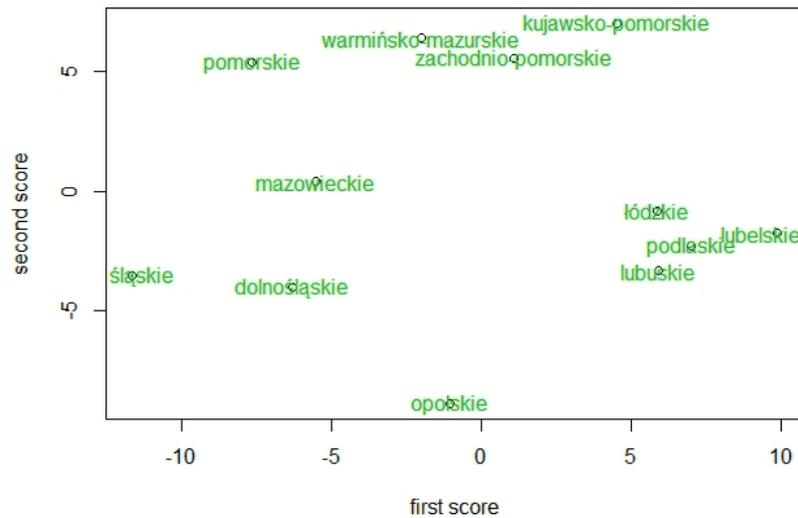


Fig. 5. Values of the first and second scores for the number of hospitalised patients during the second wave of the pandemic (voivodeships from the reference sample)

Source: own elaboration using software R and data collected by M. Rogalski (2022).

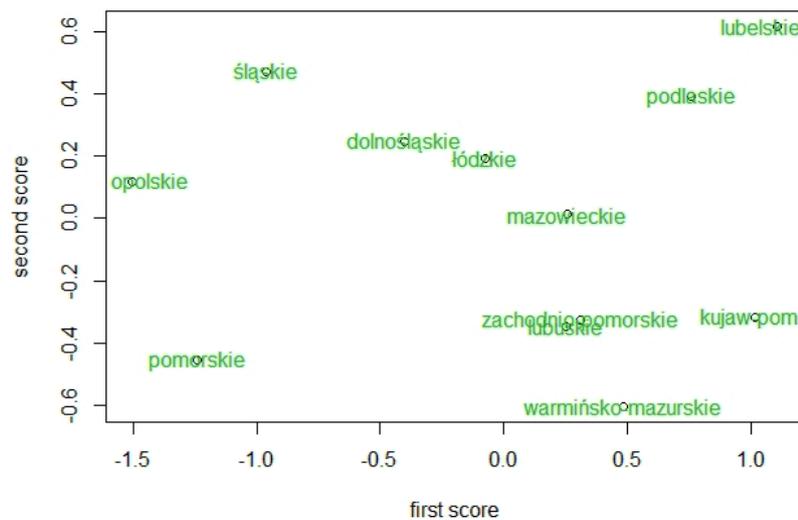


Fig. 6. Values of the first and second scores for the number of people in a critical condition during the second wave of the pandemic (voivodeships from the reference sample)

Source: own elaboration using software R and data collected by M. Rogalski (2022).

4. Prediction of the number of hospitalised and intensive care patients

In this section the author used the multiple function-on-function linear model to predict the number of hospitalised people and patients in a critical condition during the second wave of the COVID-19 pandemic in Poland. The predictions were made according to the multiple function-on-function linear model, presented by equation (1):

$$\hat{y}_{ik}(t) = \bar{y}_k(t) + \hat{\xi}_{i1}^{y_k} f_1^{y_k}(t), k = 1, 2; i = 1, \dots, 16, \quad (1)$$

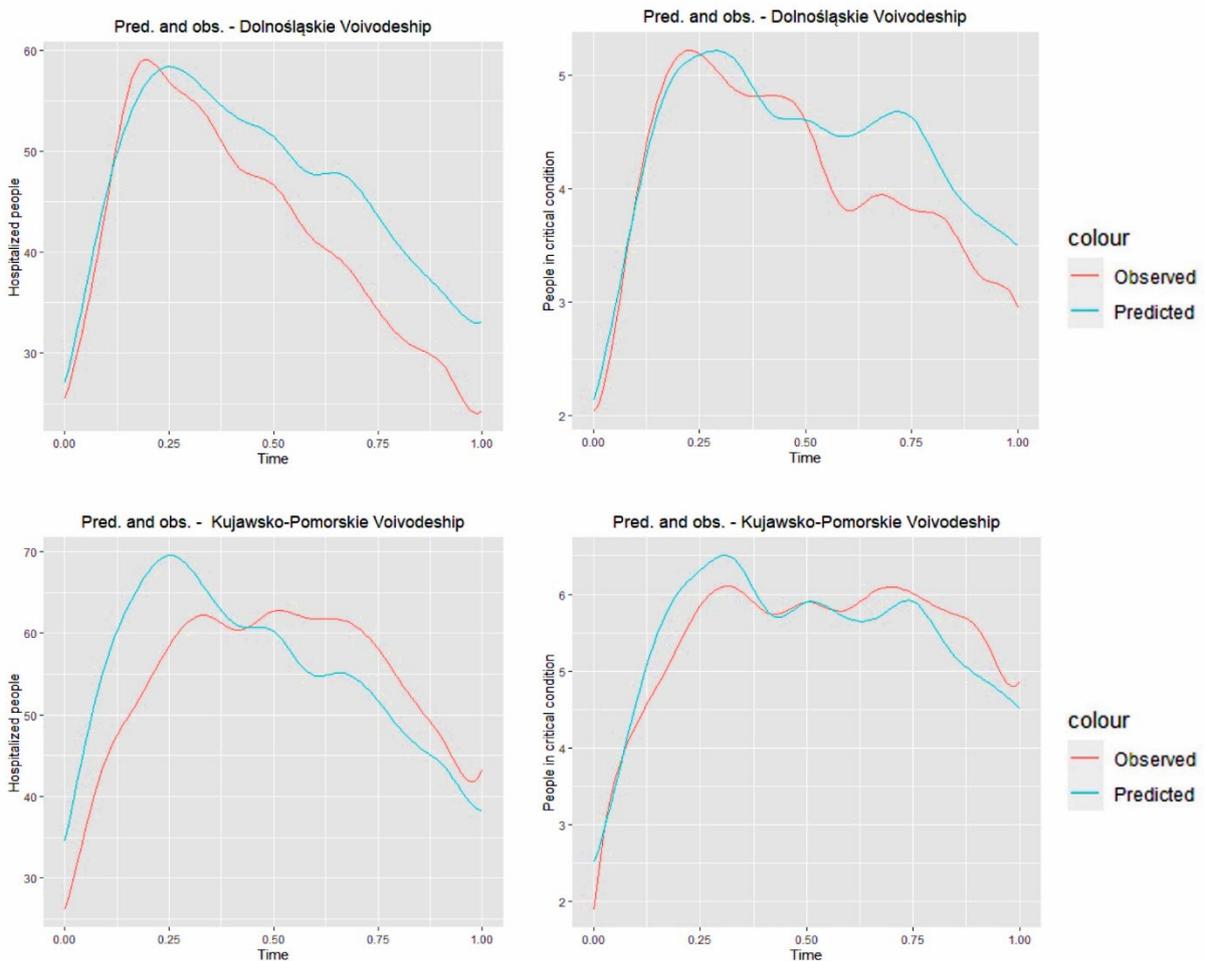
where \bar{y}_k is the mean of the number of hospitalised people (for $k = 1$) and the number of intensive care patients ($k = 2$) for voivodeships from the reference group (12 voivodeships). Functions $f_1^{y_k}(t)$ are the eigenfunctions (weight functions) of the sample covariance. The estimators of the principal component scores for y_1 and y_2 are describe by equation (2):

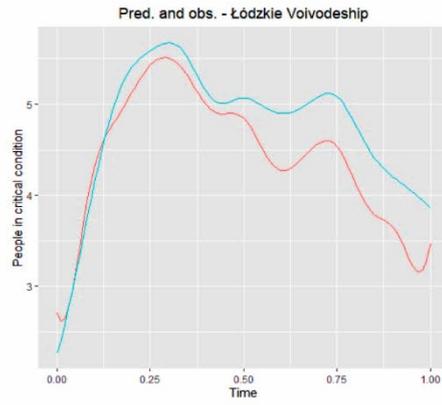
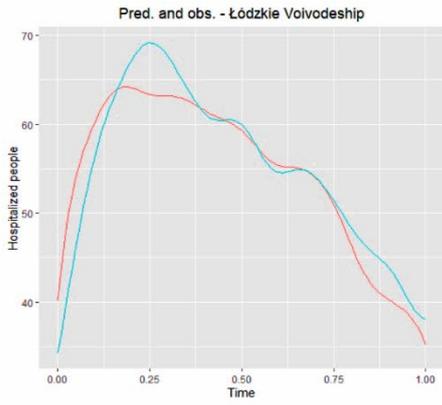
$$\hat{\xi}_{i1}^{y_k} = \delta_0 + \xi_{i1}^{x_1} \delta_1^{y_k} + \xi_{i1}^{x_2} \delta_2^{y_k} + \xi_{i1}^{x_3} \delta_3^{y_k} + \epsilon_1^{y_k}, k = 1, 2; i = 1, \dots, 16, \quad (2)$$

where $\delta_0, \delta_1^{y_k}, \delta_2^{y_k}, \delta_3^{y_k}$ are coefficients obtained by fitting linear model to the data and $\epsilon_1^{y_k}$ are independent functional errors. The multiple function-on-function linear model was described in detail by Acal, Escabias, Aguilera and Valderrama (2021) and by Cai, Xue and Cao (2022).

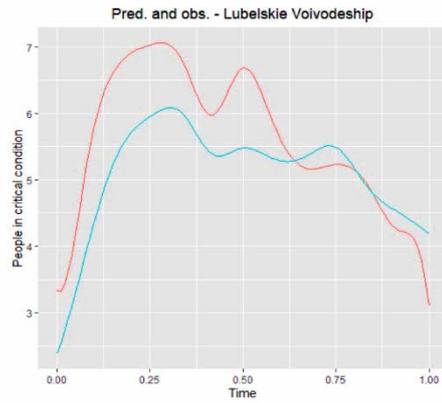
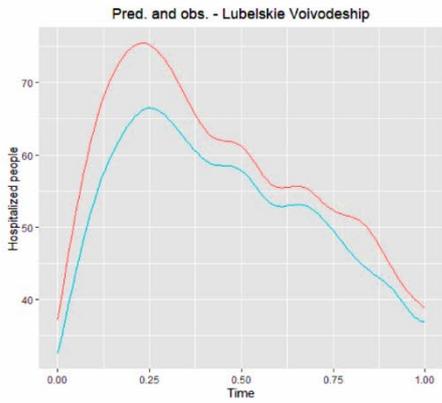
4.1. Reference sample

First, the model was tested on the reference sample. Figure 7 shows the plots of the observed curves (Figure 1) with the plots of the predicted curves for the selected voivodeships from the reference sample. Table 1 presents the values of the mean squared error calculated for all the voivodeships from the reference sample.

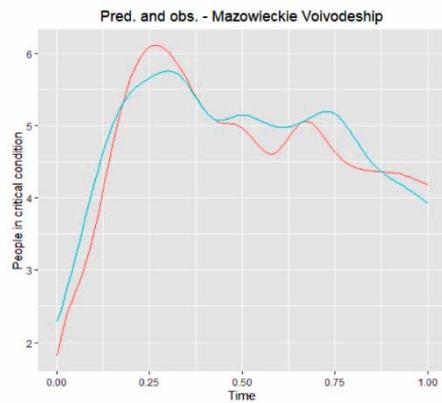
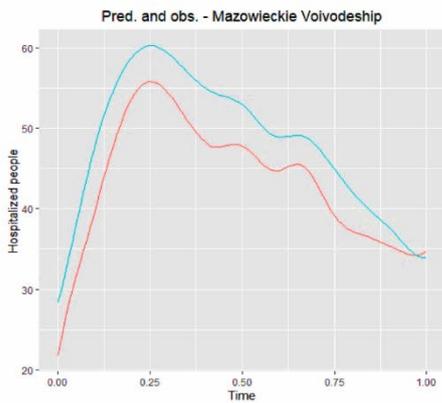




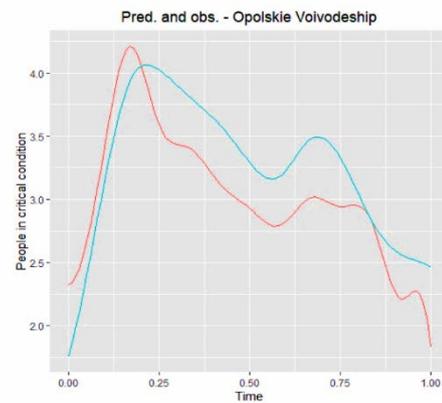
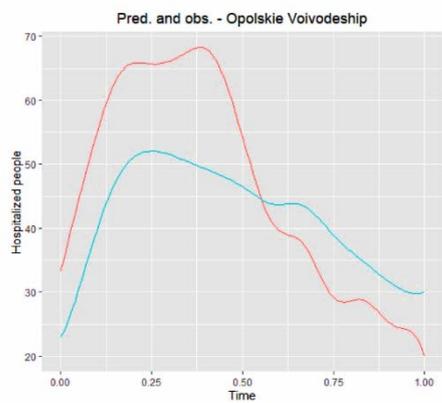
colour
— Observed
— Predicted



colour
— Observed
— Predicted



colour
— Observed
— Predicted



colour
— Observed
— Predicted

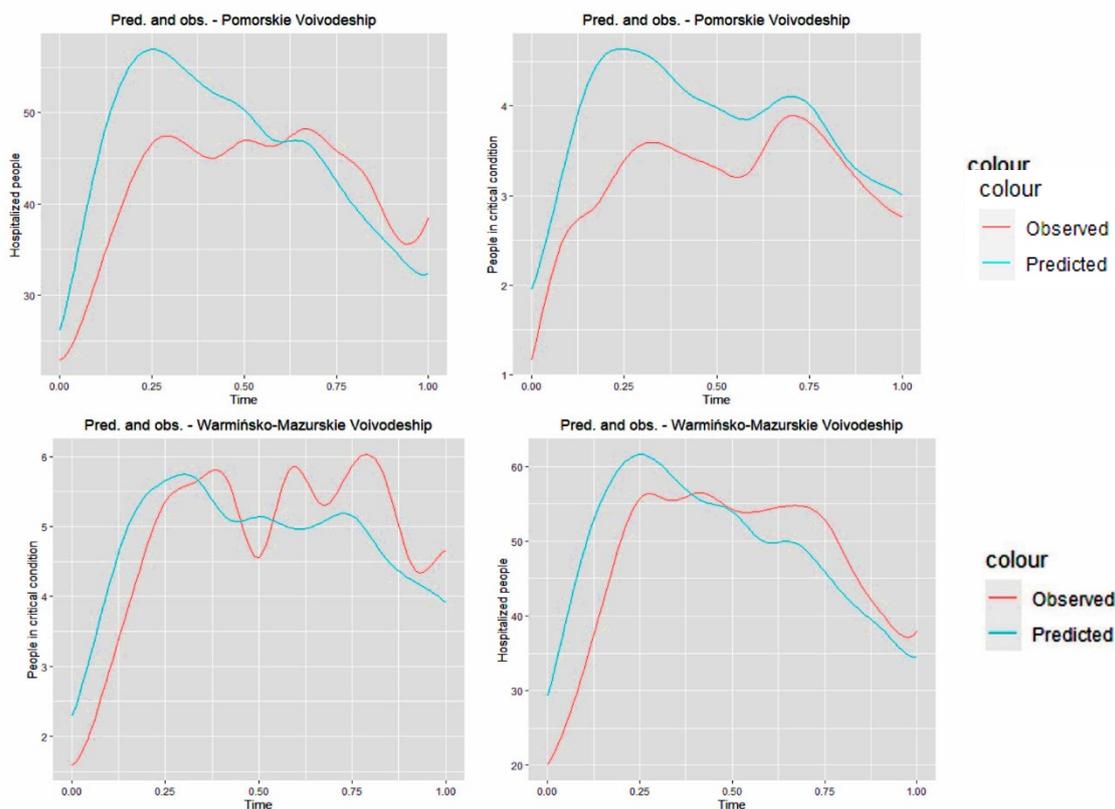


Fig. 7. Observed and predicted curves for selected voivodeships from the reference sample

Source: own elaboration using software R and data collected by M. Rogalski (2022).

Table 1. Values of the mean squared error for y_{i1} (the number of hospitalised people) and y_{i2} (the number of people in a serious condition) for the voivodeships from the reference sample for the second wave of the pandemic

Voivodeship	MSE(y_{i1})	MSE(y_{i2})
Dolnośląskie	5.998808	0.4161685
Kujawsko-Pomorskie	7.488025	0.3637523
Łódzkie	3.205443	0.4343277
Lubelskie	6.017699	0.8045232
Lubuskie	5.393380	0.6992032
Mazowieckie	5.169655	0.3131046
Opolskie	11.636583	0.3634391
Podlaskie	10.526804	1.3927908
Pomorskie	7.175780	0.7657089
Śląskie	9.430756	0.5278728
Warmińsko-Mazurskie	7.195717	0.7103345
Zachodniopomorskie	6.872950	0.4965603

Source: own elaboration using software R and data collected by M. Rogalski (2022).

For almost all the voivodeships, compared to the model for both waves (Hęćka, 2023, Table 1), MSE(y_{i2}) decreased. The exception was the Podlaskie voivodeship, where MSE(y_{i2}) was approximately equal to 1.0092404, and for the single wave model it was approximately equal to 1.3927908.

The values of $MSE(y_{i1})$ decreased only for the following voivodeships: łódzkie (Hęćka, 2023; $MSE(y_{i1}) = 8.146502$), Lubelskie (8.434905), Lubuskie (10.79992) and Pomorskie (7.526031). The other voivodeships had lower values of $MSE(y_{i1})$ during the joint analysis of the pandemic waves (Hęćka, 2023).

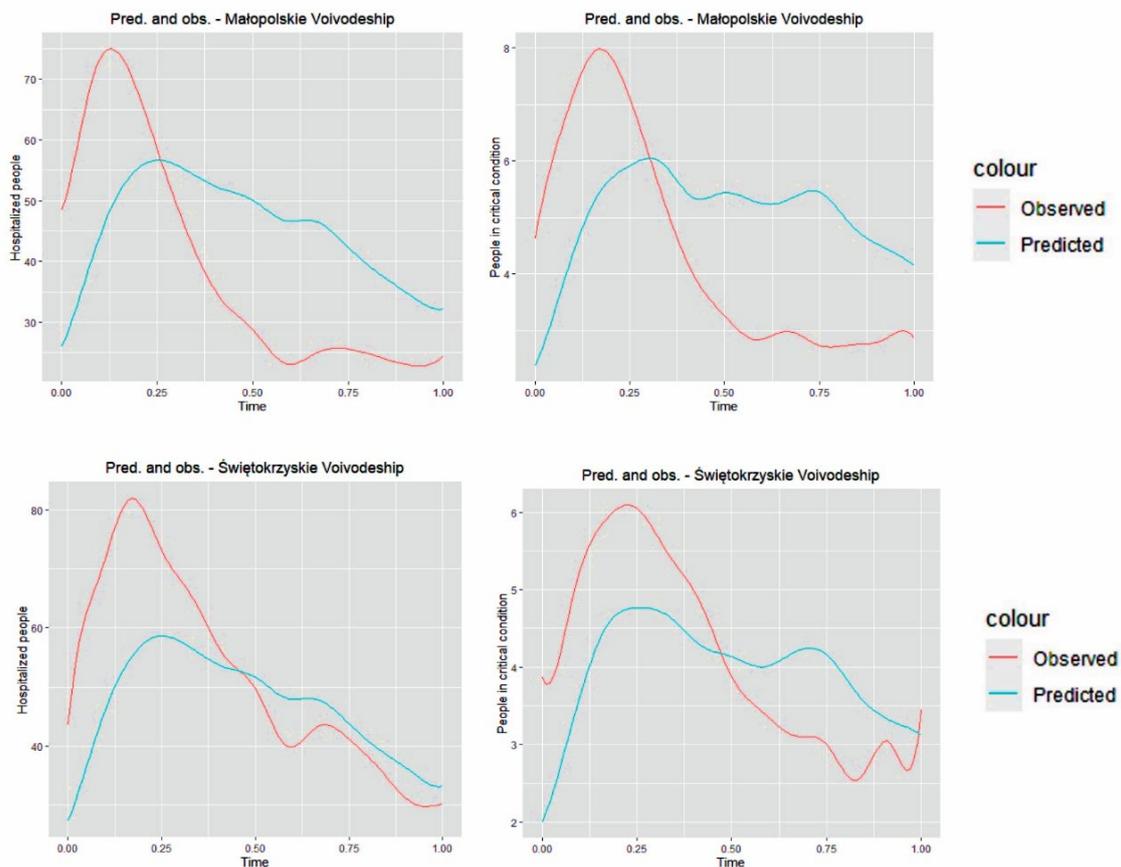
From this one can draw the conclusion that the model proposed by Hęćka (2023) turned out to be a better fit to the data of the number of people hospitalised during the second wave in Poland, than the model proposed in this article. In analysing the number of people in a serious condition during the second wave, better results were achieved by using the model proposed in this article.

The single second-wave model for the number of hospitalised people fits best to the data for łódzkie voivodeship. The received MSE was equal to 3.205443, whereas the lowest $MSE(y_{i2})$ was obtained for Mazowieckie voivodeship and the highest value of $MSE(y_{i1})$ was for Opolskie, and of $MSE(y_{i2})$ for Podlaskie.

4.2. Test sample

Analysing Figure 8 and Table 2, one can see that for Małopolskie, Podkarpackie and Wielkopolskie voivodeships, the predicted number of people in a serious condition was closer to the values collected by Rogalski (dataset), compared to the analysed data from both waves of the pandemic (Hęćka, 2023, Table 2). The exception was Świętokrzyskie, where the prediction became worse, whilst the prediction of the number of hospitalised people was worse for all the voivodeships from the test group.

In both research studies the most difficult to predict turned out to be Małopolskie voivodeship, with the lowest number of deaths. In Figure 2 one can observe that in the second half of the second wave there were few positive test results and recoveries. Based on this information, the model predicted a low number of hospitalised individuals and critically ill patients, however there were actually a lot of hospitalised people and intensive care patients.



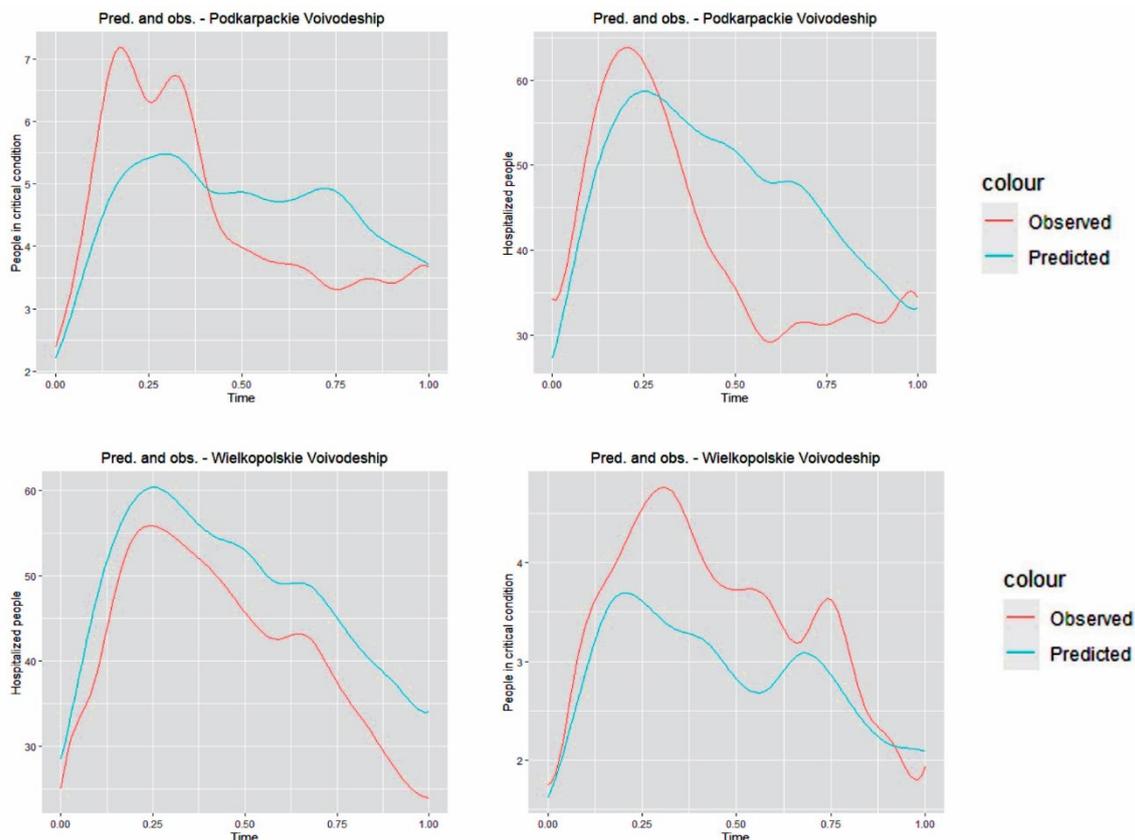


Fig. 8. Observed and predicted curves for the selected voivodeships from the test sample

Source: own elaboration using software R and data collected by M. Rogalski (2022).

Table 2. Predicted values of the number of hospitalised people and in a serious condition compared to the observed values for the second wave of the pandemic

Hospitalised people				
	Małopolskie	Podkarpackie	Świętokrzyskie	Wielkopolskie
Date	obs/pred*	obs/pred*	obs/pred*	obs/pred*
23.10	1671/884	681/579	559/334	924/994
24.10	1667/926	750/608	581/350	929/1045
25.10	1763/973	757/640	589/368	961/1101
26.10	1907/1025	796/674	704/388	1039/1160
27.10	1972/1080	819/710	742/409	1090/1222
28.10	2003/1137	813/747	750/430	1137/1285
29.10	2072/1196	848/785	792/452	1182/1349
30.10	2250/1256	883/823	789/474	1241/1412
31.10	2289/1315	917/860	779/496	1252/1475
1.11	2366/1374	976/897	829/517	1292/1535
...
6.02	786/1161	662/752	380/433	934/1280
7.02	784/1148	685/743	371/428	926/1264
8.02	818/1135	719/734	368/423	919/1248
9.02	787/1124	710/725	366/418	869/1232
10.02	772/1113	743/718	356/414	865/1218

11.02	773/1105	751/711	368/410	851/1206
12.02	808/1099	741/706	374/407	862/1196
13.02	776/1095	739/703	351/406	817/1190
14.02	806/1095	691/702	362/405	825/1187
15.02	857/1099	768/704	377/406	853/1189
People in a serious condition				
23.10	163/81	50/47	46/24	59/57
24.10	160/86	54/49	46/26	56/59
25.10	172/91	56/52	46/27	75/63
26.10	188/96	68/55	51/29	74/66
27.10	207/102	67/59	52/30	77/70
28.10	215/108	72/62	50/32	68/73
29.10	209/114	70/66	50/34	80/77
30.10	208/120	81/73	53/36	94/82
31.10	212/127	90/77	55/38	94/86
1.11	217/133	95/80	55/40	102/90
...
6.02	97/152	76/84	42/40	83/75
7.02	98/151	78/83	38/40	76/74
8.02	102/150	82/83	34/40	62/74
9.02	102/149	73/82	33/40	70/74
10.02	100/148	77/82	29/40	69/74
11.02	101/147	76/81	31/39	63/74
12.02	102/146	77/81	37/39	62/74
13.02	101/144	71/80	35/39	59/74
14.02	99/143	81/79	40/39	63/73
15.02	99/142	80/79	40/38	71/73

* Observations are denoted by "obs" and predictions by "pred".

Source: own elaboration using software R and data collected by M. Rogalski (2022).

5. Conclusion

The COVID-19 pandemic has attracted the attention of many analysts around the world. This article aimed to analyse the number of hospitalised people, patients in a serious condition, deaths, convalescents and positive COVID-19 tests using methods of functional data analysis, for which the principal component analysis proved interesting. A multiple function-on-function model was used to predict the number of hospitalised and intensive care patients during the second wave of the COVID-19 pandemic.

The model predicted well the numbers for most voivodeships yet the greatest problem was Małopolskie voivodeship. When fitting the model to combined data from both waves (Hęćka, 2023) and to the single second wave, the number of people hospitalised and patients in a serious condition for this voivodeship turned out to be too low. The model predicted values closest to the real numbers in the case of Wielkopolskie and Świętokrzyskie voivodeships.

The analysis of principal components proved interesting. During the second and third wave of the pandemic (Hęćka, 2023), the highest numbers of people hospitalised were in Lubuskie and Podlaskie voivodeships, and during a single second wave in Lubelskie and Podlaskie. The highest number of

people in a serious condition during both waves was in Kujawsko-Pomorskie and Lubuskie voivodeships, during the single second wave in Lubelskie and Kujawsko-Pomorskie. The performed analysis confirmed the real discrete observations.

The economic crisis caused by the SARS-CoV-2 virus has affected almost the entire planet, starting from when the World Health Organization declared a state of emergency in March 2020. In order to control the virus, the scientific community is busy developing models to mitigate the devastating effects of the pandemic.

References

- Acal, C., Escabias, M., Aguilera, A. M., and Valderrama, M. J. (2021). COVID-19 Data Imputation by Multiple Function-on-Function Principal Component Regression. *Mathematics*, 9(11), 1237. <https://doi.org/10.3390/math9111237>
- Boschi, T., Di Iorio, J., Testa, L., Cremona, M. A., and Chiaromonte, F. (2021). Functional Data Analysis Characterizes the Shapes of the First COVID-19 Epidemic Wave in Italy. *Scientific Reports*, 11(1), 17054. <https://doi.org/10.1038/s41598-021-95866-y>
- de Boor, C. (1978). A Practical Guide to Splines. *Mathematics of Computation*, 27(149). doi:10.2307/2006241
- Cai, X., Xue, L., and Cao, J. (2022). Variable Selection for Multiple Function-on-Function Linear Regression. *Statistica Sinica*, 32, 1435-1465. doi: <https://doi.org/10.5705/ss.202020.0473>
- Cao, L., and Liu, Q. (2022). *COVID-19 Modeling: A Review*. Preprint medRxiv 2022.08.22.22279022.
- Hećka, P. (2023). *Functional Data Analysis: Application to the Second and Third Wave of the COVID-19 Pandemic in Poland*. Preprint arXiv:2306.12390. submitted.
- Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., and Mohi Ud Din, M. (2020). Machine Learning Based Approaches for Detecting COVID-19 Using Clinical Text Data. *International Journal of Information Technology: An Official Journal of Bharati Vidyapeeth's Institute of Computer Applications and Management*, 12(3), 731-739. <https://doi.org/10.1007/s41870-020-00495-9>
- Oshinubi, K., Ibrahim, F., Rachdi, M., and Demongeot, J. (2022). Functional Data Analysis: Application to Daily Observation of COVID-19 Prevalence in France. *AIMS Mathematics*, 7(4), 5347-5385. doi:10.3934/math.2022298
- Ramsay, J. O., Graves, S., and Hooker, G. (2020). *FDA: Functional Data Analysis in R*. R package version 5.1.7. <https://CRAN.R-project.org/package=fda>
- Ramsay, J. O., and Silverman, B. W. (2010). *Functional Data Analysis, Springer Series in Statistics*. New York: Springer. Edition Number 2. <https://doi.org/10.1007/b98888>
- Rogalski, M. (2022). *COVID-19 w Polsce. Dane zebrane na podstawie raportów podawanych przez Ministerstwo Zdrowia, danych z WSSE, PSSE, Urzędów Wojewódzkich, oraz tych uzyskanych w prośbach o dostęp do informacji publicznej*. Retrieved March 1, 2022 from <http://bit.ly/covid19-poland>
- Schuller, B. W., Schuller, D. M., Qian, K., Liu, J., Zheng, H., and Li, X. (2021). COVID-19 and Computer Audition: An Overview on What Speech & Sound Analysis Could Contribute in the SARS-CoV-2 Corona Crisis. *Front. Digit. Health*, 3, 564906. doi: 10.3389/fgth.2021.564906
- Statistical Information Centre. (2021). *Area and Population in the Territorial Profile in 2021, Area, Population Number and Density, as of 1 January 2021*. Retrieved March 7, 2022 from <https://stat.gov.pl/en/topics/population/population/area-and-population-in-the-territorial-profile-in-2021,4,15.html>
- Tandon, H., Ranjan, P., Chakraborty, T., and Suhag, V. (2022). Coronavirus (COVID-19): ARIMA-based Time-series Analysis to Forecast near Future and the Effect of School Reopening in India. *Journal of Health Management*, 24(3), 373-388. doi: 10.1177/09720634221109087
- Ullah, S., and Finch, C. F. (2013). Applications of Functional Data Analysis: A Systematic Review. *BMC Medical Research Methodology*, 13(43). <https://doi.org/10.1186/1471-2288-13-43>
- Wickham, H. et al. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.
- Zhou, T., and Ji, Y. (2020). Semiparametric Bayesian Inference for the Transmission Dynamics of COVID-19 with a State-Space Model. *Contemporary Clinical Trials*, 97, 106146. <https://doi.org/10.1016/j.cct.2020.106146>

Druga fala COVID-19 w Polsce – charakterystyka z zastosowaniem metod FDA

Streszczenie: Głównym celem artykułu była analiza danych funkcjonalnych dotyczących liczby pozytywnych wyników testu, zgonów, ozdowieńców, osób hospitalizowanych oraz w stanie ciężkim podczas drugiej fali pandemii COVID-19 w Polsce. Pierwszym krokiem była konwersja danych w funkcje gładkie. Następnie przedstawiono analizę głównych składowych funkcjonalnych oraz użycie modelu *multiple function-on-function linear regression* w celu predykcji liczby osób hospitalizowanych oraz będących w stanie ciężkim z powodu COVID-19 w polskich województwach. Otrzymane wyniki porównano z wcześniej uzyskanymi dla połączonych danych z drugiej i trzeciej fali pandemii.

Słowa kluczowe: regresja *function-on-function*, analiza danych funkcjonalnych, COVID-19, analiza głównych składowych funkcjonalnych, funkcje gładkie.
