Uniwersytet Wrocławski
Wydział Matematyki i Informatyki
Instytut Matematyczny
specjalność: analiza danych

Patrycja Hęćka

Predykcja liczby osób hospitalizowanych oraz w stanie ciężkim dla drugiej i trzeciej fali pandemii COVID-19

Praca magisterska
napisana pod kierunkiem
dr. hab. Krzysztofa Topolskiego

Wrocław 2022

Spis treści

1 Wprowadzenie

Wirus SARS-CoV-2 stał się globalnym problemem od momentu, gdy ujawnił się pod koniec 2019 roku w Chinach. Jego szybkie rozprzestrzenianie się postawiło w stan gotowości wiele dziedzin nauki, nie tylko medycznych. Już 4 marca 2020 r. wykryto pierwszy przypadek koronawirusa w Polsce. Na ten moment w Polsce odnotowano około 6 milionów przypadków COVID-19.

W celu zwalczenia tej sytuacji istnieje wielka potrzeba zrozumienia sposobu rozwijania się pandemii. Znajomość zachowania choroby pozwoliłaby na ograniczenie jej rozprzestrzeniania się. W tym celu społeczność naukowa koncentruje wszystkie swoje wysiłki na opracowywaniu nowych technik modelowania i przewidywania rozwoju COVID-19. Główne, mierzone w krajach zmienne to liczba pozytywnych wyników testu, osób zmarłych oraz wyzdrowień. Innymi ważnymi zmiennymi są liczba osób hospitalizowanych oraz osób w stanie ciężkim. Znając z wyprzedzeniem takie wartości, szpitale mogłyby odpowiednio przygotować się na przyjęcie chorych.

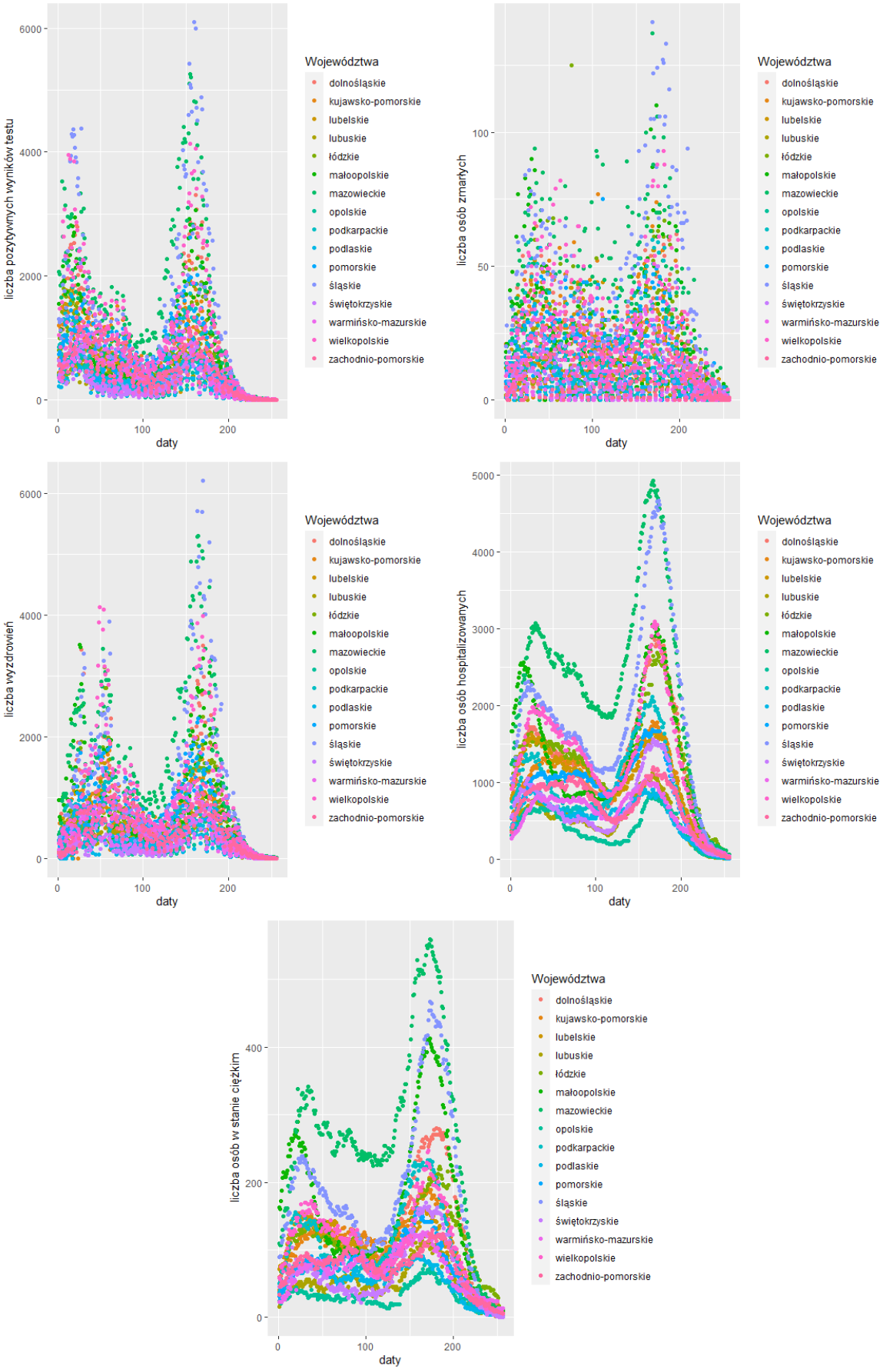
Biorąc pod uwagę naturę zmiennych, w tej pracy proponujemy użycie metod analizy danych funkcjonalnych - *Functional Data Analysis* (FDA) w celu dokonania predykcji liczby osób hospitalizowanych oraz w stanie ciężkim. FDA jest nowoczesną gałęzią statystyki, która ma na celu analizę informacji pochodzących z krzywych bądź funkcji. Jest często stosowaną metodą w różnych dziedzinach, takich jak biotechnologia, ekonomia, elektronika, a także chemia. Analiza Głównych Składowych Funkcjonalnych (FPCA) została wykorzystana między innymi w Stanach Zjednoczonych oraz w Hiszpanii w celu wyjaśnienia zmienności COVID-19.

W tej pracy zastosujemy model funkcjonalnej regresji liniowej w celu predykcji krzywych przedstawiających liczbę osób hospitalizowanych oraz osób w stanie ciężkim w wybranych, polskich województwach.

1.1 Dane

Ustalenie granic poszczególnych fal koronawirusa w Polsce jest umowne. Pierwszy przypadek zakażenia stwierdzono 4 marca 2020 roku w szpitalu w Zielonej Górze u 66-letniego mężczyzny. Za początek pierwszej fali w Polsce przyjmuje się zatem wiosnę 2020 roku. Druga fala COVID-19 w Polsce trwała pięć miesięcy - od września do stycznia. Szczyt drugiej fali przypadł na listopad z rekordowym przyrostem - 27 875 zakażeń - 7 listopada 2020 r. Trzecia fala rozpoczęła się trzy miesiące po szczycie drugiej fali. Za początek trzeciej fali uznaje się 16.02.2021 r. Jej rekord przypadł na 1 kwietnia z 35 251 nowych przypadków SARS-CoV-2 - najwyższą dobową liczbą zakażeń od początku pandemii w Polsce. Jednakże z powodu braków w danych podawanych do informacji publicznej jako drugą falę przyjmiemy okres pomiędzy 23.10.2020 a 15.02.2021 roku, zaś jako trzecią czas od 16.02 do 5.07.2021 roku.

Rysunek (1) przedstawia kolejno dzienne, dyskretne obserwacje liczby pozytywnych wyników testu na COVID-19, liczby zgonów, ozdrowieńców, osób hospitalizowanych oraz osób w stanie ciężkim.



Rysunek 1: Dzienne obserwacje liczby pozytywnych wyników testu, osób zmarłych, wyzdrowień, osób hospitalizowanych i w stanie ciężkim od 23.10.2020 do 05.07.2021 we wszystkich województwach w Polsce.

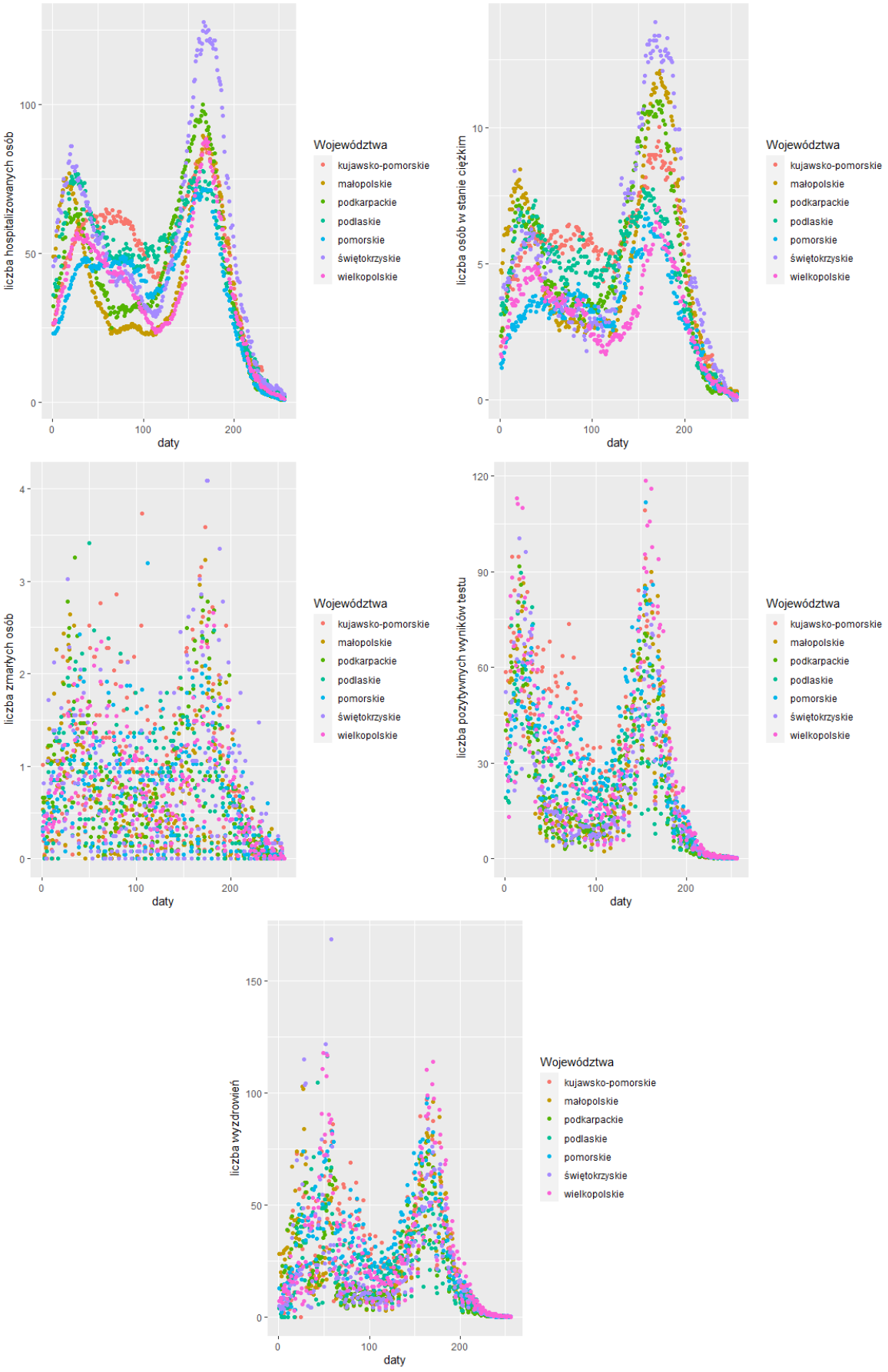
W celu redukcji wpływu liczby mieszkańców danego województwa na prowadzoną analizę podzielono liczbę przypadków, liczbę osób hospitalizowanych, ozdrowieńców, zmarłych oraz osób w stanie ciężkim przez liczbę mieszkańców danego województwa. Liczbę mieszkańców pobrano ze strony Głównego Urzędu Statystycznego (stan na 1 stycznia 2021). Następnie w celu ułatwienia analizy, przemnożono otrzymane liczby przez 100000. Na przykład w województwie wielkopolskim 23.10.2020 odnotowano liczbę osób hospitalizowanych równą 924 przypadkom. Liczba mieszkańców w województwie wielkopolskim 1 stycznia 2021 roku wyniosła 3496451 osób. Dzieląc 924 przypadki przez 3496451, a następnie mnożąc przez 100000, otrzymujemy liczbę 26.4268. W ten sam sposób skalujemy dane dla pozostałych dni i województw. Analizę w kolejnych rozdziałach przeprowadzimy zatem dla odpowiednio przeskalowanych wartości. W ten sposób liczba mieszkańców danego województwa przestanie mieć wpływ na analizę. Od teraz liczba osób np. w stanie ciężkim w danym województwie oznaczać będzie - liczbę osób w stanie ciężkim na 100000 mieszkańców tego województwa.

Z analizy odpowiednio przeskalowanych wartości okazuje się, że najwięcej osób hospitalizowanych oraz osób w stanie ciężkim przypadających na 100000 mieszkańców w czasie 23.10.2020-05.07.2021 odnotowano w województwie świętokrzyskim. Najwięcej pozytywnych testów oraz zgonów przypisano województwu kujawsko-pomorskiemu. Największa liczba ozdrowieńców została osiągnięta w województwie warmińsko-mazurskim. Najmniej osób hospitalizowanych było w województwie wielkopolskim, osób w stanie ciężkim w pomorskim, zgonów w małopolskim, pozytywnych testów w podkarpackim, a ozdrowieńców w województwie podlaskim. Na tej podstawie do próby testowej modelu analizowanego w kolejnych rozdziałach wybrano województwa: świętokrzyskie, wielkopolskie, podkarpackie oraz małopolskie.

Od tego momentu zmienność w czasie pozytywnych wyników testu, zgonów, ozdrowieńców, osób hospitalizowanych oraz osób w stanie ciężkim będzie rozważana poprzez zmienne funkcjonalne odpowiednio: $X_1(t), X_2(t), X_3(t), Y_1(t), Y_2(t)$. Zmienne X będą traktowane jak predyktory, a Y jako zmienne odpowiedzi w funkcjonalnych modelach regresji. Obserwowane dane to liczba dziennych, skumulowanych wartości tych pięciu zmiennych funkcjonalnych dla szesnastu województw w Polsce.

Rysunek (2) przedstawia stopień rozwoju choroby w wybranych, najbardziej zróżnicowanych województwach poprzez wykresy dyskretnych, dziennych, przeskalowanych obserwacji, czyli zestaw krzywych $\{(x_{ij}(t), y_{ik}(t)) : i = 1, ..., 16; j = 1, 2, 3; k = 1, 2\}$.

Ważnym ograniczeniem na metody FDA jest to, że krzywe zmiennych funkcjonalnych muszą być obserwowane na tej samej dziedzinie. Klasycznym rozwiązaniem takiego problemu jest przeniesienie wszystkich krzywych na ten sam odcinek. W tej pracy będziemy rozważać krzywe na odcinku T=[0,1]. Stąd od tego momentu x_{ij} i y_{ik} reprezentują krzywe położone w dziedzinie [0,1].



Rysunek 2: Dzienne obserwacje od 23.10.2020 do 05.07.2021 w wybranych województwach. Liczba osób została podzielona przez liczbę mieszkańców danego województwa, a następnie przemnożona przez 100000.

2 Z danych dyskretnych w dane funkcjonalne

Podstawową filozofią w analizie danych funkcjonalnych jest myślenie o funkcjach z obserwowanych danych jako pojedynczych bytach, a nie jak o sekwencji indywidualnych obserwacji. Termin „funkcjonalny" odnosi się do wewnętrznej struktury danych, rzadziej do ich wyraźnej formy. W praktyce dane funkcjonalne są zazwyczaj obserwowanymi, dyskretnymi parami (t_k, y_k), gdzie y_k jest pewną wartością funkcji w czasie t_k często zniekształconą poprzez błąd pomiaru.

W jaki sposób możemy przekształcić obserwację w funkcjonalną formę x? W praktyce często nie posiadamy wartości x dla każdego punktu czasowego t. Pomimo tego chcielibyśmy, żeby nasza funkcja x była gładka, co miałoby powodować, że pary kolejnych wartości y_k i y_{k+1} byłyby połączone ze sobą w pewien sposób i nie różniły się znacząco od siebie.

Poprzez gładkość rozumiemy, że funkcja x posiada jedną lub więcej pochodnych, które oznaczamy kolejno poprzez Dx i D^2x. Konsekwentnie poprzez D^nx oznaczymy pochodną n-tego stopnia, a $D^nx(t)$ będzie wartością n-tej pochodnej dla argumentu t.

Naszym celem jest użycie dyskretnych danych y_k, $k = 1, ..., m$ w celu oszacowania funkcji x. Należy pamiętać, że często obserwowane dane okazują się nie być gładkie ze względu na błąd pomiaru.

2.1 Rozszerzanie bazy

Istnieją różne podejścia konwersji danych. W pracy zajmiemy się klasycznym przypadkiem. Do przedstawionych w pierwszym rozdziale obserwacji dyskretnych dopasujemy wykresy, które posłużą przybliżeniu ciągłego procesu. Gdy następnie zajmiemy się analizą obiektów funkcjonalnych, obserwacje dyskretne staną się mniej istotne.

Główną procedurą w statystyce, matematyce, a także inżynierii do konwersji danych dyskretnych w funkcje gładkie jest rozszerzanie bazy.

Rozważmy wektory $x_i(t) : i = 1, ..., n; t \in T = [0, 1]$ oraz załóżmy, że obserwacje y_{ik} są dostępne dla każdego węzła $t_{i1}, t_{i2}, ..., t_{im_i} \in T$. Wówczas każdą obserwację y_{ik} możemy zapisać następująco:

$$y_{ik} = x_i(t_{ik}) + \epsilon_{ik}, i = 1, ..., n; k = 1, ..., m_i, \tag{1}$$

gdzie szumy ϵ_{ik} przyczyniają się do szorstkości analizowanych danych. Jednym z celów reprezentacji surowych danych w postaci funkcji jest odfiltrowanie szumu.

Załóżmy, że próbkowe krzywe należą do skończenie-wymiarowej przestrzeni generowanej przez zestaw funkcji bazowych $\{\phi_1(t), ..., \phi_p(t)\}$. Wtedy możemy przedstawić każdą krzywą x_i poprzez liniową rozbudowę postaci:

$$x_i(t) = \sum_{j=1}^{p} \alpha_{ij}\phi_j(t), i = 1, ..., n. \tag{2}$$

Definiujemy poprzez α_i wektory długości p, współczynników bazowych: $\alpha_i = (\alpha_{i1}, ..., \alpha_{ip})'$, które mogą być estymowane na różne sposoby. W następnym podrozdziale (2.2) zobaczymy najpopularniejszą metodę estymacji tych współczynników - metodę najmniejszych kwadratów. Okaże się, że estymatory α_i są postaci: $\hat{\alpha}_i = (\Phi_i'\Phi_i)^{-1}\Phi_i' y_i$, gdzie $\Phi_i = (\phi_j(t_{ik}))_{m_i \times p}, j = 1, ..., p; k = 1, ..., m_i$.

Zestaw funkcji bazowych jest zbiorem p znanych funkcji ϕ_j, które są matematycznie niezależne od siebie. Można je przybliżać dowolnymi, dobrymi funkcjami poprzez wzięcie ważonej sumy lub liniowej kombinacji wystarczająco dużej liczby p tych funkcji. W podrozdziale (2.3) zostanie szerzej omówiony temat funkcji bazowych.

2.2 Metoda najmniejszych kwadratów

W tym podrozdziale spróbujemy znaleźć estymator najmniejszych kwadratów dla wektorów współczynników bazowych α_i. Naszym zadaniem jest minimalizacja kryterium najmniejszych kwadratów:

$$\text{SMSSE}(y_i|\alpha_i) = \sum_{k=1}^{m_i}[y_{ik} - \sum_{j}^{p} \alpha_{ij}\phi_j(t_{ik})]^2. \tag{3}$$

W formie macierzowej możemy przedstawić je w postaci:

$$\text{SMSSE}(y_i|\alpha_i) = (y_i - \Phi_i\alpha_i)'(y_i - \Phi_i\alpha_i).$$

Biorąc pierwszą pochodną kryterium SMSSE względem współczynnika α_i, otrzymujemy:

$$\frac{\partial \text{SMSSE}}{\partial \alpha_i} = 2\Phi_i\Phi_i'\alpha_i - 2\Phi_i' y_i.$$

Stąd po przyrównaniu pochodnej do 0, estymator parametru α_i jest postaci:

$$\hat{\alpha}_i = (\Phi_i'\Phi_i)^{-1}\Phi_i' y_i. \tag{4}$$

Zaś wektor $\hat{y}_i$ dopasowanych wartości jest postaci:

$$\hat{y}_i = \Phi_i\hat{\alpha}_i = \Phi_i(\Phi_i'\Phi_i)^{-1}\Phi_i' y_i. \tag{5}$$

2.3 Typ bazy

Typ bazy jest wybierany stosownie do zbioru danych, który jest poddawany analizie. Najczęściej stosuje się *B-splines* (de Boor, 2001; Shikin Plis, 1995; Unser, 1999), funkcje trygonometryczne, *M-splines*, a także splajny naturalne. Pierwsza z metod polega na wygenerowaniu przestrzeni sklejanych ze sobą (ang. spline functions) kawałkami wielomianowych funkcji, które są następnie płynnie łączone. *B-splines* są to więc zasadniczo wielomiany połączone końcami w zbiorze granic przedziałów, nazywanych węzłami - *knots*. Taka struktura pozwala na dostosowanie do pewnego stopnia gładkości funkcji do analizowanych danych. Wygładzenie jest kontrolowane przez liczbę funkcji bazowych. Węzły są często równomiernie rozmieszczane.

Współczynniki bazowe są wybierane w taki sposób, aby skonstruowana krzywa optymalnie dopasowywała się do danych. Powszechnym podejściem wygładzania *B-splines* jest wybór dużej liczby węzłów w celu zmniejszenia liczby stopni swobody i zwiększenia gładkości estymowanej funkcji.

B-splines są charakteryzowane przez ich rząd, który jest zwykle o jeden większy od stopnia wielomianu, z którego funkcja jest skonstruowana. Na przykład *B-splines* rzędu trzy składają się z funkcji kwadratowych połączonych w węzłach, a *B-splines* drugiego rzędu są konstruowane z fragmentów linii. Funkcje *B-splines* są normalizowane tak, aby dla każdej dopasowanej wartości w momencie t, suma wszystkich funkcji bazowych w tym punkcie wynosiła 1. Ta normalizacja sprawia, że każdy współczynnik bazowy jest w przybliżeniu wartością dopasowanej krzywej w miejscu, gdzie j-ta funkcja *B-spline* osiąga maksimum.

Każda funkcja bazowa $\phi_j(t)$ jest funkcją sklejaną, definiowaną przez rząd - *„order"* i sekwencję węzłów τ. Co ciekawe każda liniowa kombinacja funkcji bazowych jest funkcją sklejaną np. suma lub różnica funkcji sklejanych jest także funkcją sklejaną.

2.4 Kompromis pomiędzy wariancją a obciążeniem

Jak już wspomniano gładkość funkcji jest kontrolowana przez liczbę użytych funkcji bazowych. Im większa liczba funkcji bazowych p, tym zbudowana krzywa lepiej dopasowuje się do dyskretnych punktów.

Im mniejsza wartość p, tym krzywa jest bardziej wygładzona. Zachodzi to jednak kosztem zmniejszenia zdolności wychwycenia pewnych, ostrzejszych cech, sygnałów w danych. Decyzja o zwiększeniu bądź zmniejszeniu liczby funkcji bazowych jest związana z osiągnięciem kompromisu pomiędzy obciążeniem a wariancją - *"bias-variance trade-off"*. Im większa liczba funkcji bazowych, tym mamy niższe obciążenie, ale mniejsze wygładzenie i wyższą wariancję. Z drugiej strony użycie małej liczby funkcji bazowych powoduje pojawienie się mniejszej wariancji, ale kosztem wzrostu obciążenia.

Przedstawimy jedno z najczęściej stosowanych rozwiązań powyższego problemu.

Obciążenie estymatora możemy zapisać następująco:

$$\text{Bias}[\hat{x}(t)] = x(t) - \text{E}[\hat{x}(t)].$$

Dla dużego p obciążenie jest małe. W sytuacji, gdy liczba funkcji bazowych jest równa liczbie obserwacji obciążenie będzie równe 0.

Jednakże powodem, dla którego nie powinno się ustalać liczby funkcji bazowych równej liczbie obserwacji, jest wariancja. Konsekwentnie jesteśmy także zainteresowani wariancją estymatora:

$$\text{Var}[\hat{x}(t)] = \text{E}[(\hat{x}(t) - \text{E}[\hat{x}(t)])^2].$$

Jeśli ustalimy, że liczba funkcji bazowych jest równa liczbie obserwacji, wariancja będzie duża. Redukcja wariancji skłania nas do wyboru małych wartości parametru p.

Jednym ze sposobów uzyskania „kompromisu" pomiędzy wariancją a obciążeniem jest minimalizacja błędu średniokwadratowego - *mean-squared error* (MSE):

$$\mathrm{MSE}[\hat{x}(t)] = \mathrm{E}[(\hat{x}(t) - x(t))^2].$$

Błąd średniokwadratowy jest równoważny sumie wariancji i kwadratu obciążenia:

$$\mathrm{MSE}[\hat{x}(t)] = \mathrm{Bias}^2[\hat{x}(t)] + \mathrm{Var}[\hat{x}(t)].$$

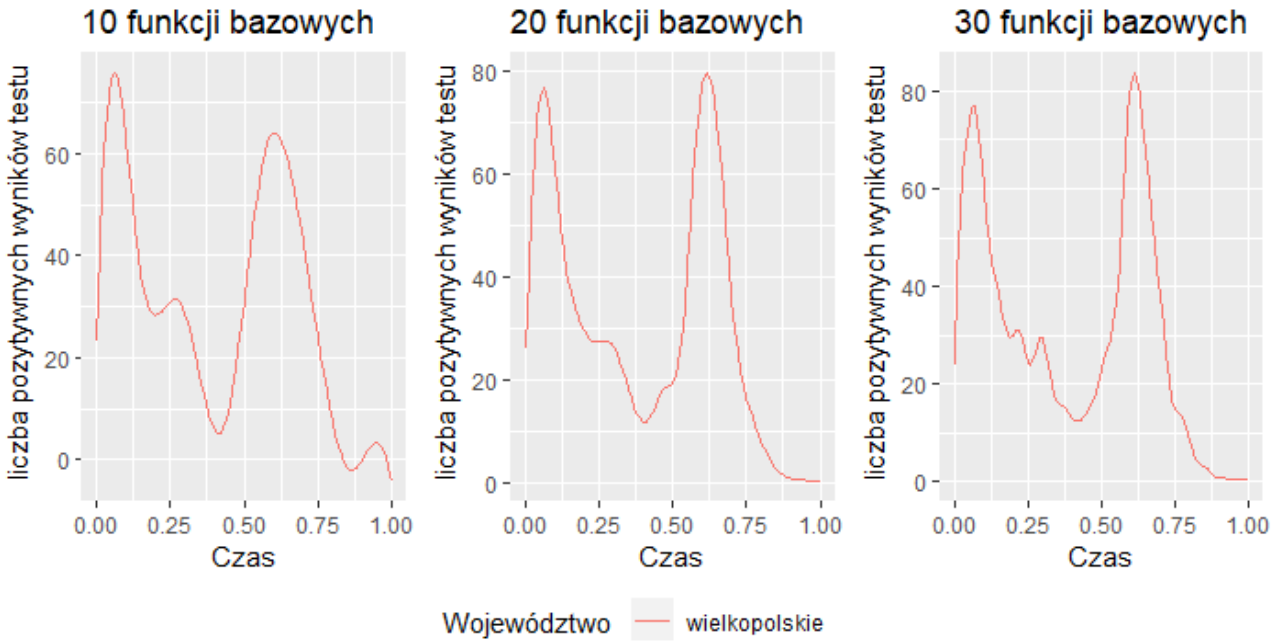
Z powyższego wzoru można uzyskać pewną intuicję odnośnie opisywanego „kompromisu". Naszym celem jest otrzymanie odpowiedniej wartości obciążenia oraz wariancji, co możemy osiągnąć, kontrolując wartość błędu średniokwadratowego.

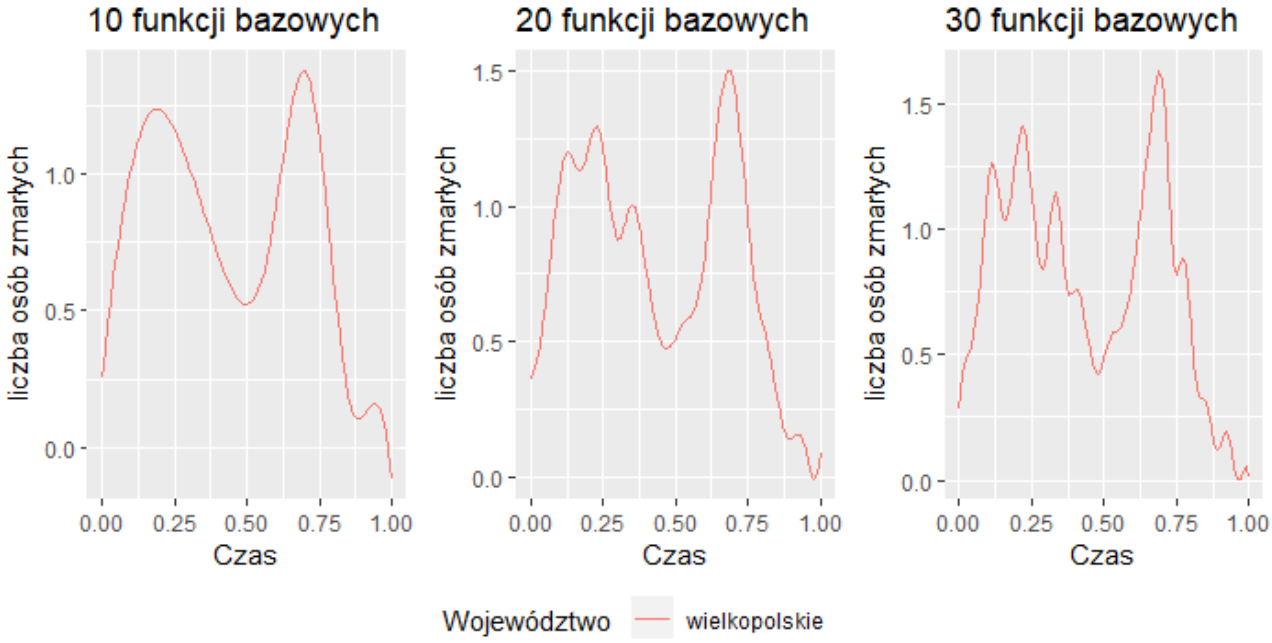
2.5 Przykład

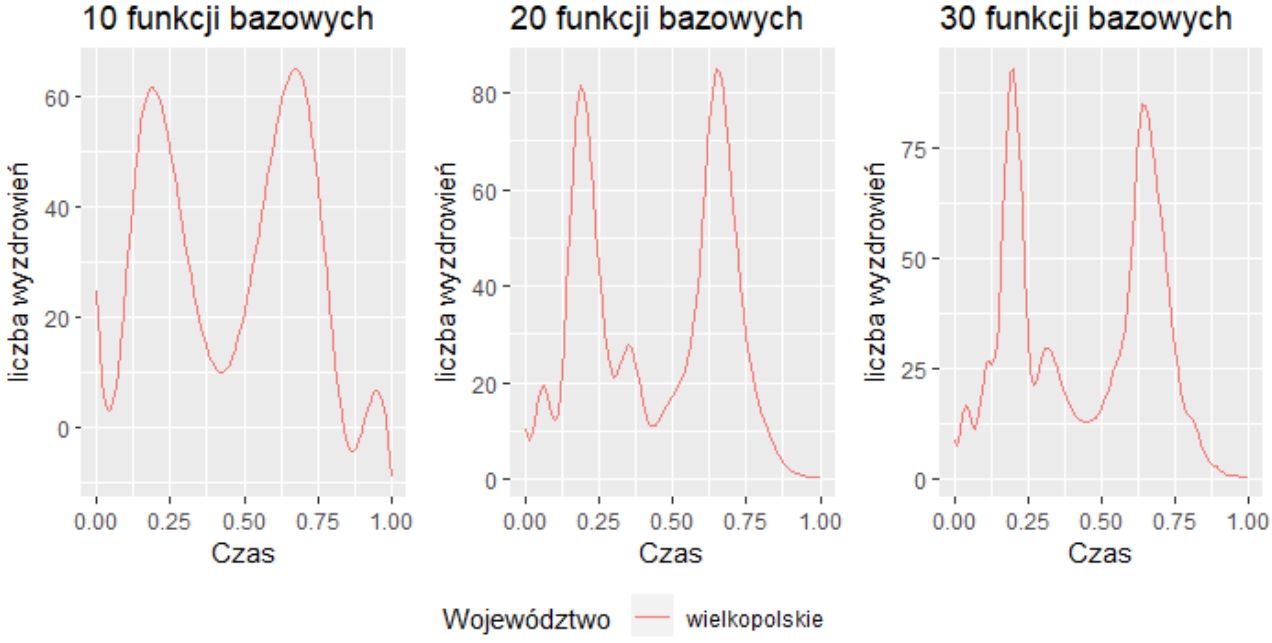
Na rysunku (3) przedstawiono dopasowane krzywe w zależności od liczby funkcji bazowych (10, 20 lub 30) dla województwa wielkopolskiego. Widzimy, że krzywe z liczbą funkcji bazowych równą 10 są najbardziej wygładzone. Jednak otrzymane tutaj obciążenie estymatora jest największe. Krzywe z liczbą funkcji bazowych równą 30, pozwalają wychwycić ostrzejsze sygnały, których nie dostrzegamy na wykresach krzywych z 10 funkcjami bazowymi. Minusem zbyt dużej liczby funkcji bazowych jest jednak wzrost wariancji.

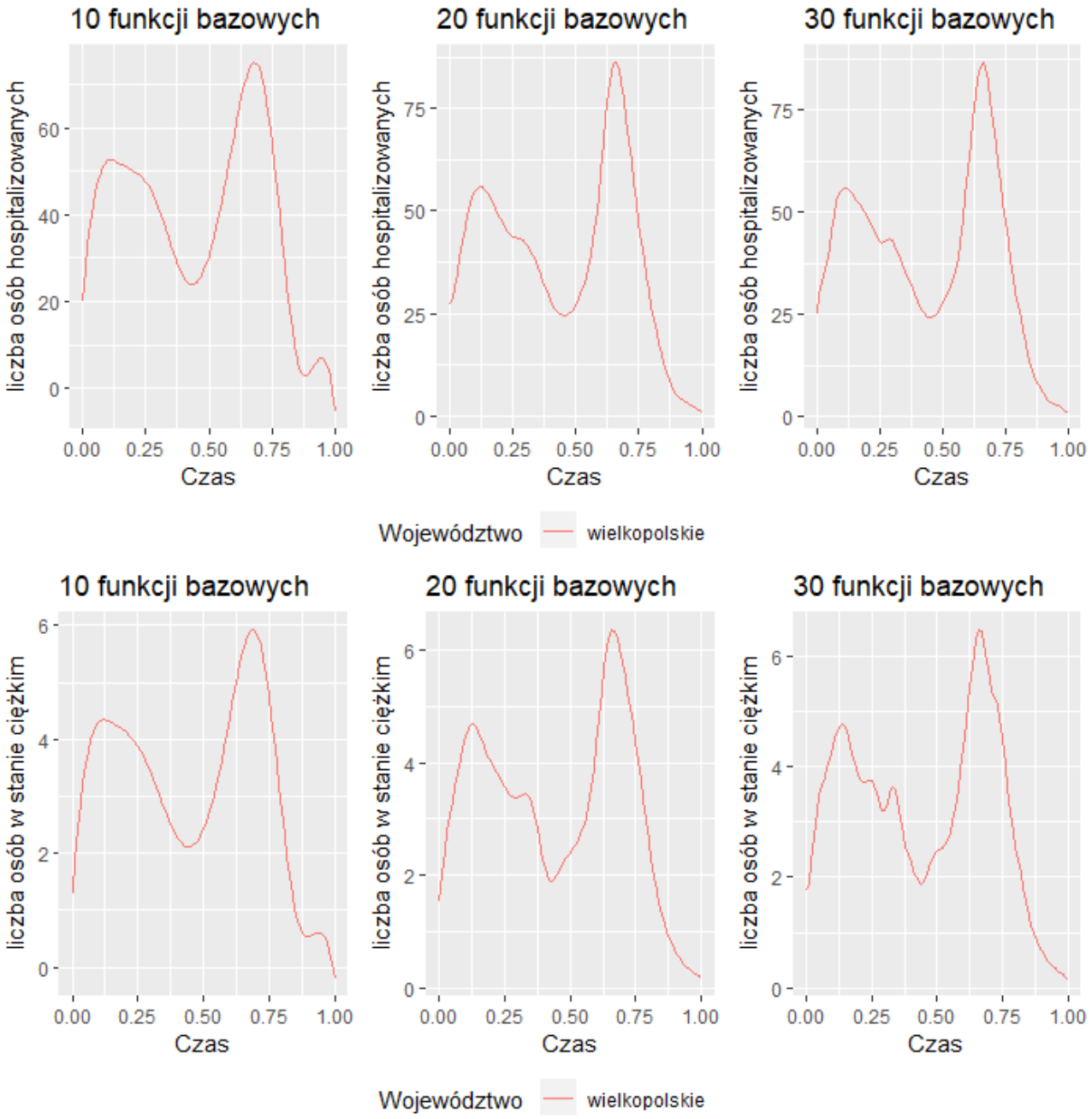
Możemy zauważyć, że dla niektórych zmiennych funkcjonalnych zwiększenie liczby funkcji bazowych ma większy wpływ na wygląd wykresu niż dla innych. Na przykład możemy dostrzec, że dla liczby osób zmarłych wykres dla dziesięciu funkcji bazowych jest znacznie bardziej wygładzony od wykresu skonstruowanego przy użyciu trzydziestu funkcji bazowych. Tymczasem wykresy dla liczby osób hospitalizowanych nie różnią się znacząco od siebie. Aby osiągnąć „kompromis" pomiędzy wariancją a obciążeniem wybrano 20 funkcji bazowych.

Rysunek (4) przedstawia dopasowanie 20 funkcji bazowych z równo rozmieszczonymi węzłami w celu przybliżenia siedmiu krzywych przedstawiających liczbę zachorowań, zgonów, wyzdrowień, osób hospitalizowanych oraz osób w stanie ciężkim z powodu COVID-19. Współczynniki dla każdej formy funkcjonalnej zostały dopasowane metodą najmniejszych kwadratów. Liczba funkcji bazowych została wybrana tak, aby wartość błędu średniokwadratowego była jak najmniejsza, ale także, aby nie doszło do przeuczenia modelu.

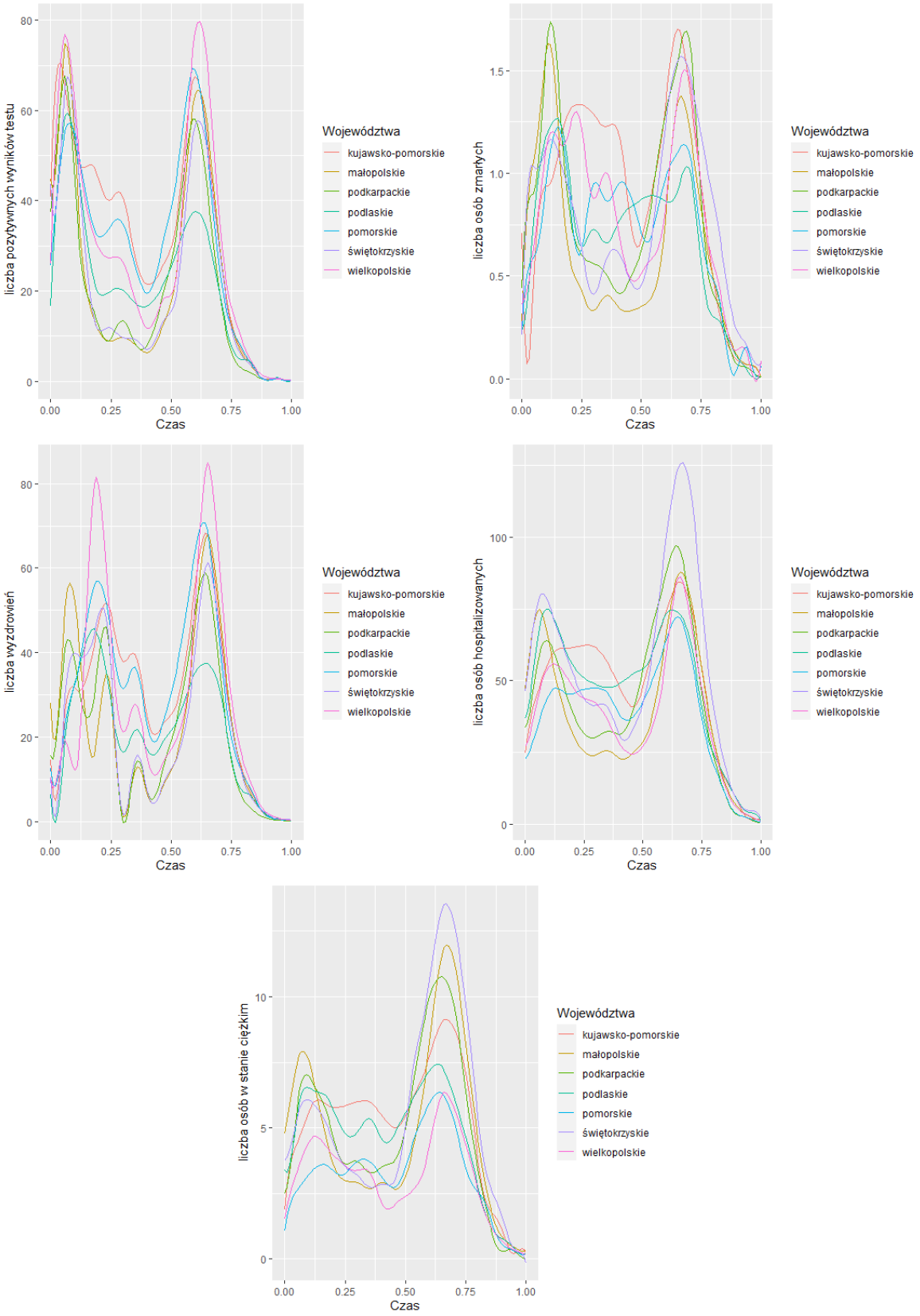

10 funkcji bazowych
20 funkcji bazowych
30 funkcji bazowych
liczba pozytywnych wyników testu
Czas
Województwo
wielkopolskie


10 funkcji bazowych
20 funkcji bazowych
30 funkcji bazowych
liczba osób zmarłych
Czas
Województwo
wielkopolskie


10 funkcji bazowych
20 funkcji bazowych
30 funkcji bazowych
liczba wyzdrowień
Czas
Województwo
wielkopolskie



Rysunek 3: Dopasowane krzywe do dziennych obserwacji liczby pozytywnych wyników testu, liczby zgonów, wyzdrowień, osób hospitalizowanych oraz w stanie ciężkim w województwie wielkopolskim wraz z liczbą funkcji bazowych.



Rysunek 4: Dopasowane krzywe do dziennych obserwacji liczby pozytywnych wyników testu, liczby zgonów, wyzdrowień, osób hospitalizowanych oraz w stanie ciężkim w wybranych województwach. Zastosowano 20 funkcji bazowych.

3 Multiple Function-on-Function Linear Model

W tym rozdziale zostanie przedstawiony *multiple function-on-function linear regression model* - MFFLR, a także jego estymacja w terminach regresji głównych składowych funkcjonalnych.

Model MFFLR pozwala na oszacowanie funkcjonalnej zmiennej odpowiedzi Y na podstawie wektora J funkcjonalnych zmiennych objaśniających oznaczanych przez $X = (X_1, ..., X_J)'$. Rozważmy losową próbę z (X,Y) oznaczaną przez $(x_i, y_i) : i = 1, ..., n$ z $x_i = (x_{i1}, x_{i2}, ..., x_{iJ})'$ i załóżmy, że wszystkie zmienne funkcjonalne przyjmują wartości w ośrodkowej przestrzeni Hilberta $L^2([0, 1])$. Jest to podstawowa przestrzeń, w której rozważa się dane funkcjonalne.

Wtedy funkcjonalny liniowy model definiujemy następująco:

$$y_i(t) = \alpha(t) + \sum_{j=1}^{J} \int_T x_{ij}(s)\beta_j(s,t)ds + \epsilon_i(t), i = 1, ..., n, \tag{6}$$

gdzie $\alpha(t)$ to funkcja incerceptu, $\beta_j(s,t)$, to funkcje współczynników, a $\epsilon_i(t)$ są niezależnymi błędami funkcjonalnymi. Poniżej przedstawiono model (6) w postaci macierzowej:

$$y_i(t) = \alpha(t) + \int_T x_i(s)'\beta(s,t)ds + \epsilon_i(t), i = 1, ..., n,$$

gdzie $x_i(s) = (x_{i1}(s), x_{i2}(s), ..., x_{iJ}(s))'$ i $\beta(s,t) = (\beta_1(s,t), \beta_2(s,t), ..., \beta_J(s,t))'$.

Analizując równanie (6), możemy zaobserwować, że wszystkie zmienne funkcjonalne są zdefiniowane na tym samym przedziale T. Nie jest to jednak wymóg i model może być uogólniany na różne, inne dziedziny zmiennych funkcjonalnych. Model przedstawiony za pomocą równania (6) jest zwykle estymowany poprzez metodę najmniejszych kwadratów z pewną karą oraz późniejsze rozszerzanie bazy parametrów funkcjonalnych. Niektóre z metod rozszerzania bazy pozwalają na zredukowanie modelu do wielowymiarowego modelu liniowego dla macierzy bazowych współczynników odpowiedzi w terminach macierzy współczynników bazowych zmiennych objaśniających. Głównym pojawiającym się problemem jest duża współliniowość modelu wielowymiarowego, która powoduje niedokładne oszacowywanie parametrów. Pomimo dobrej zdolności predykcyjnej modelu, fakt ten sprawia, że jego interpretacja jest utrudniona. W najczęściej pojawiających się rozwiązaniach unika się wykonywania kroswalidacji w celu estymacji parametru kary. Stosuje się podejścia oparte na analizie głównych składowych funkcjonalnych - PCA i funkcjonalnej metodzie częściowych najmniejszych kwadratów (ang. *Partial Least Squares* - PLS).

W tej pracy zastosujemy regresję głównych składowych. Rozszerzymy predykcyjne modele głównych składowych w celu oszacowania funkcjonalnych zmiennych odpowiedzi.

3.1 Regresja głównych składowych funkcjonalnych

Analiza głównych składowych danych funkcjonalnych jest ważną metodą, która pozwala na poznanie cech charakteryzujących typowe funkcje. Niektóre z nich nie są zaskoczeniem np. sinusoidalna natura krzywych temperatur, ale inne aspekty mogą być interesujące dla analityka.

W klasycznym przypadku wielowymiarowym funkcje wariancji-kowariancji oraz korelacji mogą być trudne do zinterpretowania. Analiza głównych składowych pozwala spojrzeć na strukturę kowariancji w szerszy sposób i przynieść więcej istotnych informacji na temat analizowanych danych.

Omawiana metoda jest również kluczowym narzędziem redukcji wymiaru wielowymiarowych danych, które zostało rozszerzone do danych funkcjonalnych i określone terminem *analizy głównych składowych funkcjonalnych* - FPCA. Dzieje się tak ponieważ FPCA ułatwia konwersję nieskończenie wymiarowych danych funkcjonalnych do skończenie wymiarowego wektora wyników. Proces stochastyczny może być wyrażony jako przeliczalny ciąg nieskorelowanych zmiennych losowych, głównych składowych funkcjonalnych bądź wyników, które w wielu praktycznych zastosowaniach są obcinane do skończonego wektora. Następnie można w prosty sposób zastosować narzędzia wielowymiarowej analizy danych do wektora wyników, redukując w ten sposób wymiar.

Rozważmy zatem dekompozycję głównych składowych zarówno funkcjonalnej zmiennej odpowiedzi, jak i funkcjonalnych zmiennych objaśniających, danych przez

$$x_{ij}(t) = \overline{x}_j(t) + \sum_{l=1}^{n-1} \xi_{il}^{x_j} f_l^{x_j}(t), \tag{7}$$

$$y_i(t) = \overline{y}(t) + \sum_{l=1}^{n-1} \xi_{il}^{y} f_l^{y}(t), \tag{8}$$

gdzie

$$\xi_{il}^{x_j} =< x_{ij} - \overline{x}_j, f_l^{x_j} >= \int_T (x_{ij}(t) - \overline{x}_j(t)) f_l^{x_j}(t) dt, \tag{9}$$

$$\xi_{il}^{y} =< y_i - \overline{y}, f_l^{y} >= \int_T (y_i(t) - \overline{y}(t)) f_l^{y}(t) dt \tag{10}$$

są głównymi funkcjonalnymi składowymi, często nazywanymi *scores*. Funkcje wag, będące funkcjami własnymi - *eigenfunctions* próbkowej kowariancji $x_{ij}(t)$ i $y_i(t)$, oznaczamy przez $f_l^{x_j}$ i f_l^{y}. *Scores* głównych składowych są wyśrodkowanymi, nieskorelowanymi zmiennymi skalarnymi z maksymalną wariancją daną przez wartości własne i ich funkcje wagowe:

$$Var(\xi_{il}^{x_j}) = \lambda_l^{x_j},$$

$$Var(\xi_{il}^{y}) = \lambda_l^{y}$$

oraz wartością oczekiwaną równą 0.

Dekompozycja głównych składowych dana przez wzór (7) pozwala na przekształcenie modelu MFFLR, opisanego wzorem (6) w liniowy model regresji dla każdego składnika głównego zmiennej odpowiedzi Y ze składników głównych funkcjonalnych predyktorów, dany przez wzór (11).

$$\hat{\xi}_{ik}^{y} = \sum_{j=1}^{J} \sum_{l=1}^{n-1} b_{kl}^{x_j} \xi_{il}^{x_j} + \epsilon_{ik}, i = 1, ..., n; k = 1, ..., n-1. \tag{11}$$

Funkcjonalne współczynniki dane są tutaj przez $\beta_j(s,t) = \sum_{k=1}^{n-1}\sum_{l=1}^{n-1} b_{kl}^{x_j} f_k^{x_j}(s) f_l^y(t)$.

Obcinając każdy główny składnik dekompozycji, otrzymujemy następujący model PC-MFFLR dla odpowiedzi funkcjonalnej:

$$\hat{y}_i(s) = \overline{y}(s) + \sum_{k=1}^{K} \hat{\xi}_{ik}^y f_k^y(s) = \overline{y}(s) + \sum_{k=1}^{K}(\sum_{j=1}^{J}\sum_{l\in L_{kj}} \hat{b}_{kl}^{x_j}\xi_{il}^{x_j}) f_k^y(s), \tag{12}$$

gdzie K oznacza liczbę wybranych do modelu głównych składowych, zaś $\hat{b}_{kl}^{x_j}$ są liniowymi estymatorami najmniejszych kwadratów współczynników regresji b_{kl}.

W celu oszacowania głównych składowych dla zmiennej odpowiedzi, opracowano różne podejścia wyboru modelu i selekcji optymalnych głównych składowych dla każdej zmiennej objaśniającej (podzbiory L_{kj}). Główne składowe są uporządkowywane zgodnie z wyjaśnianą przez nie zmiennością. Co zaskakujące, składowe, które najwięcej objaśniają czasami nie są najbardziej skorelowane ze zmienną odpowiedzi. Wersja FPCA, która estymuje główne składowe, biorąc pod uwagę korelację funkcjonalnego predyktora i zmiennej odpowiedzi została opracowana w modelu regresji *scalar-on-function*. Najpopularniejsza procedura wyboru modelu oparta na metodzie krokowej - *stepwise* i wyborze najlepszego podzbioru - *best subset regression*, połączona z kroswalidacją jest często stosowana w kontekście funkcjonalnej regresji.

3.2 Imputacja brakujących krzywych odpowiedzi

Rozważmy sytuację, w której wszystkie zmienne objaśniające X_j są zaobserwowane, a tylko zmienna odpowiedzi Y ma brakujące wartości. Załóżmy, że w próbie, posiadamy pierwsze n wartości zmiennej odpowiedzi, a ostatnie m wartości są brakujące. Ogólniej oznacza to, że mamy n kompletnie zaobserwowanych krzywych dla wszystkich zmiennych i m brakujących krzywych dla zmiennej odpowiedzi.

Estymując brakujące krzywe odpowiedzi, parametry b_{kl} w modelu (11) są szacowane za pomocą kompletnych n próbkowych krzywych odpowiedzi i predyktorów. Wtedy brakujące krzywe odpowiedzi $y_i^{miss}(s) : i = n+1, ..., n+m$ są estymowane poprzez obliczenie *scores* głównych składowych predyktorów: $\xi_{il}^{x_j} : i = n+1, ..., n+m, l = 1, ..., n-1$ danych przez wzór (9) i wstawienie ich w równanie (12). Wtedy wyestymowany PC-MFFLR model może być użyty do predykcji wartości nowej zmiennej odpowiedzi Y na próbie testowej.

Model regresji, przedstawiony za pomocą równania (12) może być również stosowany w celu predykcji zmiennej odpowiedzi Y(s) w przyszłym przedziale czasowym długości u, oznaczanym przez [T, T+u]. W przypadku danych o COVID-19 parametr u może być wybrany poprzez wzięcie średniej liczby dni potrzebnych, aby u osoby rozwinęły się ciężkie objawy i pojawiła się konieczność przyjęcia do szpitala.

Problem imputacji możemy rozwiązać poprzez użycie modelu *function-on-function* wielokrotnej regresji liniowej dla każdej zmiennej odpowiedzi $Y_1(t)$ (oznaczającej osoby hospitalizowane) oraz $Y_2(t)$ (oznaczającej osoby w ciężkim stanie). Oba funkcjonalne modele regresji są estymowane z pełnych danych 12 województw, które wyznaczają próbę treningową. Następnie predykcje są dokonywane dla czterech pozostałych województw z brakującymi wartościami. W tym celu wybrano województwa: małopolskie, wielkopolskie, świętokrzyskie oraz podkarpackie.

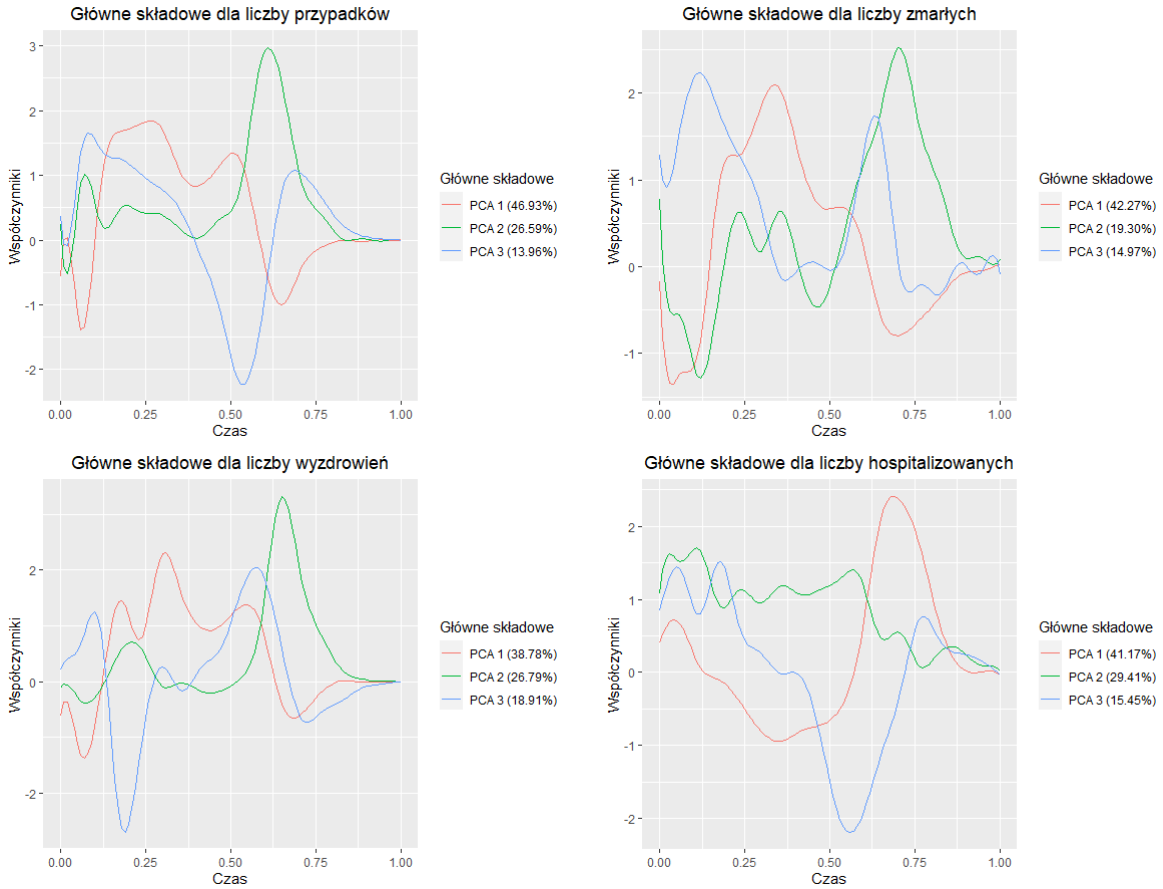
4 Analiza danych dotyczących COVID-19

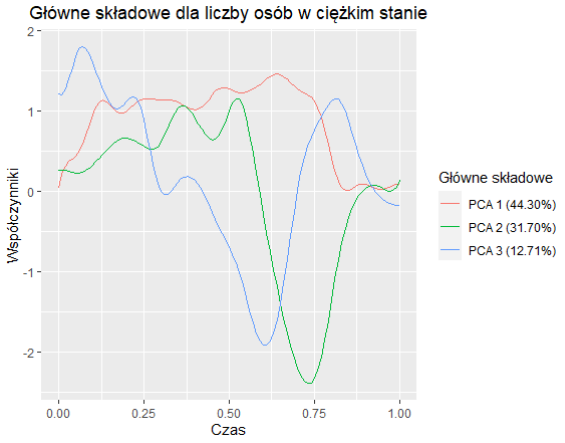
W tym rozdziale zastosujemy analizę głównych składowych funkcjonalnych oraz model MFFLR w celu przeprowadzenia analizy danych funkcjonalnych dotyczących COVID-19 oraz imputacji brakujących krzywych odpowiedzi.

4.1 Analiza Głównych Składowych Funkcjonalnych

Pierwszym krokiem jest estymacja głównych składowych funkcjonalnych dla każdego z pięciu funkcjonalnych predyktorów. Okazuje się, że pierwsze główne składowe wyjaśniają kolejno 46.93%, 42.27%, 38.78%, 41.17%, 44.30% zmienności X_1, X_2, X_3, Y_1, Y_2 z próby treningowej. Drugie główne składowe wyjaśniają kolejno: 26.59%, 19.30%, 26.79%, 29.41%, 31.70% wariancji. Trzecie główne składowe objaśniają znacznie mniej, odpowiednio: 13.96%, 14.97%, 18.91%, 15.45%, 12.71%.

Poniżej przedstawiono funkcje wagowe, harmoniczne związane z 3 pierwszymi, głównymi składowymi. Przedstawione funkcje wagowe są dodatkowo współczynnikami, umożliwiającymi obliczanie wektorów własnych.





Rysunek 5: Funkcje wagowe $f_l^{x_j}; j = 1, 2, 3$ oraz $f_l^{y_k}; k = 1, 2,\ l = 1, 2, 3$ dla trzech pierwszych głównych składowych.

Dla liczby osób hospitalizowanych pierwsza, druga i trzecia główna składowa wyjaśniają razem około 86.03, a dla liczby osób w ciężkim stanie 88.71 procent wariancji. Pozostałe główne składowe wyjaśniają niewielki procent informacji.

Wykresy funkcji wagowych są trudne do bezpośredniej interpretacji. W dalszej części rozdziału zostaną przedstawione rysunki, które pomogą przeprowadzić analizę głównych składowych funkcjonalnych. Jednakże, patrząc na powyższe wykresy, możemy zyskać pewną intuicję na ich temat.

Pierwsza główna składowa pokazuje ogólny wskaźnik zmienności liczby osób hospitalizowanych w zależności od fali pandemii. Możemy zauważyć, że największa różnica pomiędzy drugą i trzecią falą w liczbie osób hospitalizowanych przypada na drugą połowę analizowanego czasu, czyli na trzecią falę pandemii COVID-19. Najmniejsze różnice pomiędzy województwami są podczas drugiej fali pandemii. Może to sugerować, że dokonanie predykcji liczby osób hospitalizowanych podczas trzeciej fali COVID-19 może być najtrudniejsze. W przejściu pomiędzy drugą a trzecią falą dostrzegamy negatywne wartości współczynników, które mówią nam o spadku w liczbie osób hospitalizowanych. Województwa dla których wartość $\xi_{i1}^{y_1}$ jest wysoka będą miały duże różnice pomiędzy liczbą osób hospitalizowanych w drugiej i trzeciej fali. Podczas trzeciej fali liczba osób hospitalizowanych znacznie przekroczy liczbę opisywanych osób podczas drugiej fali pandemii. Okazuje się, że najwyższa wartość tego współczynnika zostaje osiągnięta dla województwa łódzkiego.

Ponieważ funkcja wagowa $f_2^{y_k}; k = 1, 2$ musi być ortogonalna do $f_1^{y_k}$, nie możemy oczekiwać, że pierwsza główna składowa będzie wyjaśniała większy procent wariancji od drugiej głównej składowej. W przypadku liczby osób hospitalizowanych druga główna składowa tłumaczy około 29.41 procent zmienności wariancji. Widzimy, że osiąga pozytywne wartości przez cały okres trwania drugiej i trzeciej fali pandemii. Druga główna składowa może być interpretowana jako wskaźnik liczby osób hospitalizowanych podczas całego analizowanego czasu.

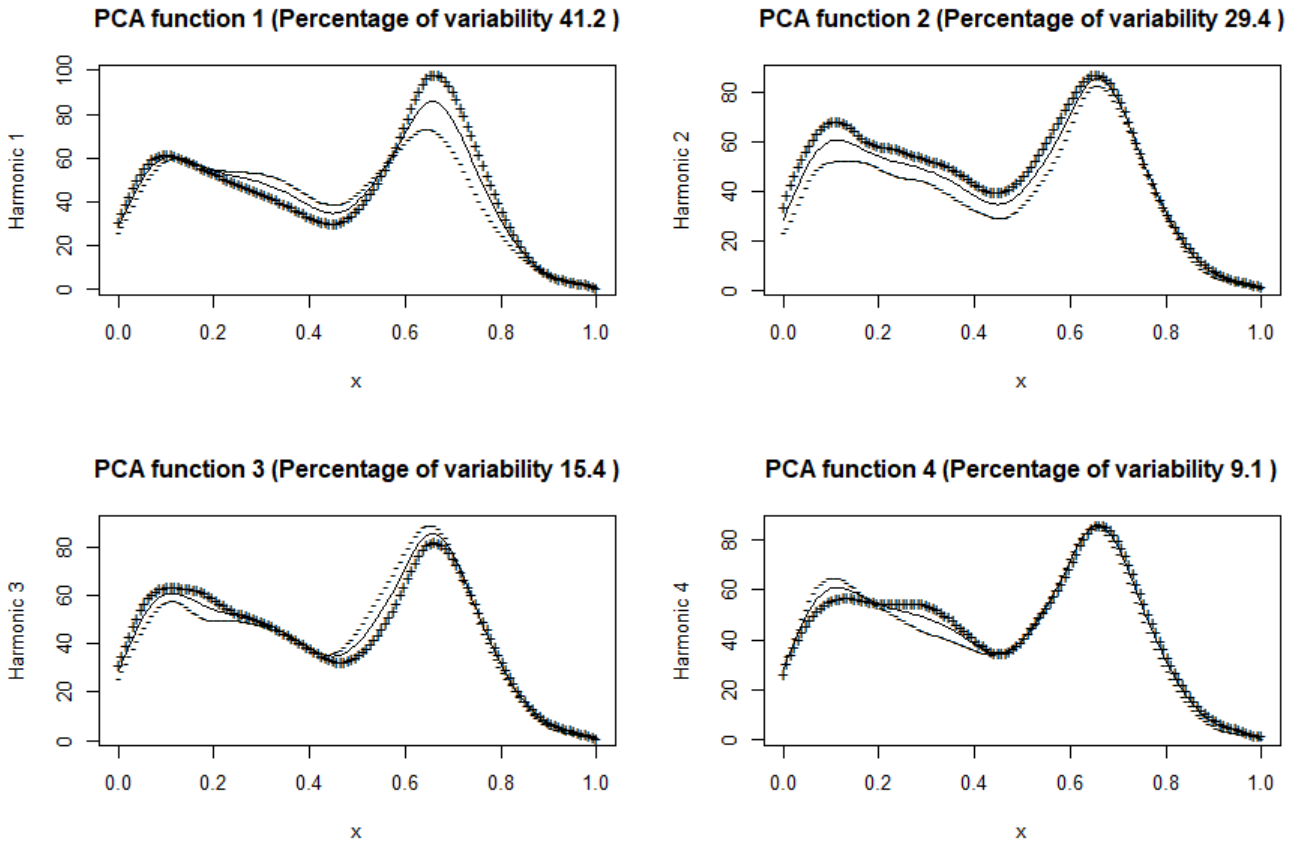
Z kolei pierwsza główna składowa dla liczby osób w ciężkim stanie sugeruje, że zmienność w liczbie osób w stanie ciężkim pomiędzy województwami była na podobnym poziomie przez prawie cały czas trwania drugiej i trzeciej fali COVID-19. W tym czasie wartości współ-

czynników były pozytywne. Najmniejsze różnice dostrzegamy na końcu trzeciej fali pandemii. Najwyższą wartość $\xi_{i1}^{y_2}$ otrzymujemy dla województwa kujawsko-pomorskiego.

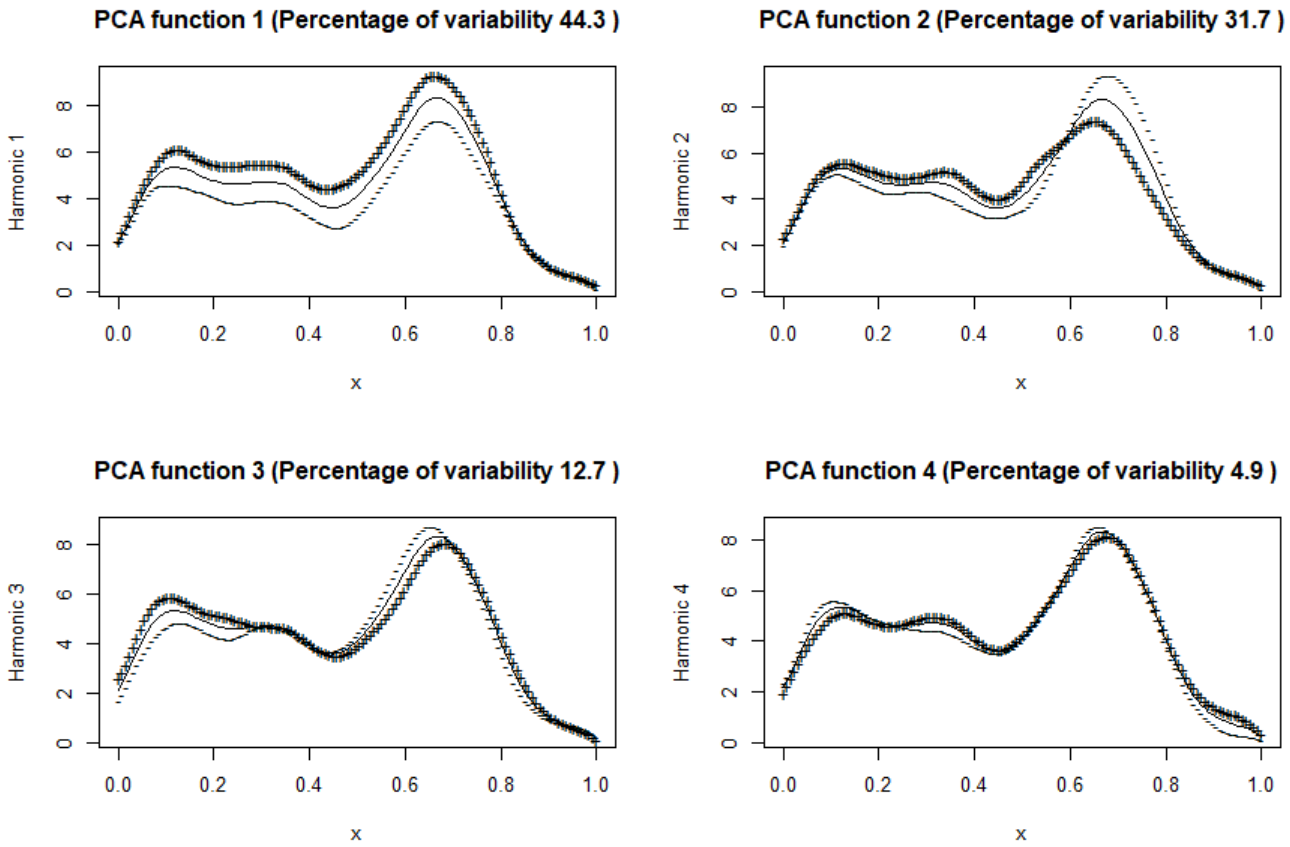
Trzecia oraz dalsze główne składowe wyjaśniają znacznie mniejszą proporcję wariancji od pierwszych dwóch składowych. Ma na to wpływ fakt, że muszą być one ortogonalne do pierwszych dwóch składowych głównych, a także ortogonalne względem siebie. Są one trudniejsze do zinterpretowania.

Interpretacja współczynników wagowych, harmonicznych nie zawsze jest prosta w kontekście funkcjonalnego PCA. Często stosuje się następującą formę prezentacji wyników.

Metodą, która jest pomocna podczas analizy funkcjonalnych głównych składowych jest przedstawienie funkcji średniej wraz z funkcjami powstałymi poprzez dodanie i odjęcie odpowiednio zwielokrotnionych funkcji harmonicznych składowych głównych od średniej. Podczas konstrukcji takiego wykresu, ważne jest, aby odpowiednio dopasować wielkość zwielokrotnienia funkcji. Taki wykres jest sensowny, ponieważ główne składowe reprezentują wariację wokół średniej.



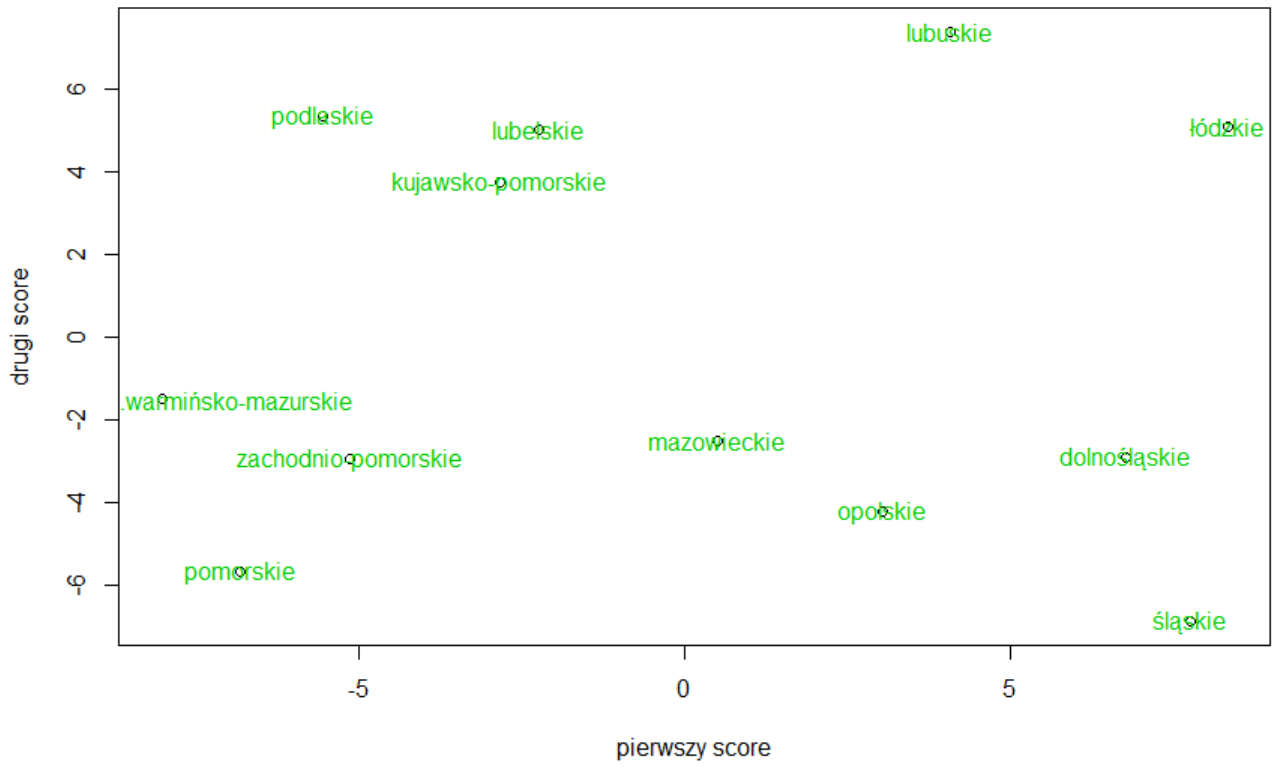
Rysunek 6: Średnia liczba osób hospitalizowanych wraz z krzywymi będącymi wynikiem dodawania (+) i odejmowania (-) odpowiednio przeskalowanych współczynników harmonicznych od średniej.



Rysunek 7: Średnia liczba osób w stanie ciężkim wraz z krzywymi będącymi wynikiem dodawania (+) i odejmowania (-) odpowiednio przeskalowanych współczynników harmonicznych od średniej.

Analizując rysunek (6), możemy zauważyć, że w kontekście liczby osób hospitalizowanych, pierwsza główna składowa wskazuje na różnice pomiędzy drugą a trzecią falą pandemii, zaś druga główna składowa pokazuje ogólny zarys pandemii. Z kolei rysunek (7) sugeruje, że dla liczby osób w stanie ciężkim sytuacja była odwrotna. Pierwsza główna składowa skupia się na ogólnym zarysie pandemii, zaś druga główna składowa pokazuje różnice pomiędzy falami. Kolejne główne składowe tłumaczą już mniejszy procent wariancji. Można to wywnioskować, analizując wykresy czwartej funkcji harmonicznej, gdzie wykresy (+) i (-) często pokrywają się z wykresem średniej. Im mniejszy procent wariancji jest wyjaśniany przez główną składową, tym bardziej wykresy (+) i (-) pokrywają się z funkcją średniej.

Rysunki (8) i (9) pokazują kolejne wykresy interesujące z punktu widzenia analizy FPCA. Prezentują one wartości zmiennych osiąganych na pierwszym i drugim *score*.



Rysunek 8: Wartości pierwszego i drugiego score dla liczby osób hospitalizowanych (województwa z próby treningowej).

Sugerując się wnioskami pochodzącymi z analizy rysunku (6), z prawej strony powyższego wykresu znajdują się województwa, w których jest największa różnica pomiędzy drugą, a trzecią falą pandemii. Z lewej strony wykresu takie, dla których różnica w liczbie osób hospitalizowanych nie jest aż tak duża w obu falach.

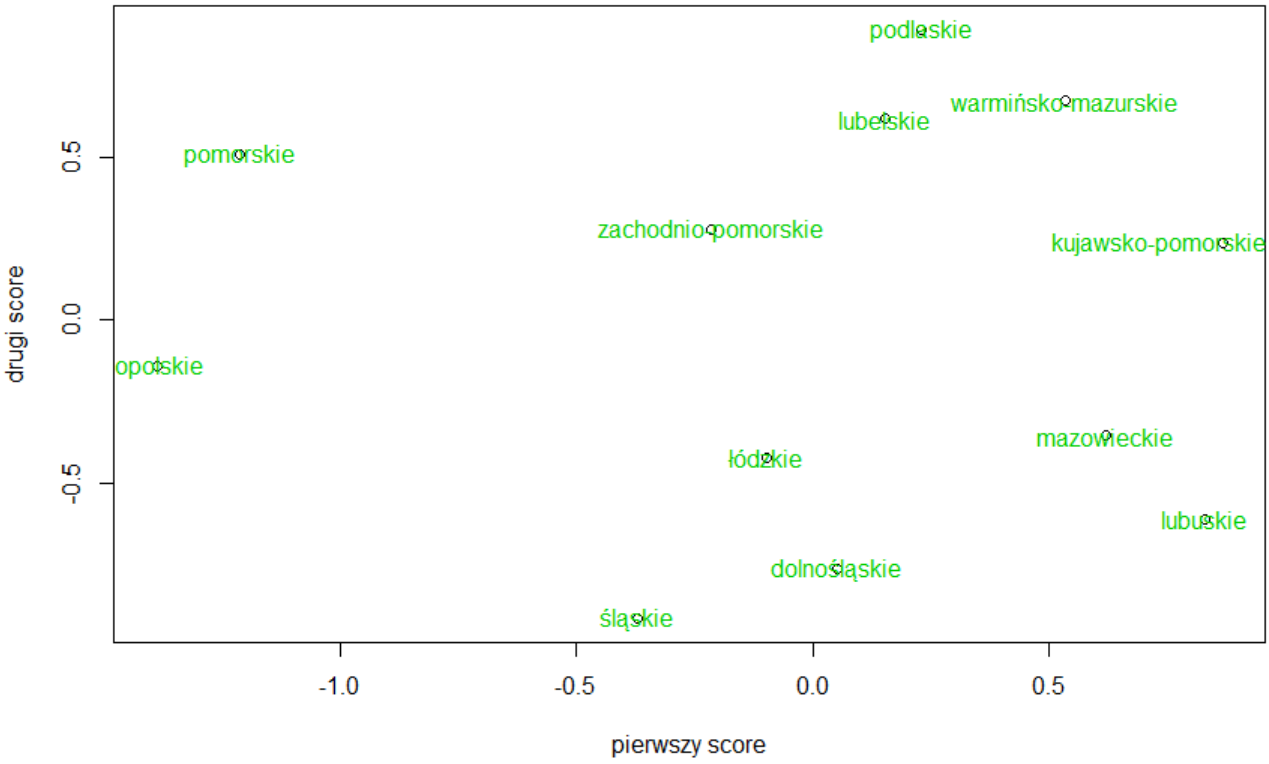
U góry wykresu będą zatem województwa, w których liczba osób hospitalizowanych była najwyższa podczas trwania drugiej i trzeciej fali pandemii. Konsekwentnie na dole wykresu obserwujemy województwa z małą liczbą osób hospitalizowanych na 100000 mieszkańców. Sugerując się rysunkiem (8), największa liczba osób hospitalizowanych została osiągnięta przez województwa lubuskie, podlaskie, lubelskie oraz łódzkie. Porównując z odpowiednio przeskalowanymi, obserwacjami dyskretnymi podawanymi przez Ministerstwo Zdrowia, największa liczba osób hospitalizowanych na 100000 mieszkańców, osiągnięta była kolejno w województwach: łódzkim, lubuskim oraz lubelskim. Województwa z najmniejszą liczbą osób hospitalizowanych, czyli śląskie, pomorskie i opolskie również pokrywają się z prawdziwymi danymi.

Stąd możemy potwierdzić wniosek, pojawiający się podczas analizy rysunku (5), że pierwsza główna składowa odpowiada za różnice pomiędzy drugą, a trzecią falą, zaś druga główna składowa jest związana z liczbą osób hospitalizowanych.

W prawym, górnym rogu mamy zatem województwa o dużej liczbie przypadków osób

hospitalizowanych i z dużymi różnicami w przebiegu drugiej i trzeciej fali. W województwach lubuskim i łódzkim było zatem najwięcej osób hospitalizowanych na 100000 mieszkańców oraz dodatkowo obie fale różniły się znacząco od siebie.

Przeprowadzimy teraz analogiczną analizę dla rysunku (9), przedstawiającego wartości pierwszego i drugiego *score* dla liczby osób w stanie ciężkim.



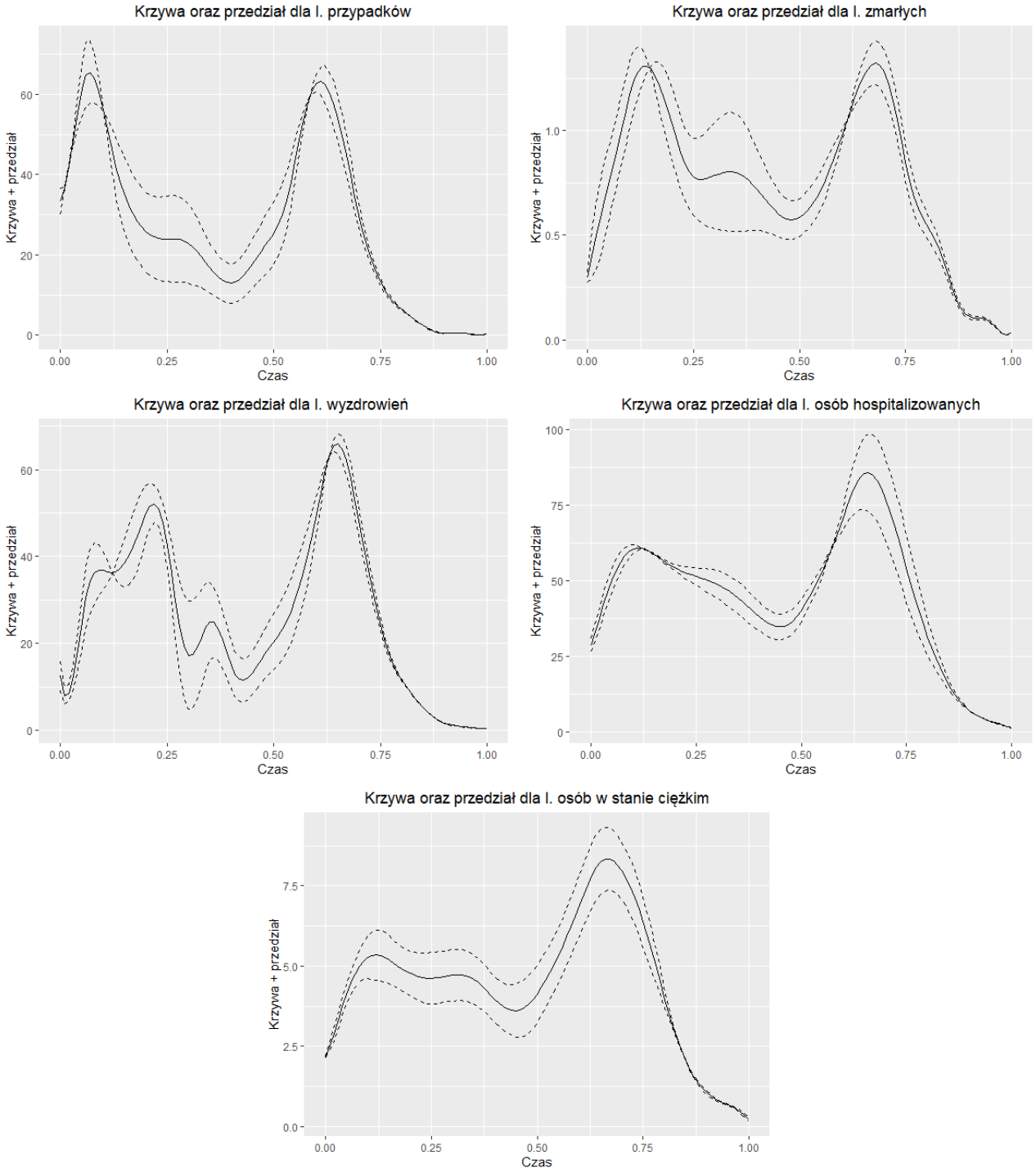
Rysunek 9: Wartości pierwszego i drugiego score dla liczby osób w stanie ciężkim (województwa z próby treningowej).

W tym przypadku u góry wykresu możemy zobaczyć województwa, dla których różnica pomiędzy drugą a trzecią falą pandemii jest wysoka, a na dole niska. Z lewej strony mamy województwa, dla których liczba osób w stanie ciężkim była niska, a z prawej wysoka. Tutaj sytuacja jest odwrotna w porównaniu z analizą przeprowadzoną dla *scores* dla liczby osób hospitalizowanych.

Według danych Ministerstwa Zdrowia największa liczba osób w stanie ciężkim na 100000 mieszkańców danego województwa została osiągnięta w województwach kujawsko-pomorskim, lubuskim oraz mazowieckim, a najmniejsza w: opolskim, pomorskim i śląskim. Zatem widzimy, że analiza rysunków (8) i (9) zdaje się być prawidłowo przeprowadzona.

Stąd w województwie kujawsko-pomorskim oraz lubuskim było najwięcej osób w stanie ciężkim, zaś województwa podlaskie i warmińsko-mazurskie miały największe różnice pomiędzy drugą i trzecią falą pandemii w liczbie osób w stanie ciężkim.

Podobne analizy można przeprowadzić dla liczby pozytywnych wyników testu, wyzdrowień oraz osób zmarłych, analizując rysunek (10) i odpowiednie wykresy *scores*.



Rysunek 10: Wykresy średnich wraz z krzywymi będącymi wynikiem dodawania (+) i odejmowania (-) odpowiednio zwielokrotnionych, współczynników harmonicznych, pierwszych głównych składowych od średnich.

4.2 Model Function-on-Function

Rozważmy próbę treningową złożoną ze wszystkich województw z wyjątkiem świętokrzyskiego, małopolskiego, wielkopolskiego oraz podkarpackiego. Wymienione województwa posłużą jako próba testowa i na nich dokonamy predykcji.

Równanie (13) przedstawia redukcję liniowego modelu *function-on-function*, przedstawionego za pomocą równania (11), do modelu liniowego dla pierwszych głównych składowych odpowiedzi w terminach pierwszych głównych składowych każdego z predyktorów:

$$\hat{\xi}_{i1}^{y_k} = \gamma_0 + \xi_{i1}^{x_1}\gamma_1^{y_k} + \xi_{i1}^{x_2}\gamma_2^{y_k} + \xi_{i1}^{x_3}\gamma_3^{y_k} + \epsilon_i^{y_k},\ k = 1, 2, i = 1, ..., 16, \tag{13}$$

gdzie $\gamma_0, \gamma_1^{y_k}, \gamma_2^{y_k}, \gamma_3^{y_k}$ oznaczają odpowiednie współczynniki uzyskane poprzez dopasowanie do danych modelu liniowego. Na podstawie takich modeli dokonamy estymacji pierwszych składowych $Y_1(t)$ i $Y_2(t)$ z pierwszych składowych $X_1(t)$, $X_2(t)$ i $X_3(t)$.

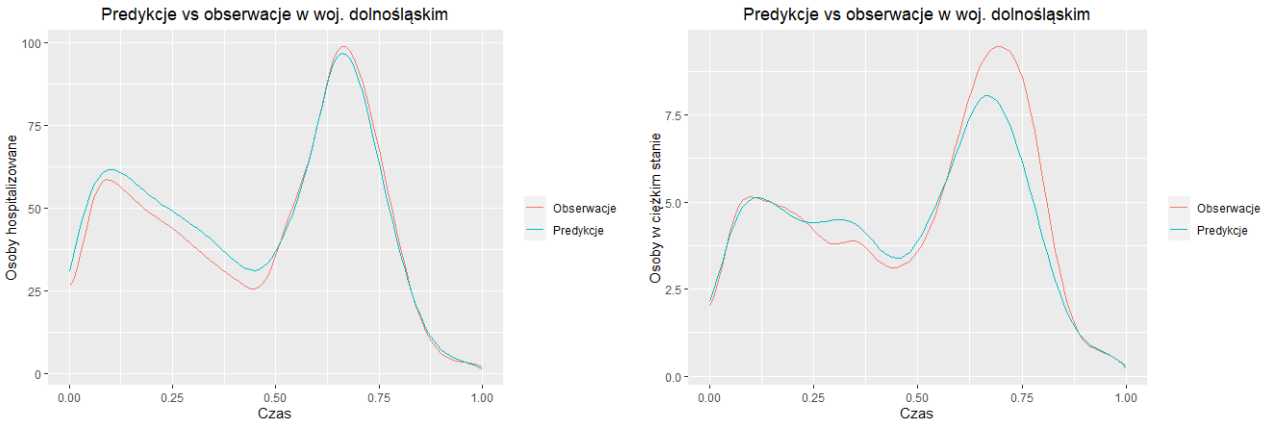
Predykcję $Y_1(t)$ i $Y_2(t)$ dokonamy za pomocą następującego równania:

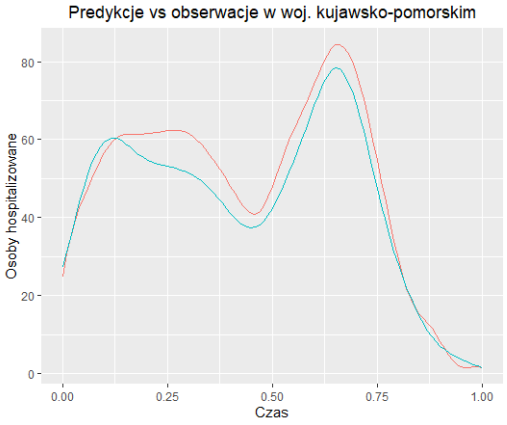
$$\hat{y}_{ik}(t) = \overline{y}_k(t) + \hat{\xi}_{i1}^{y_k} f_1^{y_k}(t),\ k = 1, 2, i = 1, ..., 16. \tag{14}$$

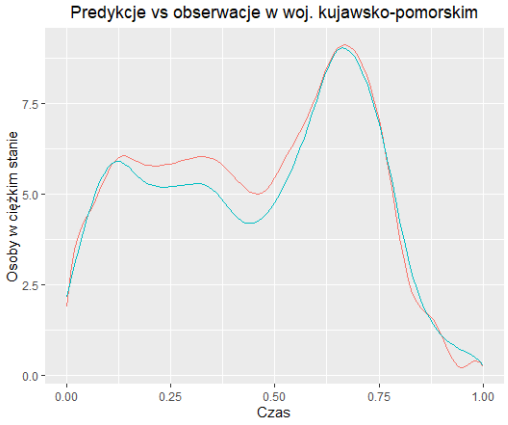
W celu przetestowania modeli na próbie treningowej posłużymy się błędem średniokwadratowym, wyrażonym równaniem:

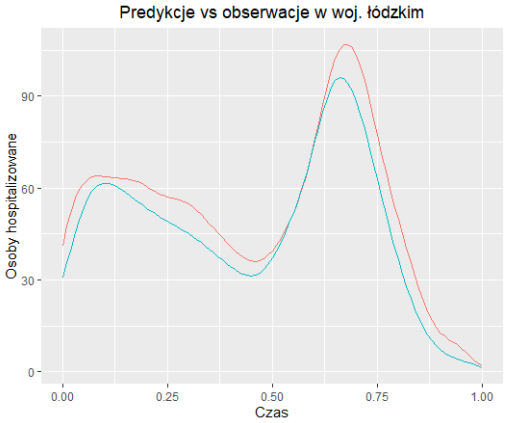
$$\text{MSE}(y_{ik}) = \int_0^1 (y_{ik}(t) - \hat{y}_{ik}(t))^2 dt,\ k = 1, 2, i = 1, ...12. \tag{15}$$

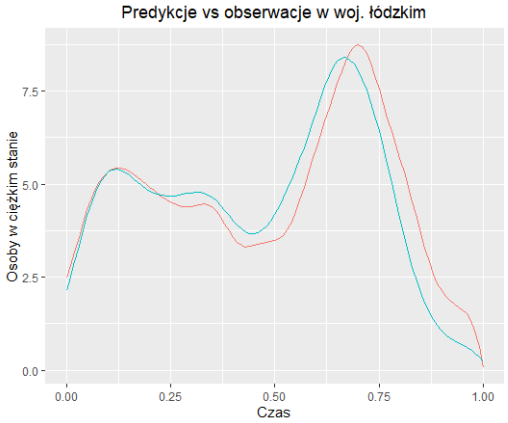
Rysunek (11) przedstawia wykresy obserwowanych krzywych, dopasowanych w podrozdziale 2.5 wraz z przewidywanymi krzywymi otrzymanymi poprzez zastosowanie wzoru (14) dla kilku województw wylosowanych z próby treningowej. W tabeli (1) można zobaczyć wartości błędu średniokwadratowego obliczonego za pomocą równania (15) dla wszystkich województw z próby treningowej.

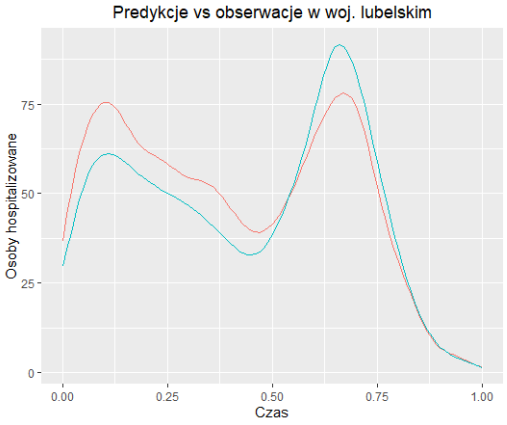


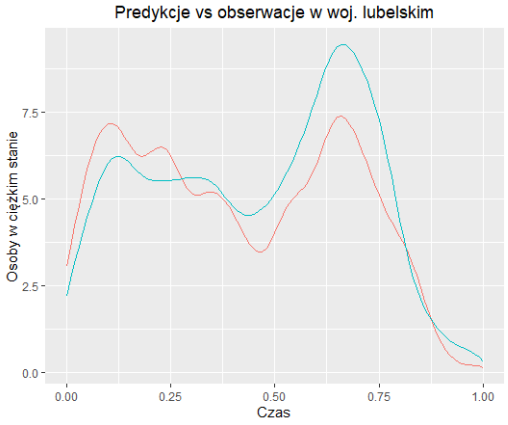

Predykcje vs obserwacje w woj. kujawsko-pomorskim
Osoby hospitalizowane
Czas
Obserwacje
Predykcje

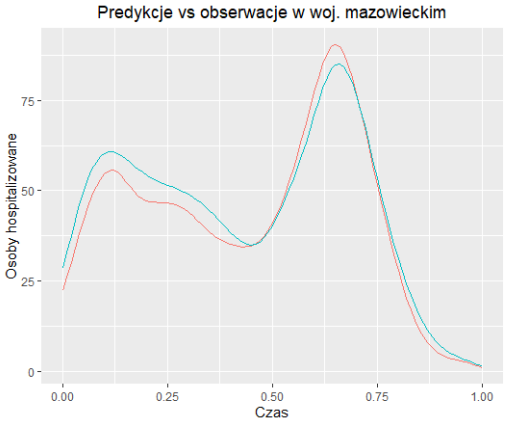

Predykcje vs obserwacje w woj. kujawsko-pomorskim
Osoby w ciężkim stanie
Czas
Obserwacje
Predykcje

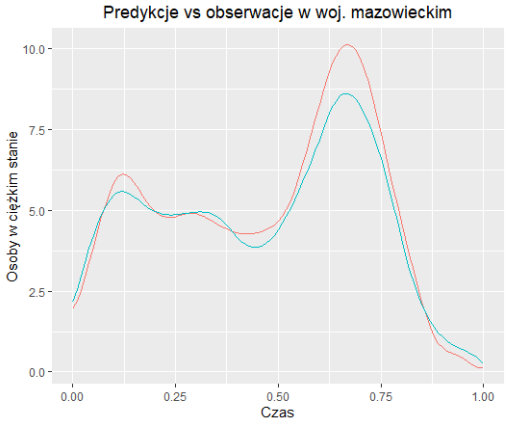

Predykcje vs obserwacje w woj. łódzkim
Osoby hospitalizowane
Czas
Obserwacje
Predykcje

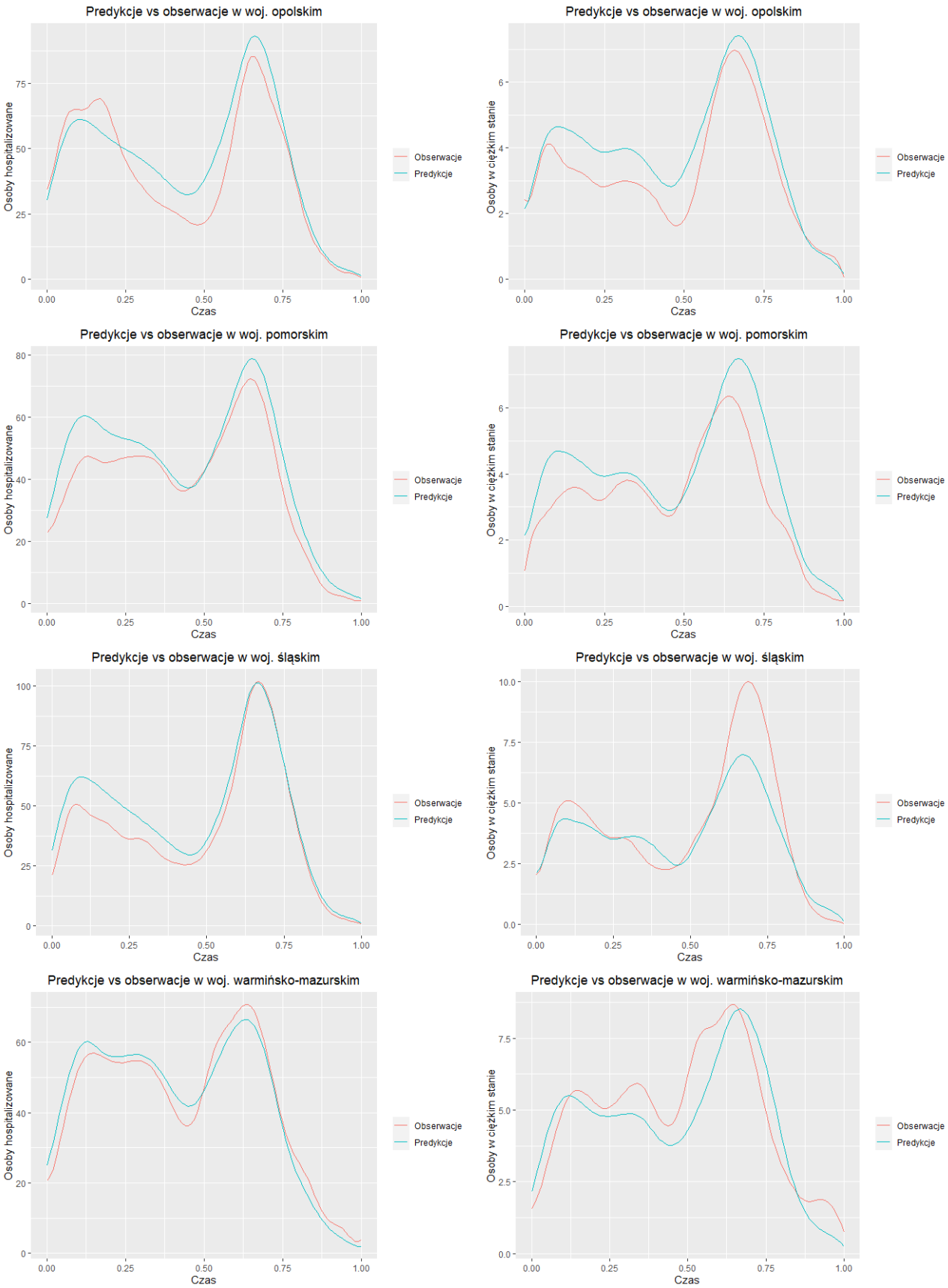

Predykcje vs obserwacje w woj. łódzkim
Osoby w ciężkim stanie
Czas
Obserwacje
Predykcje


Predykcje vs obserwacje w woj. lubelskim
Osoby hospitalizowane
Czas
Obserwacje
Predykcje


Predykcje vs obserwacje w woj. lubelskim
Osoby w ciężkim stanie
Czas
Obserwacje
Predykcje


Predykcje vs obserwacje w woj. mazowieckim
Osoby hospitalizowane
Czas
Obserwacje
Predykcje


Predykcje vs obserwacje w woj. mazowieckim
Osoby w ciężkim stanie
Czas
Obserwacje
Predykcje



Rysunek 11: Obserwowane i przewidywane krzywe dla wybranych województw z próby treningowej.

Tabela 1: Wartości błędu średniokwadratowego dla y_{i1} (liczby osób hospitalizowanych) i y_{i2} (liczby osób w stanie ciężkim) dla województw z próby treningowej.

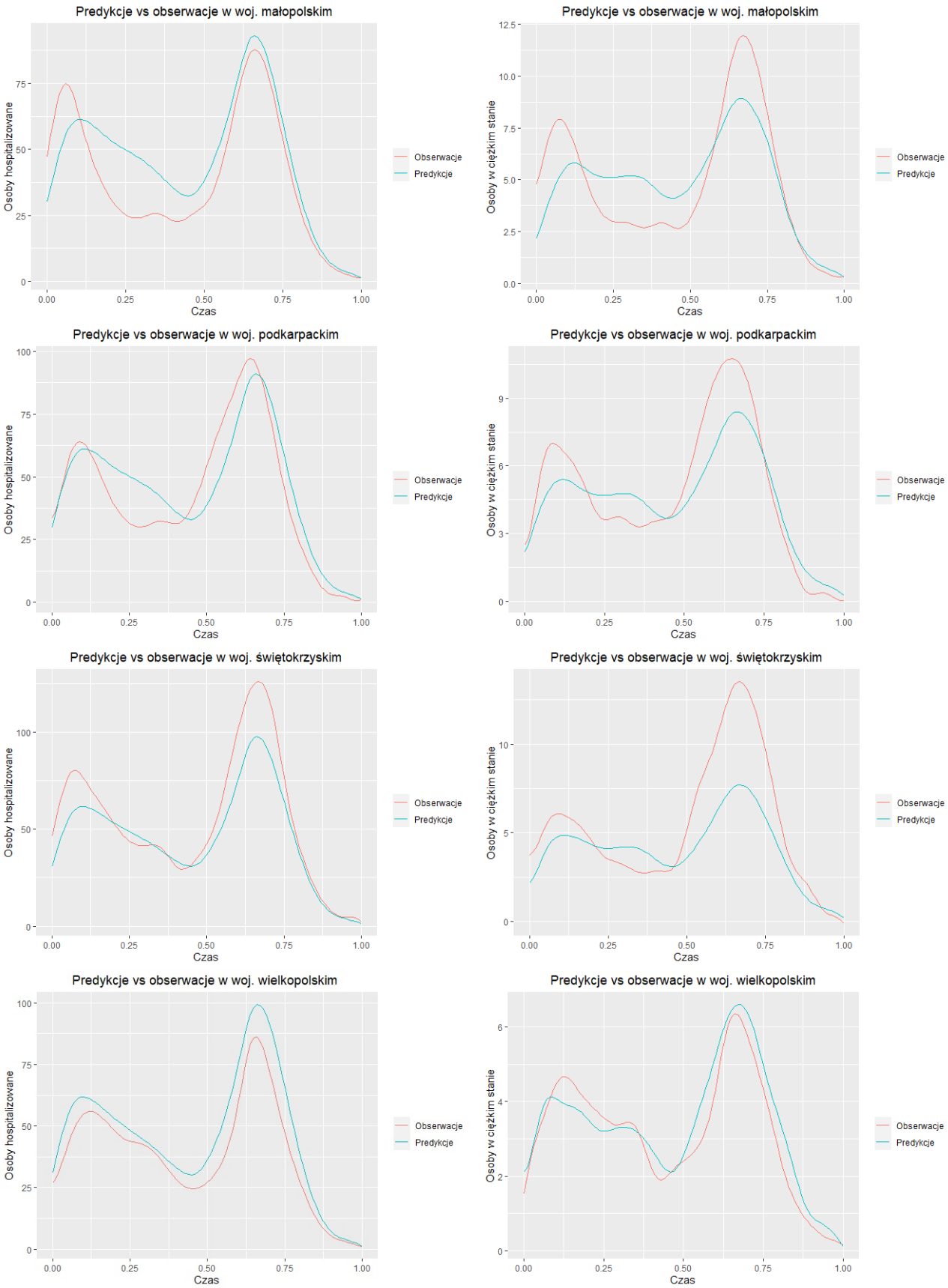
województwo	MSE(y_{i1})	MSE(y_{i2})
dolnośląskie	**4.087584**	0.8424024
kujawsko-pomorskie	5.801758	**0.5078765**
łódzkie	8.146502	0.7999353
lubelskie	8.434905	**1.1796471**
lubuskie	**10.79992**	0.9807714
mazowieckie	4.569325	0.6134904
opolskie	9.170555	0.7959462
podlaskie	9.612532	1.0092404
pomorskie	7.526031	0.9541513
śląskie	7.451072	1.1456427
warmińsko-mazurskie	3.906237	1.0330436
zachodnio-pomorskie	6.692964	0.9105502

Czcionką wytłuszczoną zaznaczono odpowiednio najniższe i najwyższe wartości MSE. Widzimy, że wartości błędu średniokwadratowego dla liczby osób w stanie ciężkim są znacznie niższe od wartości błędu średniokwadratowego dla liczby osób hospitalizowanych. Najniższą wartość MSE(y_{i2}) otrzymano dla województwa kujawsko-pomorskiego, a najwyższą dla województwa lubelskiego. Z kolei wartość błędu średniokwadratowego dla y_{i1} jest najniższa dla województwa dolnośląskiego, a najwyższa dla lubuskiego.

Rysunek (12) przedstawia predykcje dla województw z próby testowej, a tabela (2) wartości przewidywane wraz z obserwowanymi, podawanymi przez Ministerstwo Zdrowia w trakcie kilku pierwszych i ostatnich dni analizowanego czasu pandemii. W celu porównania dokonanych przewidywań z obserwowanymi danymi, predykcje umieszczone w tabeli (2) zostały odpowiednio przeskalowane i prezentują liczbę osób hospitalizowanych oraz w stanie ciężkim przypadającą na dane województwo danego dnia, a nie liczbę osób przypadającą na 100000 mieszkańców.

Analizując rysunek (12) oraz tabelę (2) możemy zauważyć, że predykcje dla województwa wielkopolskiego okazały się bardzo bliskie prawdziwym wartościom. Najgorzej model poradził sobie z województwem małopolskim. Przykładowo pierwszego listopada odnotowano w województwie wielkopolskim 102 osoby w stanie ciężkim, a model przewidział 107 osób; 217 obserwowano w województwie małopolskim, zaś model przewidział 126. Tego samego dnia w województwie podkarpackim stwierdzono 95 osób w stanie ciężkim przeciwko 76 przewidywanym, a w województwie świętokrzyskim 55 przeciwko 41. Z kolei liczba osób hospitalizowanych pierwszego listopada w województwie małopolskim wyniosła 2366 osób, zaś model przewidział 1591 przypadków; w województwie podkarpackim 976 przeciwko 977 przewidywanym; w województwie świętokrzyskim 829 obserwowano na 589 przewidywanych, a w województwie wielkopolskim 1292 przeciwko 1700.

Analizując rysunek (12) oraz tabelę (2) możemy zauważyć, że najbliższe obserwowanym wartościom wyniki przewidziano na początku oraz na końcu analizowanego czasu. Największa różnica pomiędzy wartościami obserwowanymi a przewidywanymi jest dla województwa małopolskiego. Tutaj model przewiduje czasem niemal o około 2 razy niższe wartości niż zaobserwowano.



Rysunek 12: Obserwowane i przewidywane krzywe dla próby testowej.

Tabela 2: Przewidywane wartości liczby osób hospitalizowanych oraz w stanie ciężkim w porównaniu z obserwowanymi wartościami podawanymi przez Ministerstwo Zdrowia. Obserwacje oznaczono przez „obs", a predykcje - „pred".

Osoby hospitalizowane				
	małopolskie	podkarpackie	świętokrzyskie	wielkopolskie
czas	obs/pred	obs/pred	obs/pred	obs/pred
23.10	1671/1023	681/629	559/377	924/1088
24.10	1667/1086	750/668	581/401	929/1158
25.10	1763/1151	757/707	589/426	961/1229
26.10	1907/1216	796/747	704/450	1039/1300
27.10	1972/1281	819/786	742/474	1090/1370
28.10	2003/1345	813/826	750/498	1137/1439
29.10	2072/1408	848/865	792/522	1182/1507
30.10	2250/1471	883/903	789/545	1241/1574
31.10	2289/1532	917/941	779/567	1252/1638
1.11	2366/1591	976/977	829/589	1292/1700
...	...	...	...	...
26.06	72/106	31/66	69/39	81/111
27.06	72/100	26/62	61/37	79/105
28.06	66/95	25/58	57/34	82/99
29.06	63/89	20/55	54/32	82/93
30.06	57/82	15/51	51/30	63/86
1.07	53/76	13/47	47/27	42/79
2.07	51/69	13/43	39/25	44/71
3.07	53/62	13/39	36/22	44/62
4.07	48/54	15/34	32/19	44/53
5.07	41/46	14/29	33/15	39/43
Osoby w stanie ciężkim				
23.10	163/74	50/46	46/26	59/74
24.10	160/80	54/49	46/27	56/75
25.10	172/85	56/52	46/29	75/77
26.10	188/91	68/55	51/30	74/80
27.10	207/97	67/58	52/32	77/83
28.10	215/103	72/62	50/34	68/87
29.10	209/109	70/65	50/35	80/92
30.10	208/115	81/69	53/37	94/97
31.10	212/121	90/72	55/39	94/102
1.11	217/126	95/76	55/41	102/107
...	...	...	...	...
26.06	11/21	5/13	4/7	11/18
27.06	11/20	4/12	3/7	10/17
28.06	14/19	3/11	2/6	9/16
29.06	12/18	2/11	2/6	9/14
30.06	11/17	2/10	2/5	9/13
1.07	11/16	0/9	1/5	7/11
2.07	11/14	0/8	0/4	7/9
3.07	11/13	0/7	0/4	7/8
4.07	11/11	1/6	0/3	7/6
5.07	10/9	0/5	0/2	5/4

5 Porównanie fal pandemii w Polsce

Przeprowadzając analizę w rozdziale 4, można dostrzec istotne różnice pomiędzy drugą, a trzecią falą pandemii. Rozdzielenie i analiza dwóch fal osobno, mogłaby przynieść ciekawe wnioski dotyczące województw w których rozwijała się choroba. W tym rozdziale zostanie przeprowadzona analiza głównych składowych funkcjonalnych oraz zostanie dopasowany model MFFLR do dwóch fal pandemii osobno.

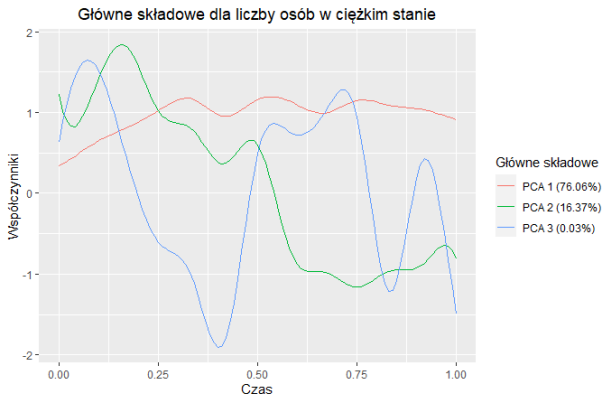
5.1 Druga fala pandemii

Analizując wartości błędu średniokwadratowego w zależności od liczby funkcji bazowych dla drugiej fali pandemii wybrano 15 funkcji bazowych. Jest to niższa wartość niż w przypadku, gdy analizowaliśmy obie fale razem, ale wciąż dość wysoka, aby nie utracić istotnych informacji.

5.1.1 Analiza Głównych Składowych Funkcjonalnych

Przeprowadzono estymację funkcjonalnych głównych składowych. Okazało się, że pierwsze główne składowe wyjaśniają kolejno 68%, 54%, 57%, 58%, 76% zmienności X_1, X_2, X_3, Y_1, Y_2.

Poniżej przedstawiono wykresy funkcji harmonicznych pierwszych 3 głównych składowych X_1, X_2, X_3, Y_1, Y_2.



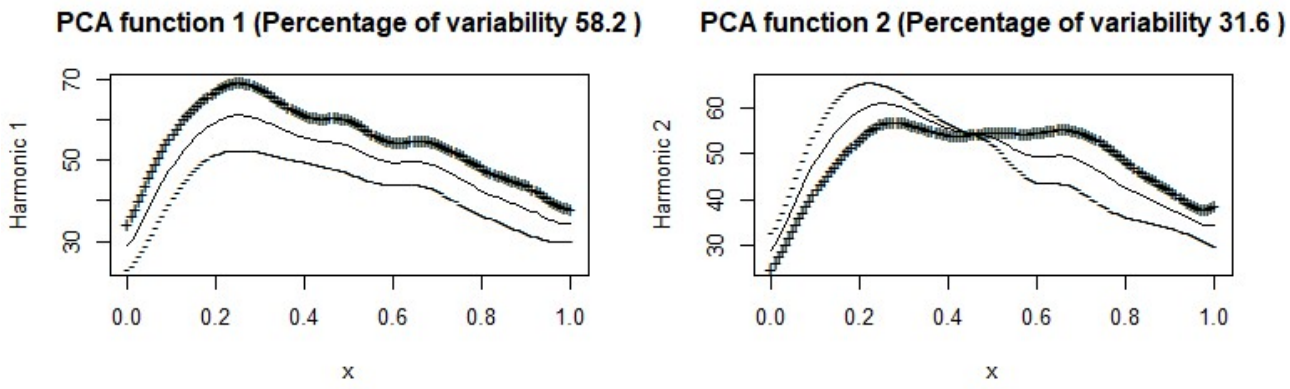
Rysunek 13: Funkcje wagowe dla 3 pierwszych głównych składowych dla drugiej fali pandemii.

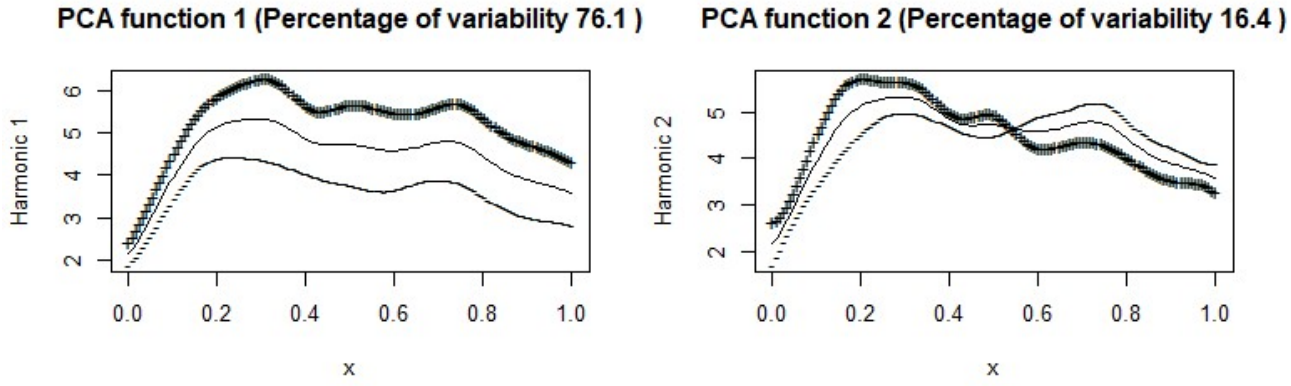
Pierwsza i druga główna składowa dla liczby osób hospitalizowanych tłumaczą razem 89.77 procent wariancji. Z kolei dla liczby osób w ciężkim stanie już pojedyncza pierwsza składowa tłumaczy aż 76.05%, a razem z drugą wyjaśniają 92.43%. Są to znacznie wyższe wartości niż w przypadku przeprowadzanej wcześniej, wspólnej analizy dla drugiej i trzeciej fali, przedstawionej na rysunku (5).

Ponownie, aby lepiej zrozumieć powyższe wykresy, spróbujemy przeanalizować wartości pierwszego i drugiego *score* oraz wykresy funkcji średniej wraz z odpowiednimi wariacjami funkcji składowych głównych.

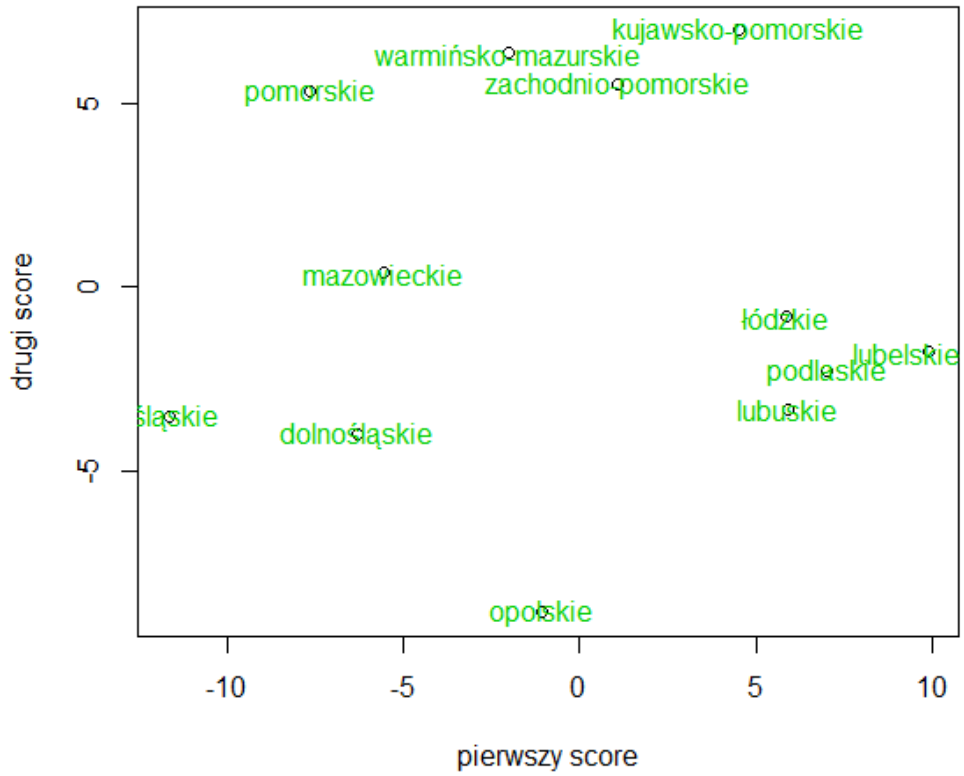
Analizując rysunek (14) możemy zauważyć, że pierwsze główne składowe tłumaczą ogólną tendencję liczby osób hospitalizowanych oraz osób w stanie ciężkim. Z kolei drugie główne składowe pokazują różnice pomiędzy początkiem a końcem drugiej fali pandemii.

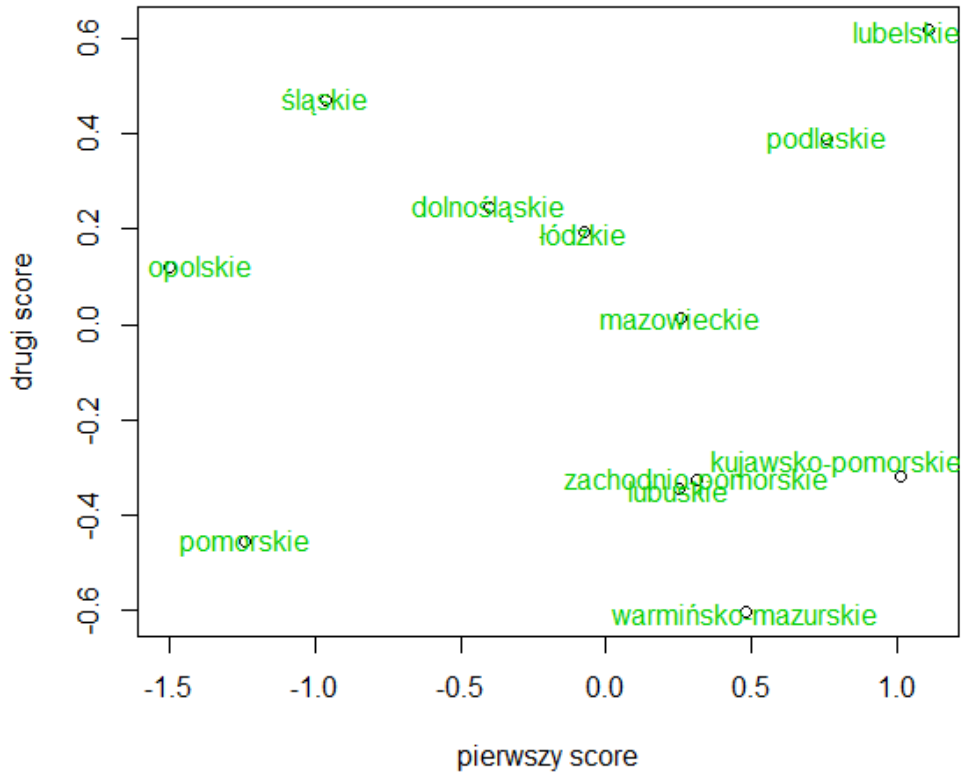
Możemy zatem wyciągnąć wniosek do rysunku (15), że województwa umieszczone z prawej strony rysunków, powinny osiągnąć w sumie największą liczbę osób hospitalizowanych i w stanie ciężkim, a te znajdujące się po lewej stronie wykresów najniższą w trakcie trwania drugiej fali pandemii. Co ciekawe województwo lubelskie miało zarówno najwięcej osób hospitalizowanych jak i w stanie ciężkim w trakcie trwania drugiej fali pandemii. Najmniej osób hospitalizowanych było w województwie śląskim, a osób w stanie ciężkim w opolskim.





Rysunek 14: Wykresy średnich wraz z krzywymi będącymi wynikiem dodawania (+) i odejmowania (-) odpowiednio przeskalowanych, współczynników harmonicznych pierwszych głównych składowych od średnich. Górne rysunki dotyczą liczby osób hospitalizowanych, zaś dolne osób w stanie ciężkim.





Rysunek 15: Wartości pierwszego i drugiego score dla liczby osób hospitalizowanych (górny wykres) i liczby osób w stanie ciężkim (dolny wykres) dla drugiej fali pandemii.

5.1.2 Model Function-on-Function

Posłużymy się tą samą metodą predykcji liczby osób hospitalizowanych i liczby osób w stanie ciężkim jak w podrozdziale 4.2. W tabeli (3) widzimy wartości MSE, obliczone za pomocą wzoru (15), zaś rysunek (16) przedstawia obserwowane i przewidywane krzywe dla kilku województw z próby treningowej. Rysunek (17) oraz tabela (4) ilustrują predykcje dokonane na próbie testowej.

W tabeli (3) czcionką wytłuszczoną zaznaczono odpowiednio najniższe i najwyższe wartości $\mathrm{MSE}(y_{i1})$ oraz $\mathrm{MSE}(y_{i2})$.

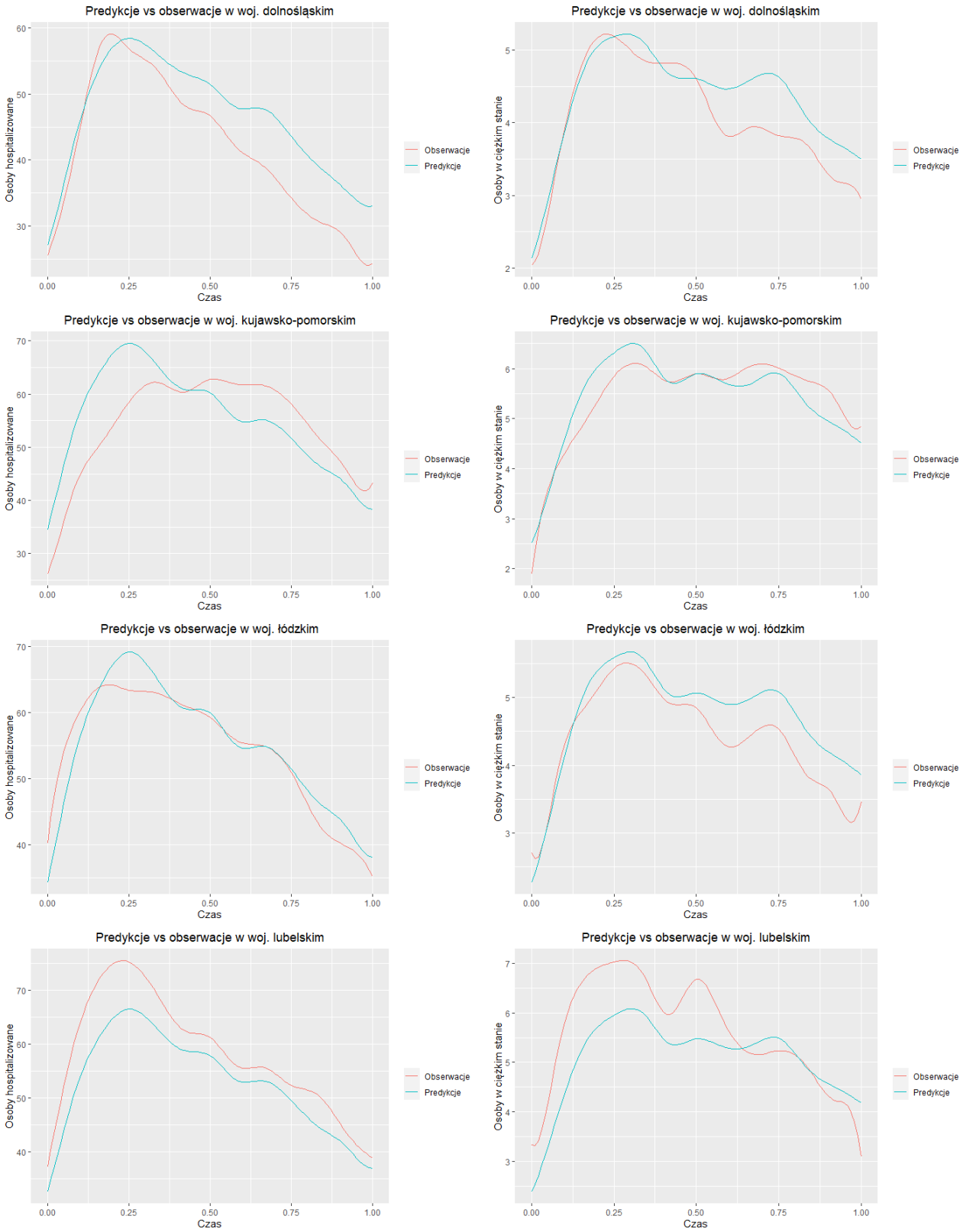
Tabela 3: Wartości błędu średniokwadratowego dla y_{i1} (liczby osób hospitalizowanych) i y_{i2} (liczby osób w stanie ciężkim) dla województw z próby treningowej dla drugiej fali.

województwo	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	9.430756	0.7103345
zachodnio-pomorskie	6.872950	0.4965603

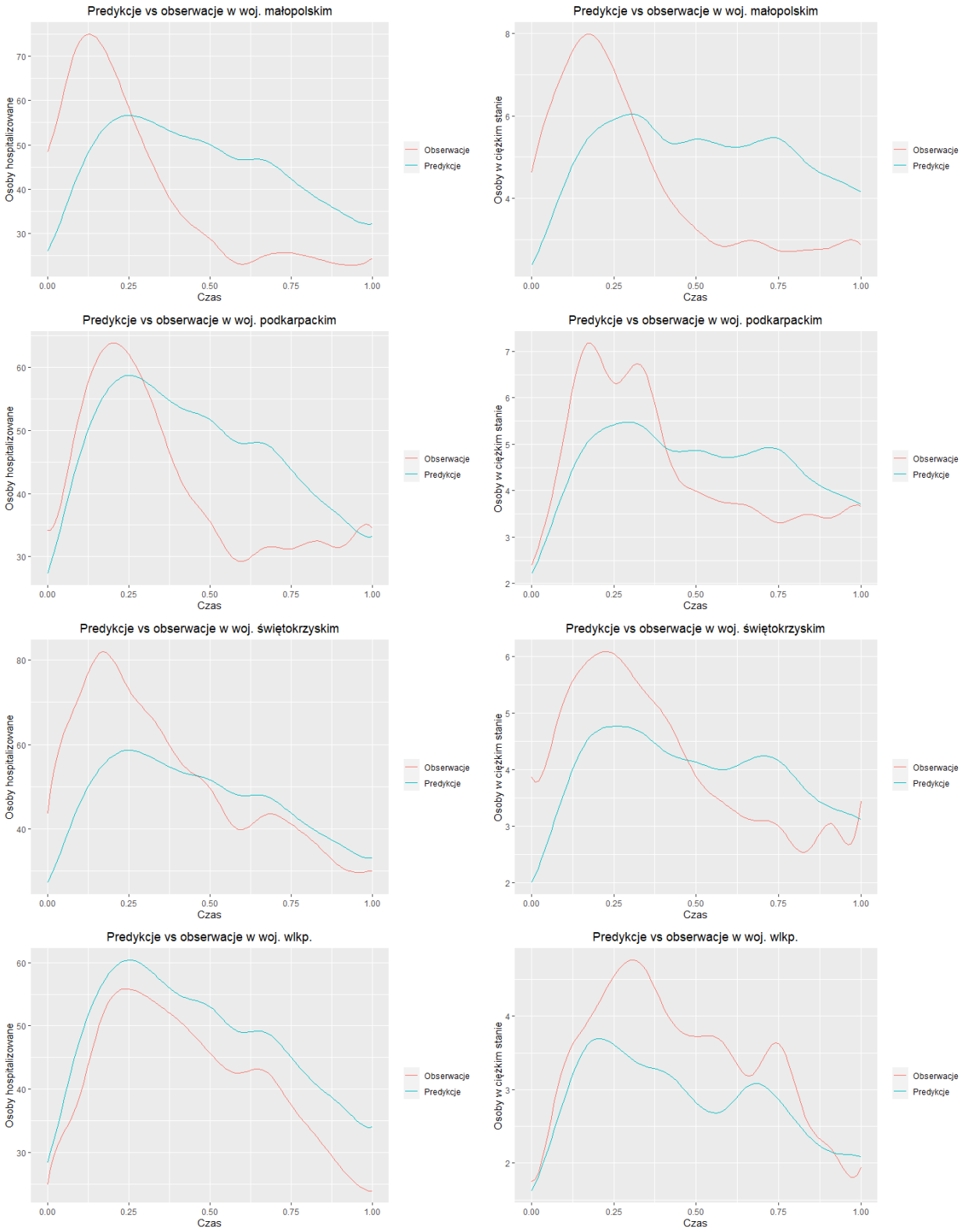
Dla niemal wszystkich województw MSE(y_{i2}), w porównaniu z modelem dla obu fal razem przedstawionym za pomocą tabeli (1), spadło. Wyjątkiem jest województwo podlaskie, gdzie wcześniej MSE(y_{i2}) wyniosło 1.0092404, zaś teraz dla modelu pojedynczej fali 1.3927908. Z kolei MSE(y_{i1}) poprawiło się tylko dla województw: łódzkiego (wcześniej 8.146502), lubelskiego (8.434905), lubuskiego (10.79992) oraz pomorskiego (7.526031). Pozostałe województwa miały niższe wartości MSE(y_{i1}) podczas wspólnej analizy fal pandemii. Stąd możemy wyciągnąć wniosek, że przewidując liczbę osób hospitalizowanych dla drugiej fali warto brać pod uwagę również trzecią falę. Z kolei przeprowadzając analizę liczby osób w stanie ciężkim najlepsze rezultaty osiągniemy analizując pojedynczą, drugą falę.

Najlepiej model pojedynczej, drugiej fali dla liczby osób hospitalizowanych dopasował się do danych dla województwa łódzkiego. Otrzymane MSE wyniosło jedynie 3.205443. Z kolei najniższe MSE(y_{i2}) otrzymano dla województwa mazowieckiego.

Analizując rysunek (17) oraz tabelę (4) widzimy, że dla województwa małopolskiego, podkarpackiego i wielkopolskiego przewidywana liczba osób w stanie ciężkim jest bliższa wartościom podawanym przez Ministerstwo Zdrowia niż podczas analizy danych pochodzących z obu fal pandemii. Wyjątkiem jest województwo świętokrzyskie, gdzie predykcja pogorszyła się nieznacznie. Z kolei predykcja liczby osób hospitalizowanych pogorszyła się dla wszystkich województw z grupy testowej.



Rysunek 16: Obserwowane i przewidywane krzywe dla wybranych województw z próby treningowej dla drugiej fali pandemii.



Rysunek 17: Obserwowane i przewidywane krzywe dla województw z próby testowej dla drugiej fali pandemii.

Tabela 4: Przewidywane wartości liczby osób hospitalizowanych oraz w stanie ciężkim w trakcie drugiej fali pandemii w porównaniu z obserwowanymi wartościami podawanymi przez Ministerstwo Zdrowia.

Osoby hospitalizowane				
	małopolskie	podkarpackie	świętokrzyskie	wielkopolskie
czas	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
Osoby w stanie ciężkim				
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

5.2 Trzecia fala pandemii

W tym podrozdziale przeprowadzimy analogiczną analizę dla trzeciej fali, co w poprzednim podrozdziale przeznaczonym drugiej fali.

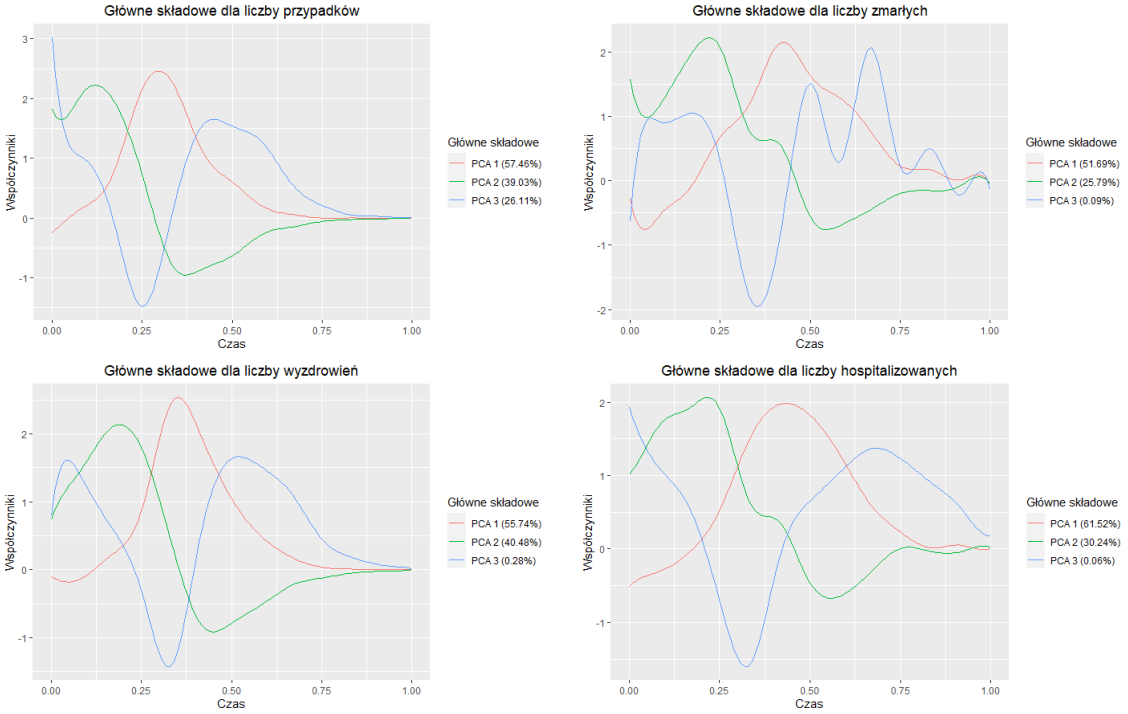
5.2.1 Analiza Głównych Składowych Funkcjonalnych

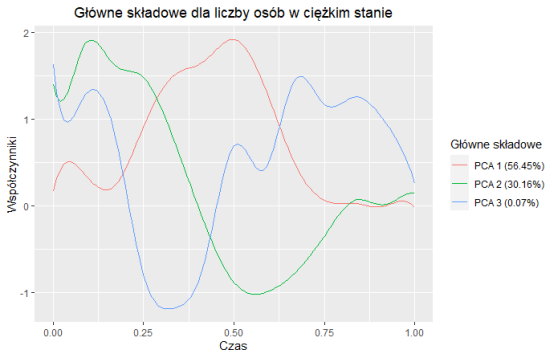
W tym podrozdziale przedstawimy wyniki analizy danych dotyczących trzeciej fali pandemii. Analizę przeprowadzimy w analogiczny sposób jak w przypadku przeprowadzonej w podrozdziale 5.1.1 analizy dla drugiej fali.

Podobnie jak dla drugiej fali pandemii na podstawie wartości błędu średniokwadratowego ustalono liczbę funkcji bazowych na 15.

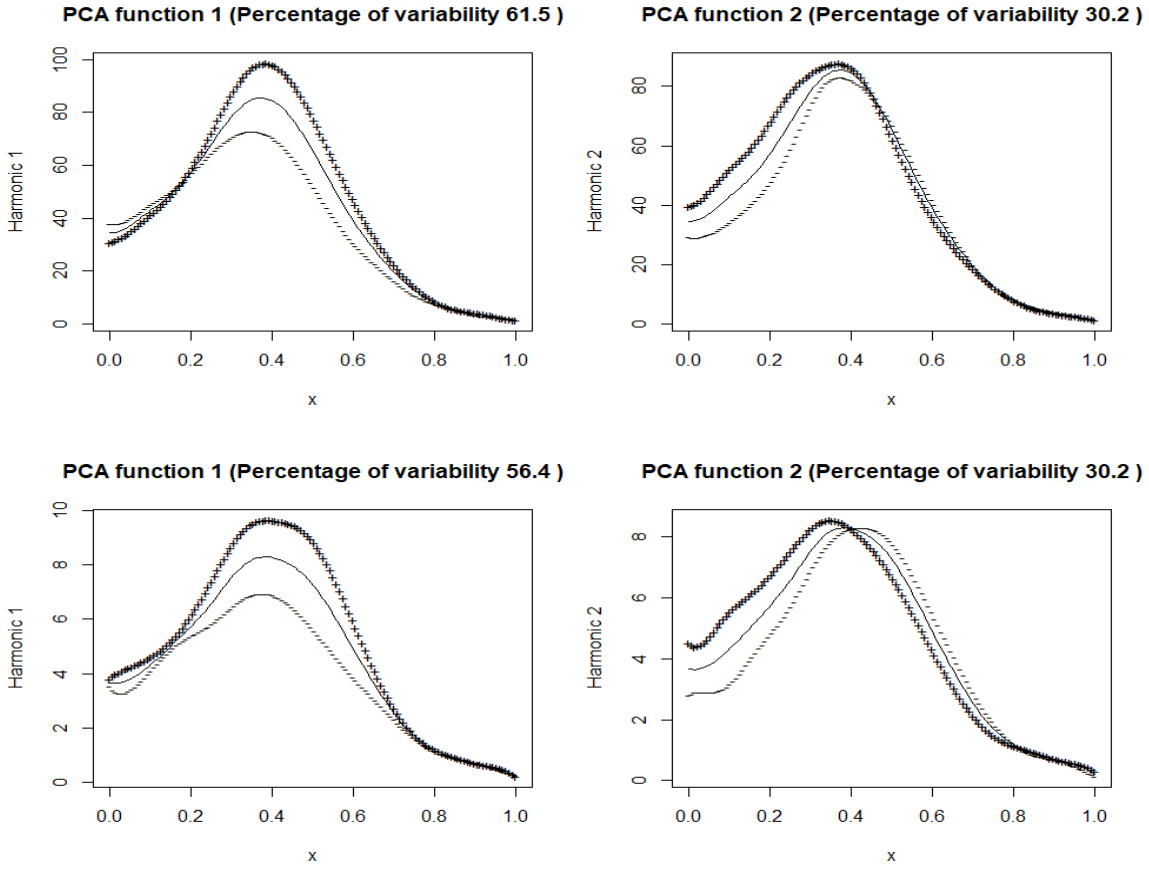
Zmienność X_1, X_2, X_3, Y_1, Y_2 wyjaśniana przez pierwsze główne składowe to kolejno: 57%, 52%, 56%, 62%, 56%.

Na rysunku (18) przedstawiono wykresy funkcji harmonicznych głównych składowych, na rysunku (19) wykresy średnich wraz z odpowiednimi wariacjami wokół średnich oraz na rysunku (20) wykres pierwszego i drugiego *score*.

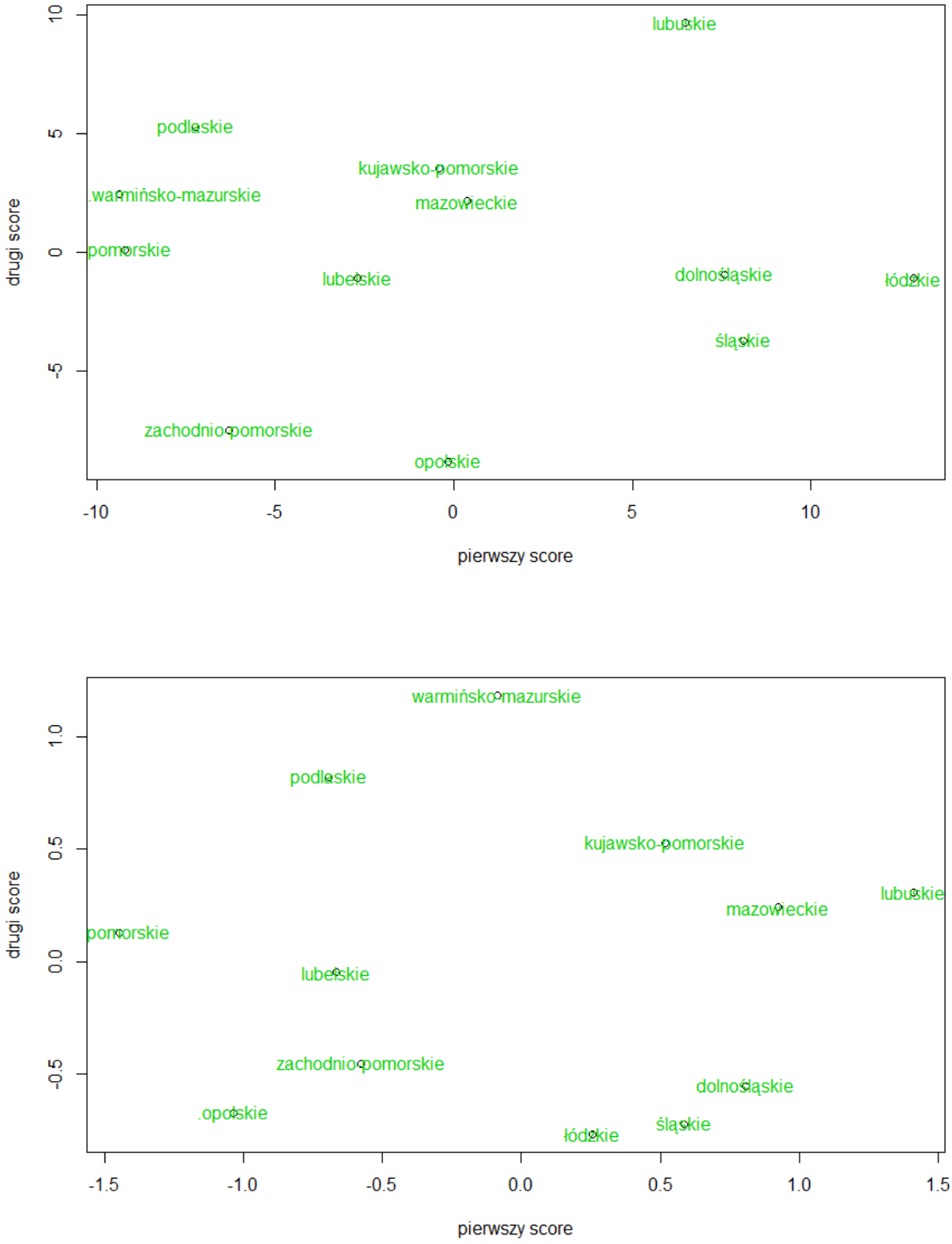




Rysunek 18: Funkcje wagowe dla 3 pierwszych głównych składowych dla trzeciej fali.



Rysunek 19: Średnia liczba osób hospitalizowanych (u góry wykresu) oraz osób w stanie ciężkim (u dołu wykresu) wraz z krzywymi będącymi wynikiem dodawania (+) i odejmowania (-) odpowiednio zwielokrotnionych współczynników harmonicznych od średniej.



Rysunek 20: Wartości pierwszego i drugiego score dla liczby osób hospitalizowanych (wykres u góry strony) i dla liczby osób w stanie ciężkim (wykres u dołu strony) dla trzeciej fali pandemii.

Tutaj za zmienność zarówno w liczbie osób hospitalizowanych jak i w liczbie osób w stanie ciężkim odpowiada pierwszy *score*, zaś za różnicę pomiędzy pierwszą a drugą połową trzeciej fali drugi *score*. Jest to analogiczna sytuacja do drugiej fali pandemii. Inne wyniki otrzymywaliśmy podczas analizy liczby osób hospitalizowanych dla obu fal razem. Wówczas pierwsza główna składowa odpowiadała za różnice pomiędzy falami, a druga za ogólną tendencję w liczbie przypadków.

Z analizy rysunków (19) i (20) możemy wywnioskować, że największa liczba osób hospitalizowanych była w województwie łódzkim, zaś osób w ciężkim stanie w lubuskim. Najmniej osób hospitalizowanych, podczas trzeciej fali pandemii, spodziewamy się, że było w województwie warmińsko-mazurskim, zaś w stanie ciężkim w województwie pomorskim.

5.2.2 Model Function-on-Function

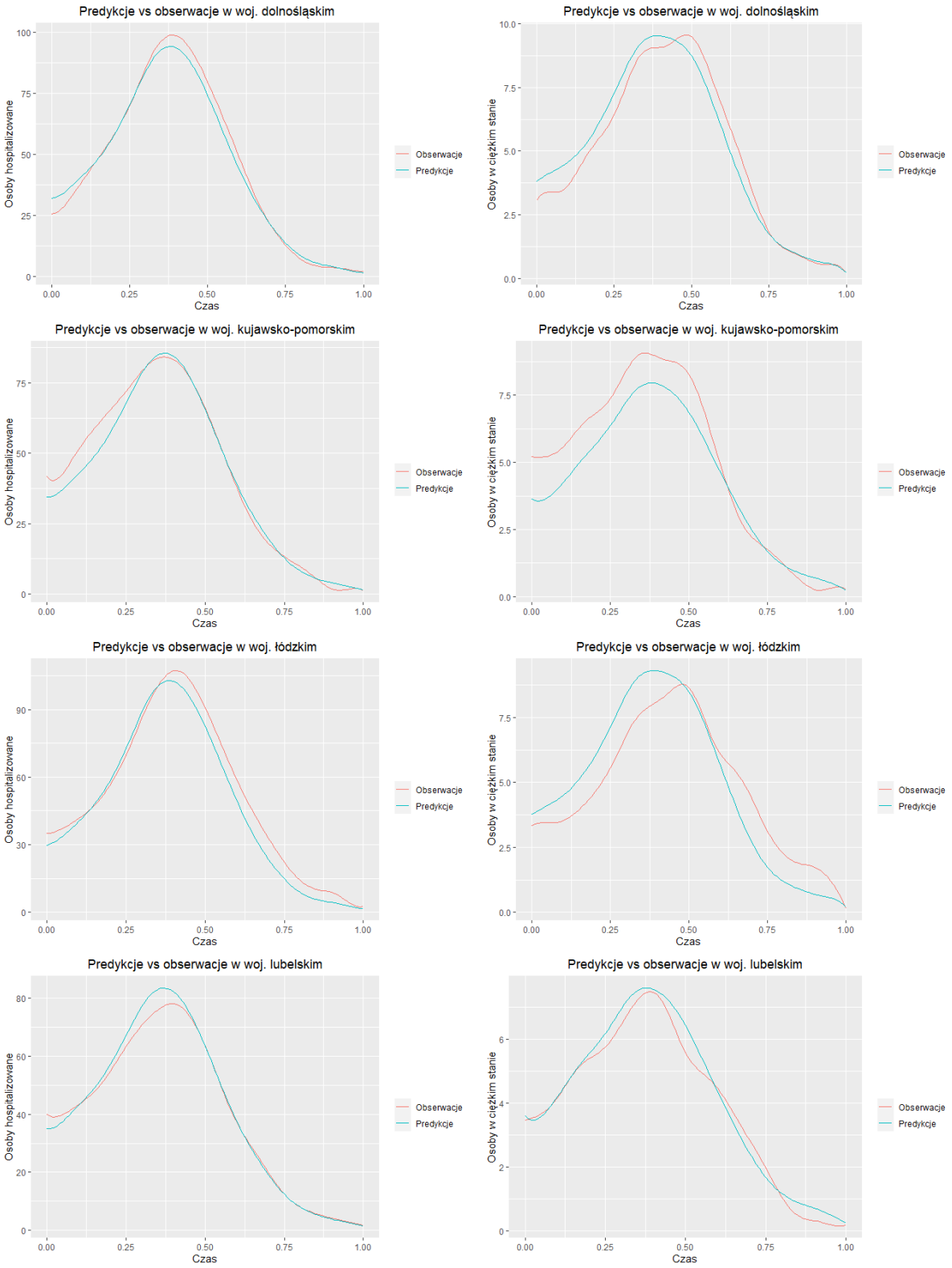
W poniższej tabeli (5) widzimy wartości MSE, obliczone za pomocą wzoru (15), zaś rysunek (21) przedstawia obserwowane i przewidywane krzywe dla kilku województw z próby treningowej. Rysunek (22) oraz tabela (6) ilustrują predykcje na próbie testowej.

Tabela 5: Wartości błędu średniokwadratowego dla y_{i1} (liczby osób hospitalizowanych) i y_{i2} (liczby osób w stanie ciężkim) dla województw z próby treningowej.

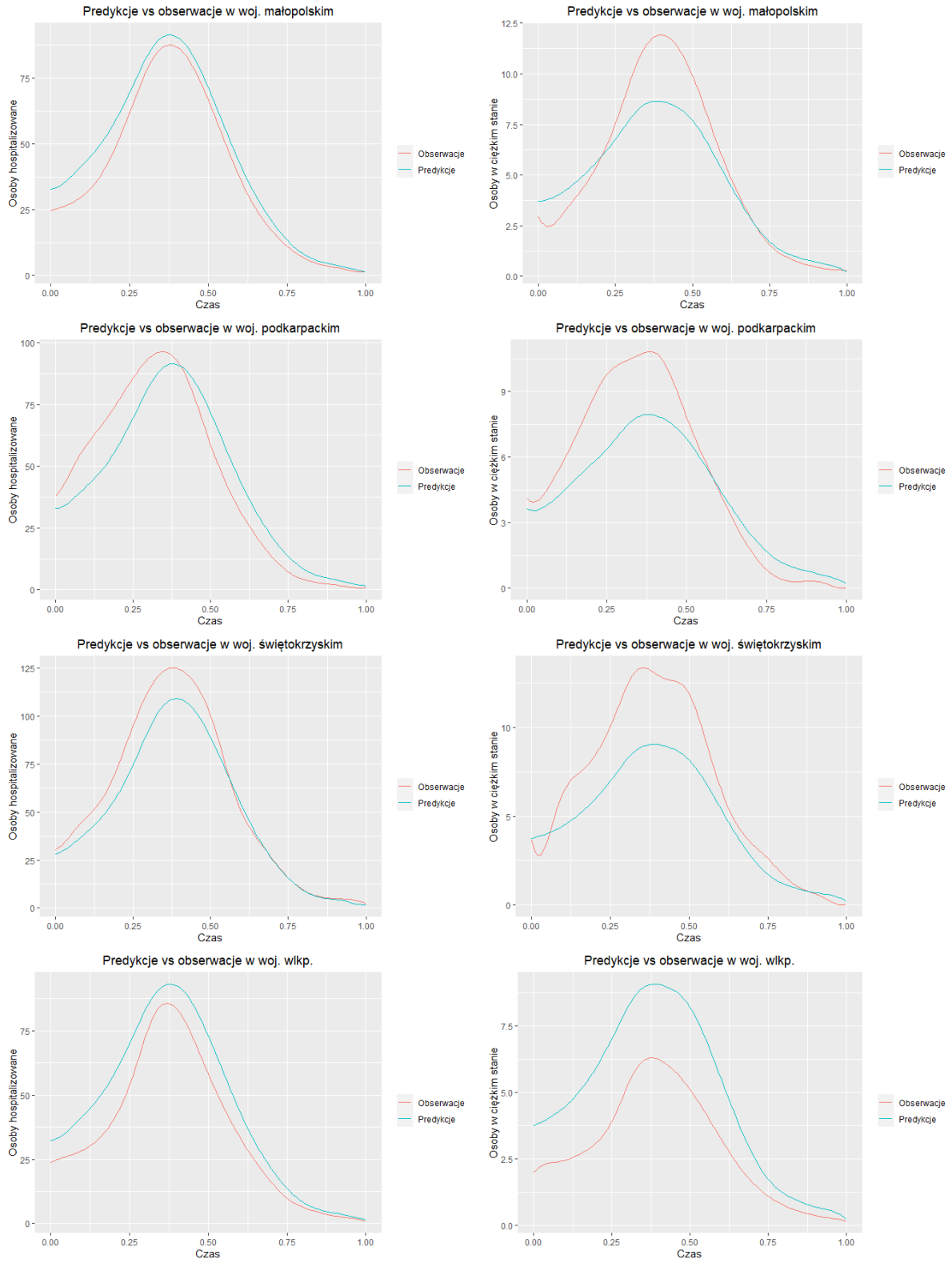
województwo	MSE(y_{i1})	MSE(y_{i2})
dolnośląskie	3.305934	0.6009604
kujawsko-pomorskie	4.043124	0.9548898
łódzkie	5.810232	1.0920511
lubelskie	**2.735627**	**0.3562136**
lubuskie	**10.226623**	1.0394366
mazowieckie	3.535219	0.5104856
opolskie	9.569692	1.1204369
podlaskie	6.109919	0.8874620
pomorskie	2.959356	1.0975004
śląskie	4.493135	0.8706739
warmińsko-mazurskie	4.586885	**1.5028614**
zachodnio-pomorskie	8.482927	0.7540102

W powyższej tabeli wytłuszczono odpowiednio najniższe oraz najwyższe wartości błędu średniokwadratowego.

Najniższe wartości MSE(y_{i1}) i MSE(y_{i2}) są przyjmowane dla województwa lubelskiego. Z kolei najwyższą wartość MSE(y_{i1}) otrzymano dla województwa lubuskiego, a MSE(y_{i2}) dla warmińsko-mazurskiego. W porównaniu z MSE(y_{i1}) otrzymanym podczas analizy danych z obu fal razem, wartość MSE(y_{i1}) spadła dla prawie wszystkich województw (wyjątek: opolskie, warmińsko-mazurskie oraz zachodnio-pomorskie). Podobnie MSE(y_{i2}), dla którego wartość wzrosła jedynie w województwach: łódzkim, lubuskim, opolskim oraz warmińsko-mazurskim.



Rysunek 21: Obserwowane i przewidywane krzywe dla próby treningowej dla trzeciej fali pandemii.



Rysunek 22: Obserwowane i przewidywane krzywe dla próby testowej dla trzeciej fali pandemii.

Tabela 6: Przewidywane wartości liczby osób hospitalizowanych w porównaniu z obserwowanymi wartościami podawanymi przez Ministerstwo Zdrowia w trakcie trzeciej fali.

Osoby hospitalizowane				
	małopolskie	podkarpackie	świętokrzyskie	wielkopolskie
czas	obs/pred	obs/pred	obs/pred	obs/pred
16.02	857/1120	790/697	377/344	837/1132
17.02	831/1124	833/699	375/349	837/1137
18.02	857/1132	872/704	391/355	845/1146
19.02	874/1144	888/712	391/362	868/1160
20.02	884/1161	906/722	417/370	896/1177
21.02	884/1181	925/734	433/379	892/1198
22.02	897/1204	956/749	464/389	923/1222
23.02	939/1229	990/765	467/400	926/1249
24.02	910/1257	1073/782	482/411	925/1277
25.02	956/1285	1108/799	494/423	916/1307
...	...	...	...	...
247	72/108	31/67	69/43	81/112
248	72/101	26/63	61/40	79/105
249	66/94	25/59	57/36	82/97
250	63/88	20/54	54/32	82/90
251	57/81	15/50	51/29	63/83
252	53/74	13/46	47/26	42/76
253	51/68	13/42	39/23	44/69
254	53/62	13/38	36/21	44/63
255	48/56	15/35	32/20	44/57
256	41/51	14/32	33/19	39/52
Osoby w stanie ciężkim				
16.02	99/126	84/77	40/46	71/131
17.02	100/127	83/76	35/46	72/132
18.02	81/127	88/76	40/47	74/134
19.02	85/128	87/75	40/47	75/135
20.02	86/129	87/76	41/48	81/137
21.02	89/130	85/76	41/48	78/138
22.02	86/132	93/77	41/49	81/140
23.02	87/133	83/78	46/49	85/141
24.02	88/135	88/79	40/50	85/143
25.02	88/137	96/81	49/51	85/145
...	...	...	...	...
26.06	11/20	5/13	4/7	11/21
27.06	11/19	4/12	3/7	10/20
28.06	14/18	3/11	2/7	9/19
29.06	12/17	2/10	2/6	9/18
30.06	11/16	2/10	2/6	9/17
1.07	11/15	0/9	1/6	7/16
2.07	11/13	0/8	0/5	7/14
3.07	11/12	0/7	0/4	7/12
4.07	11/10	1/6	0/4	7/10
5.07	10/7	0/5	0/3	5/7

Porównując, przedstawioną w tabeli (6), przewidywaną liczbę osób w stanie ciężkim dla trzeciej fali pandemii z tabelą (2) widzimy, że wartości nie zmieniły się znacząco. Predykcje dla województwa małopolskiego delikatnie poprawiły się, jednakże dla województwa wielkopolskiego nieznacznie uległy pogorszeniu. Województwo podkarpackie i świętokrzyskie uzyskało bardzo podobne rezultaty, co wcześniej.

Porównując liczbę osób hospitalizowanych widzimy, że wyniki są podobne. Można jednak zauważyć, że model otrzymany na podstawie danych jedynie z trzeciej fali przewiduje większą liczbę osób hospitalizowanych podczas trzeciej fali pandemii niż model otrzymany z wykorzystaniem połączonych danych z obu fal.

6 Wnioski i podsumowanie

Pandemia COVID-19 wstrząsnęła całym światem. Epidemia na taką skalę jest stosunkowo nowym zjawiskiem, dlatego przyciągnęła uwagę sporej liczby analityków na całym świecie. W przedstawionej pracy podjęto próbę dopasowania modelu opartego na funkcjonalnej analizie danych do danych dotyczących drugiej i trzeciej fali pandemii COVID-19 na terenie Polski. Celem tak skonstruowanego modelu była predykcja liczby osób w stanie ciężkim oraz osób hospitalizowanych. Jako predyktory posłużyły: liczba zgonów, ozdrowieńców oraz pozytywnych wyników testu. Na grupie treningowej złożonej z dwunastu województw dokonano estymacji parametrów modelu oraz przetestowano jego jakość. Następnie dokonano predykcji dla województw: małopolskiego, podkarpackiego, świętokrzyskiego oraz wielkopolskiego. Model bardzo dobrze przewidywał liczby w większości województw. Największym problemem okazało się województwo małopolskie. Zarówno w przypadku dopasowywania modelu do obu fal razem, do pojedynczej drugiej lub trzeciej fali, liczba przewidywanych osób hospitalizowanych i osób w stanie ciężkim dla tego województwa okazywała się zbyt niska. Model świetnie poradził sobie natomiast z województwem wielkopolskim i świętokrzyskim.

Rozdzielenie fal nie przyniosło dużych efektów podczas porównywania przewidywanych wartości i wartości obserwowanych (tabele 2, 4, 6) w grupie testowej. Jednakże dokonując analizy wartości błędów średniokwadratowych, policzonych na podstawie próby treningowej (tabele 1, 3, 5), można wyciągnąć ciekawe wnioski na temat województw. Analizując MSE otrzymane dla liczby osób hospitalizowanych najniższe wartości otrzymujemy podczas trzeciej fali dla województw: dolnośląskiego, kujawsko-pomorskiego, lubelskiego, lubuskiego, mazowieckiego oraz pomorskiego. Podczas drugiej fali: łódzkiego, podlaskiego oraz śląskiego. Co ciekawe w województwach opolskim, warmińsko-mazurskim oraz zachodnio-pomorskim najniższe MSE otrzymaliśmy podczas analizy obu fal razem.

Z kolei błąd średniokwadratowy dla liczby osób w stanie ciężkim jest bardziej ujednolicony. Dla wszystkich województw z wyjątkiem lubelskiego i podlaskiego druga fala miała najniższe wartości MSE.

Interesująca okazała się analiza głównych składowych. Rysunki (8), (9), (15) i (20) przyniosły ciekawe informacje na temat województw z grupy treningowej. Podczas drugiej i trzeciej fali pandemii najwięcej osób hospitalizowanych było w województwach lubuskim i podlaskim, podczas pojedynczej drugiej fali w lubelskim i podlaskim, zaś podczas trzeciej fali w łódzkim i śląskim. Najwięcej osób w stanie ciężkim podczas obu fal razem było w województwach kujawsko-pomorskim i lubuskim, podczas drugiej fali w lubelskim i kujawsko-pomorskim, zaś podczas trzeciej fali w lubuskim i mazowieckim. Przeprowadzona analiza zgadza się z prawdziwymi, dyskretnymi obserwacjami.

Największa różnica pomiędzy drugą i trzecią falą w liczbie osób hospitalizowanych była w województwie łódzkim i śląskim, największa różnica pomiędzy początkiem i końcem drugiej fali w województwach kujawsko-pomorskim i warmińsko-mazurskim, zaś pomiędzy początkiem i końcem trzeciej fali w lubuskim i podlaskim.

Z kolei dla liczby osób w stanie ciężkim podobna analiza przedstawia się następująco. Największa różnica pomiędzy drugą i trzecią falą była w województwach podlaskim i warmińsko-mazurskim, największa różnica pomiędzy początkiem i końcem drugiej fali w województwach

lubelskim (jest to również województwo z największą liczbą przypadków osób w stanie ciężkim) i śląskim, zaś pomiędzy początkiem i końcem trzeciej fali w warmińsko-mazurskim i podlaskim.

Co również jest ciekawe, prawie we wszystkich modelach, pierwsza składowa główna opisywała ogólną tendencję oraz liczbę przypadków osób w różnych stanach. Wyjątkiem jest model wyjaśniający liczbę osób hospitalizowanych w obu falach. Tutaj pierwsza główna składowa zwracała uwagę na różnice pomiędzy dwoma analizowanymi falami.

Ekonomiczny kryzys wywołany przez wirusa SARS-CoV-2 objął niemal całą planetę, od kiedy Światowa Organizacja Zdrowia (WHO) ogłosiła stan wyjątkowy w połowie marca 2020 roku. Chcąc w pewnej części kontrolować wirusa, społeczność naukowa jest pochłonięta opracowywaniem modeli, które pozwoliłyby na łagodzenie niszczycielskich skutków pandemii. Ważne jest zatem budowanie skutecznych modeli w celu zagwarantowania poprawnych predykcji. Biorąc pod uwagę naturę zmiennych - liczby przypadków, zgonów, ozdrowień, hospitalizacji oraz osób w stanie ciężkim, warto wziąć pod uwagę metody analizy danych funkcjonalnych. Niemniej jednak konstruując modele, ważna jest jakość danych, które w przypadku rozwijającej się pandemii nie są zbyt dokładne. W tej pracy zaproponowano model MFFLR - *multiple function-on-function linear regression model* w celu imputacji brakujących danych. Motywacją tego była prognoza krzywych osób hospitalizowanych i osób w stanie ciężkim z krzywych pozytywnych przypadków, zgonów oraz wyzdrowień.

7 Dodatek - kod programu R

W tym rozdziale został przedstawiony kod umożliwiający analizę przedstawioną w pracy.

```
# instalacja pakietów:
library('ggplot2')
library('fda')

# przykładowy wykres obserwacji dyskretnych dla liczby osób
# hospitalizowanych:

ggplot(hospitalizacje, mapping = aes(x = daty,
y = hospitalizowani)) +
  geom_point(aes(x=daty, y=kujawskopom.hospitalizacje,
  col="kujawsko-pomorskie")) +
  geom_point(aes(x=daty, y=malopolskie.hospitalizacje,
  col="małopolskie")) +
  geom_point(aes(x=daty, y=podkarpackie.hospitalizacje,
  col="podkarpackie")) +
  geom_point(aes(x=daty, y=podlaskie.hospitalizacje,
  col="podlaskie")) +
  geom_point(aes(x=daty, y=pomorskie.hospitalizacje,
  col="pomorskie")) +
  geom_point(aes(x=daty, y=swietokrzyskie.hospitalizacje,
  col="świętokrzyskie")) +
  geom_point(aes(x=daty, y=wielkopolskie.hospitalizacje,
  col="wielkopolskie")) +
  scale_color_discrete(name="Województwa")

# I CZĘŚĆ: WYGŁADZENIE DANYCH

nbas = 20 # wybór liczby funkcji bazowych
# utworzenie bazy funkcji bazowych B-splines:
Basis20 = create.bspline.basis(rangeval = c(0,1), nbasis=nbas)

# wylistowanie przypadków;
# nie bierzemy pod uwagę kolumny odpowiedzialnej za daty (17):
indywidualne.przypadki.fd = as.list(przypadki[-17])
indywidualne.hospitalizacje.fd = as.list(hospitalizacje[-17])
indywidualne.stanyciezkie.fd = as.list(stan_ciezki[-17])
indywidualne.zgony.fd = as.list(zgony[-17])
indywidualne.wyzdrowienia.fd = as.list(wyzdrowienia[-17])

# konwersja danych w obiekty funkcjonalne
for (i in 1:16){
  indywidualne.przypadki.fd[[i]] = Data2fd(argvals =seq(0,1,
```

```
  length=length(indywidualne.przypadki.fd[[i]])),
  y=as.matrix(indywidualne.przypadki.fd[[i]]),basisobj = Basis20)

  indywidualne.hospitalizacje.fd[[i]] = Data2fd(argvals =seq(0,1,
  length=length(indywidualne.hospitalizacje.fd[[i]])),
  y=as.matrix(indywidualne.hospitalizacje.fd[[i]]),
  basisobj = Basis20)

  indywidualne.stanyciezkie.fd[[i]] = Data2fd(argvals =seq(0,1,
  length=length(indywidualne.stanyciezkie.fd[[i]])),
  y=as.matrix(indywidualne.stanyciezkie.fd[[i]]),basisobj = Basis20)

  indywidualne.zgony.fd[[i]] = Data2fd(argvals =seq(0,1,
  length=length(indywidualne.zgony.fd[[i]])),
  y=as.matrix(indywidualne.zgony.fd[[i]]),basisobj = Basis20)

  indywidualne.wyzdrowienia.fd[[i]] = Data2fd(argvals =seq(0,1,
  length=length(indywidualne.wyzdrowienia.fd[[i]])),
  y=as.matrix(indywidualne.wyzdrowienia.fd[[i]]),basisobj = Basis20)
}

wspolcz.przypadki = indywidualne.przypadki.fd[[1]]$coefs
wspolcz.hospitalizacje = indywidualne.hospitalizacje.fd[[1]]$coefs
wspolcz.stanciezki = indywidualne.stanyciezkie.fd[[1]]$coefs
wspolcz.zgony = indywidualne.zgony.fd[[1]]$coefs
wspolcz.ozdrowienia = indywidualne.wyzdrowienia.fd[[1]]$coefs

for (i in 2:16){
  wspolcz.przypadki = cbind(wspolcz.przypadki,
  indywidualne.przypadki.fd[[i]]$coefs)
  wspolcz.hospitalizacje = cbind(wspolcz.hospitalizacje,
  indywidualne.hospitalizacje.fd[[i]]$coefs)
  wspolcz.stanciezki = cbind(wspolcz.stanciezki,
  indywidualne.stanyciezkie.fd[[i]]$coefs)
  wspolcz.zgony = cbind(wspolcz.zgony,
  indywidualne.zgony.fd[[i]]$coefs)
  wspolcz.ozdrowienia = cbind(wspolcz.ozdrowienia,
  indywidualne.wyzdrowienia.fd[[i]]$coefs)
}

przypadki.fd = fd(wspolcz.przypadki,basisobj = Basis20)
hospitalizacje.fd = fd(wspolcz.hospitalizacje,basisobj = Basis20)
stanciezki.fd = fd(wspolcz.stanciezki,basisobj = Basis20)
zgony.fd = fd(wspolcz.zgony,basisobj = Basis20)
ozdrowiency.fd = fd(wspolcz.ozdrowienia,basisobj = Basis20)
```

```
przypadki.fd.predykcja = data.frame(predict(przypadki.fd,
seq(0,1,0.01)))
hospitalizacje.fd.predykcja = data.frame(predict(hospitalizacje.fd,
seq(0,1,0.01)))
stanciezki.fd.predykcja = data.frame(predict(stanciezki.fd,
seq(0,1,0.01)))
zgony.fd.predykcja = data.frame(predict(zgony.fd,
seq(0,1,0.01)))
ozdrowienia.fd.predykcja = data.frame(predict(ozdrowiency.fd,
seq(0,1,0.01)))

wojewodztwa = c("kujawsko-pomorskie", "małopolskie", "podkarpackie",
"podlaskie", "pomorskie", "świętokrzyskie","wielkopolskie")

# przykładowy wykres dopasowanych krzywych dla l. przypadków

ggplot(, aes(x=seq(0,1,0.01))) +
  geom_line(aes(, y=przypadki.fd.predykcja[,2],
  colour=wojewodztwa[1])) +
  geom_line(aes(, y=przypadki.fd.predykcja[,6],
  colour=wojewodztwa[2])) +
  geom_line(aes(, y=przypadki.fd.predykcja[,9],
  colour=wojewodztwa[3])) +
  geom_line(aes(, y=przypadki.fd.predykcja[,10],
  colour=wojewodztwa[4])) +
  geom_line(aes(, y=przypadki.fd.predykcja[,11],
  colour=wojewodztwa[5])) +
  geom_line(aes(, y=przypadki.fd.predykcja[,13],
  colour=wojewodztwa[6])) +
  geom_line(aes(, y=przypadki.fd.predykcja[,15],
  colour=wojewodztwa[7])) +
  xlab("Czas")+
  ylab("Przypadki")+
  scale_color_discrete(name="Województwa")+
  theme(plot.title = element_text(hjust = 0.5))

# II CZĘŚĆ: FUNKCJONALNA ANALIZA GŁÓWNYCH SKŁADOWYCH

# FPCA dla grupy treningowej

# liczba głównych składowych nharm=10
przypadki.fpca = pca.fd(przypadki.fd,nharm=10)
hospitalizacje.fpca=pca.fd(hospitalizacje.fd[c(1:5,7:8,10:12,14,16)],
nharm=10)
```

```
stanciezki.fpca = pca.fd(stanciezki.fd[c(1:5,7:8,10:12,14,16)],
nharm=10)
zgony.fpca = pca.fd(zgony.fd,nharm=10)
ozdrowiency.fpca = pca.fd(ozdrowiency.fd,nharm=10)

przypadki.fpca$varprop # procent wyjaśnianej wariancji
hospitalizacje.fpca$varprop
stanciezki.fpca$varprop
zgony.fpca$varprop
ozdrowiency.fpca$varprop

# Przykładowy wykres harmonicznych - składowych głównych

ggplot(, aes(x=seq(0,1,0.01))) +
  geom_line(aes(, y=eval.fd(przypadki.fpca$harmonics[1],
  seq(0,1,0.01)),col="PCA 1 (46.93%)")) +
  geom_line(aes(, y=eval.fd(przypadki.fpca$harmonics[2],
  seq(0,1,0.01)), col="PCA 2 (26.59%)")) +
  geom_line(aes(, y=eval.fd(przypadki.fpca$harmonics[3],
  seq(0,1,0.01)), col="PCA 3 (13.96%)")) +
  xlab("Czas")+
  ylab("Współczynniki")+
  ggtitle("Główne składowe dla liczby przypadków")+
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_color_discrete(name="Główne składowe")

# Wykresy średniej +/- składowa główna

plot.pca.fd(przypadki.fpca)

# Wykresy 1 vs 2 score

plot(przypadki.fpca$scores[,1], przypadki.fpca$scores[,2],
xlab="pierwszy score", ylab="drugi score")

wojewodztwa = c("dolnośląskie", "kujawsko-pomorskie","łódzkie",
"lubelskie", "lubuskie", "mazowieckie","opolskie",
"podlaskie","pomorskie","śląskie", "warmińsko-mazurskie",
"zachodnio-pomorskie")

text(przypadki.fpca$scores[,1], przypadki.fpca$scores[,2],
wojewodztwa, col=3)

# Wykres dla średniej +/- składowa główna
```

```
ggplot(, aes(x=seq(0,1,0.01))) +
  geom_line(aes(, y=eval.fd(przypadki.fpca$meanfd,seq(0,1,0.01)))) +
  geom_line(aes(, y=eval.fd(przypadki.fpca$meanfd-
  sqrt(przypadki.fpca$values[1])*przypadki.fpca$harmonics[1],
  seq(0,1,0.01))),linetype="dashed") +
  geom_line(aes(, y=eval.fd(przypadki.fpca$meanfd+
  sqrt(przypadki.fpca$values[1])*przypadki.fpca$harmonics[1],
  seq(0,1,0.01))),linetype="dashed") +
  xlab("Czas")+
  ylab("Krzywa + przedział")+
  ggtitle("Krzywa oraz przedział dla l. przypadków")+
  theme(plot.title = element_text(hjust = 0.5))

# CZĘŚĆ III: PREDYKCJA

# Macierz "scores" dla próby treningowej:

Macierzwynikow = cbind(stanciezki.fpca$scores[,1:3],
przypadki.fpca$scores[c(1:5,7:8,10:12,14,16),1:3],
hospitalizacje.fpca$scores[,1:3],
zgony.fpca$scores[c(1:5,7:8,10:12,14,16),1:3],
ozdrowiency.fpca$scores[c(1:5,7:8,10:12,14,16),1:3])

colnames(Macierzwynikow)=c("stanciezki1","stanciezki2","stanciezki3",
"przypadki1","przypadki2","przypadki3",
"hospi1","hospi2","hospi3",
"zgony1","zgony2","zgony3",
"ozdrow1","ozdrow2","ozdrow3")

Macierzwynikow = data.frame(Macierzwynikow)

# Dopasowanie modeli

Model.stanciezki = lm(stanciezki1~przypadki1+zgony1+ozdrow1,
data=Macierzwynikow)

Model.hospi = lm(hospi1~przypadki1+zgony1+ozdrow1,
data=Macierzwynikow)

# Macierz planu dla próby testowej (osoby w stanie ciężkim)

Macierz_stanciezki = matrix(c(1,1,1,1,
przypadki.fpca$scores[c(6,9,13,15),1],
zgony.fpca$scores[c(6,9,13,15),1],
ozdrowiency.fpca$scores[c(6,9,13,15),1]),4,)
```

```
# Predykcja osób w stanie ciężkim dla próby testowej

stanciezki_estymacja = fd(matrix(rep(mean.fd(stanciezki.fd[c(1:5,
7:8,10:12,14,16)])$coefs,4),nbas)+
t(Macierz_stanciezki%*%Model.stanciezki$coefficients%*%
stanciezki.fpca$harmonics$coefs[,1]),basisobj = Basis20)

# Przykładowy wykres predykcji vs obserwacji w grupie testowej

stanciezki_wykres = eval.fd(stanciezki_estymacja,seq(0,1,0.01))
ggplot(, aes(x=seq(0,1,0.01))) +
geom_line(aes(,y=stanciezki.fd.predykcja[,6],colour="Obserwowane")) +
  geom_line(aes(, y=stanciezki_wykres[,1],colour="Przewidywane")) +
  xlab("Czas")+
  ylab("Osoby w stanie ciężkim")+
  ggtitle("Predykcje vs obserwacje w woj. małopolskim")+
  theme(plot.title = element_text(hjust = 0.5))

# Macierz planu dla próby treningowej
# (osoby w stanie ciężkim)

Macierz_stanciezki_treningowa = matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,
przypadki.fpca$scores[c(1:5,7:8,10:12,14,16),1],
zgony.fpca$scores[c(1:5,7:8,10:12,14,16),1],
ozdrowiency.fpca$scores[c(1:5,7:8,10:12,14,16),1]),12,)

# Predykcja osób w stanie ciężkim dla próby treningowej

stanciezki_estymacja_treningowa = fd(matrix(rep(mean.fd(
stanciezki.fd[c(1:5,7:8,10:12,14,16)])$coefs,12),
nbas)+t(Macierz_stanciezki_treningowa%*%
Model.stanciezki$coefficients%*%stanciezki.fpca$harmonics$coefs[,1]),
basisobj = Basis20)

# Obliczenie błędu średniokwadratowego

sqrt(diag(t(stanciezki_estymacja_treningowa$coefs-
stanciezki.fd$coefs[,c(1:5,7:8,10:12,14,16)])%*%
inprod(Basis20,Basis20)%*%(stanciezki_estymacja_treningowa$coefs-
stanciezki.fd$coefs[,c(1:5,7:8,10:12,14,16)])))

# Przykładowy wykres predykcji vs obserwacji w grupie treningowej

stanciezki_wykres_treningowa=eval.fd(stanciezki_estymacja_treningowa,
```

```
seq(0,1,0.01))

ggplot(, aes(x=seq(0,1,0.01))) +
geom_line(aes(,y=stanciezki.fd.predykcja[,1],colour="Obserwowane")) +
  geom_line(aes(, y=stanciezki_wykres_treningowa[,1],
  colour="Przewidywane")) +
  xlab("Czas")+
  ylab("Osoby w stanie ciężkim")+
  ggtitle("Predykcje vs obserwacje w woj. dolnośląskim")+
  theme(plot.title = element_text(hjust = 0.5))

# Macierz planu dla próby testowej - osoby hospitalizowane

Macierz_hospitalizacje = matrix(c(1,1,1,1,
przypadki.fpca$scores[c(6,9,13,15),1],
zgony.fpca$scores[c(6,9,13,15),1],
ozdrowiency.fpca$scores[c(6,9,13,15),1]),4,)

# Predykcja osób hospitalizowanych dla próby testowej

Hospitalizacje_estymacja = fd(matrix(rep(mean.fd(
hospitalizacje.fd[c(1:5,7:8,10:12,14,16)])$coefs,4),
nbas)+t(Macierz_hospitalizacje%*%Model.hospi$coefficients%*%
hospitalizacje.fpca$harmonics$coefs[,1]),basisobj = Basis20)

# Przykładowy wykres predykcji vs obserwacji osób hospitalizowanych

hospitalizacje_wykres=eval.fd(Hospitalizacje_estymacja,seq(0,1,0.01))
ggplot(, aes(x=seq(0,1,0.01))) +
  geom_line(aes(, y=hospitalizacje.fd.predykcja[,6],
  colour="Obserwowane")) +
  geom_line(aes(, y=hospitalizacje_wykres[,1],
  colour="Przewidywane")) +
  xlab("Czas")+
  ylab("Osoby hospitalizowane")+
  ggtitle("Predykcje vs obserwacje w woj. małopolskim")+
  theme(plot.title = element_text(hjust = 0.5))

# Macierz planu dla próby treningowej, osoby hospitalizowane

Macierz_hospitalizacje_treningowa = matrix(c(1,1,1,1,1,1,1,1,1,1,1,1,
przypadki.fpca$scores[c(1:5,7:8,10:12,14,16),1],
zgony.fpca$scores[c(1:5,7:8,10:12,14,16),1],
ozdrowiency.fpca$scores[c(1:5,7:8,10:12,14,16),1]),12,)
```

```
# Predykcja osób hospitalizowanych dla próby treningowej

Hospitalizacje_estymacja_treningowa = fd(matrix(rep(mean.fd(
hospitalizacje.fd[c(1:5,7:8,10:12,14,16)])$coefs,12),nbas)
+t(Macierz_hospitalizacje_treningowa%*%Model.hospi$coefficients%*%
hospitalizacje.fpca$harmonics$coefs[,1]),basisobj = Basis20)

# Obliczenie błędu średniokwadratowego

sqrt(diag(t(Hospitalizacje_estymacja_treningowa$coefs-
hospitalizacje.fd$coefs[,c(1:5,7:8,10:12,14,16)])%*%inprod(Basis20,
Basis20)%*%(Hospitalizacje_estymacja_treningowa$coefs-
hospitalizacje.fd$coefs[,c(1:5,7:8,10:12,14,16)])))

# Przykładowy wykres predykcji vs obserwacji w grupie treningowej

hospi_wykres_treningowa=eval.fd(Hospitalizacje_estymacja_treningowa,
seq(0,1,0.01))
ggplot(, aes(x=seq(0,1,0.01))) +
  geom_line(aes(, y=hospitalizacje.fd.predykcja[,1],
  colour="Obserwowane")) +
  geom_line(aes(, y=hospi_wykres_treningowa[,1],
  colour="Przewidywane")) +
  xlab("Czas")+
  ylab("Osoby hospitalizowane")+
  ggtitle("Predykcje vs obserwacje w woj. dolnośląskim")+
  theme(plot.title = element_text(hjust = 0.5))
```

Literatura

[1] Autor zbioru danych: Michał Rogalski, *COVID-19 w Polsce*, Źródło: 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. Kontakt: contact.micalrg@gmail.com; bit.ly/covid19-poland; dostęp: 01.03.2022.

[2] Główny Urząd Statystyczny, *Powierzchnia i ludność w przekroju terytorialnym w 2021 roku*, Powierzchnia, liczba ludności i gęstość zaludnienia wg stanu na 1 stycznia 2021 roku. Przekroje: województwa, powiaty, gminy, miasta. Data publikacji: 22.07.2021. Dostęp: 07.03.2022. Link: https://stat.gov.pl/obszary-tematyczne/ludnosc/ludnosc/powierzchnia-i-ludnosc-w-przekroju-terytorialnym-w-2021-roku,7,18.html.

[3] Christian Acal, Manuel Escabias, Ana M. Aguilera, Mariano J. Valderrama, *COVID-19 Data Imputation by Multiple Function-on-Function Principal Component Regression*, Mathematics 2021, 9, 1237. https:// doi.org/10.3390/math9111237, Academic Editor: Jin-Ting Zhang, 28 May 2021.

[4] Jaroslaw Harezlak, David Ruppert, Matt P. Wand, *Semiparametric Regression with R, Use R*, Springer, New York, NY, Edition Number 1, 2008, https://doi.org/10.1007/978-1-4939-8853-2.

[5] Shahid Ullah, Caroline F Finch, *Applications of functional data analysis: A systematic review*, BMC Medical Research Methodology 2013, http://www.biomedcentral.com/1471-2288/13/43.

[6] De Boor, C. *A practical guide to splines*, 2001, revised edition, Berlin: Springer-Verlag.

[7] Shikin, E. V., Plis, A. I. *Handbook on splines for the user*, Boca Raton, FL: CRC Press, 1995.

[8] Unser, M. *Splines: A perfect fit for signal and image processing. IEEE Signal Processing,* 16, 22-38, 1999.

[9] Daniel Levitin, Bradley W. Vines, Regina L. Nuzzo, James O. Ramsay, *Introduction to Functional Data Analysis,* Canadian Psychology, Copyright 2007 by the Canadian Psychological Association, August 2007, Vol. 48, No. 3, 135-155, DOI: 10.1037/cp2007014.

[10] Trevor Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning*, Data Mining, Inference, and Prediction, Second Edition, Springer Series in Statistics, Stanford, California, August 2008.

[11] Jane-Ling Wang, Jeng-Min Chiou, Hans-Georg Müller, *Review of functional data analysis,* Annu. Rev. Statist. 2015, doi: 10.1146, Copyright 2015 by Annual Reviews.

[12] Joseph Rickert, *Functional PCA with R*, 10.06.2021, https://rviews.rstudio.com/2021/06/10/functional-pca-with-r/ [dostęp: 21.04.2022].

[13] Acal, C.; Aguilera, A.M.; Escabias, *M. New Modeling Approaches Based on Varimax Rotation of Functional Principal Components,* Mathematics 2020, 8, 2085.

[14] Aguilera, A.M., Escabias, M., Ocaña, F.A. *Functional Wavelet-Based Modelling of Dependence Between Lupus and Stress,* Methodol Comput Appl Probab 17, 1015–1028 (2015), https://doi.org/10.1007/s11009-014-9424-5.

[15] Florence Nicol, *Functional Principal Component Analysis of Aircraft Trajectories.* [Research Report], RR/ENAC/2013/02, ENAC. 2013. ffhal-01349113f.

[16] Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, R package version 3.3.0, published: 2022-05-03, URL: https://CRAN.R-project.org/package=ggplot2.

[17] J. O. Ramsay, Hadley Wickham, Spencer Graves, Giles Hooker, *fda: Functional Data Analysis in R*, R package version 5.1.7, published: 2022-04-26, URL: https://CRAN.R-project.org/package=fda.