WIADOMOŚCI STATYSTYCZNE

PL ISSN 0043-518X
e-ISSN 2543-8476

Rok LXII
5 (672) 2017, 19—37

STUDIA METODOLOGICZNE

Karol DERĘGOWSKI
Mirosław KRZYŚKO
Łukasz WASZAK
Waldemar WOŁYŃSKI

Zastosowanie funkcjonalnej analizy kanonicznej w badaniu zależności między wydatkami konsumpcyjnymi w europejskich gospodarstwach domowych

Streszczenie. *Celem artykułu jest zbadanie zależności między wydatkami na napoje alkoholowe i wyroby tytoniowe a pozostałymi wydatkami konsumpcyjnymi gospodarstw domowych w 27 krajach Europy w latach 2000—2010. Wybór krajów i przedziału czasowego podyktowany został dostępnością i kompletnością danych pochodzących z Eurostatu. Nowością zaprezentowaną w tej pracy jest rozpatrywanie lat łącznie, a nie każdego roku oddzielnie. Stało się to możliwe dzięki przekształceniu danych pierwotnych na wielowymiarowe dane funkcjonalne oraz skonstruowaniu korelacji i zmiennych kanonicznych dla danych przekształconych.*

Z badania wynika, że wydatki na napoje alkoholowe i wyroby tytoniowe są bardzo silnie skorelowane z pozostałymi wydatkami konsumpcyjnymi (współczynnik korelacji kanonicznej między dwiema pierwszymi funkcjonalnymi zmiennymi kanonicznymi wynosi 0,99). Wydatki na napoje alkoholowe i wyroby tytoniowe mają prawie jednakowy wkład w budowę funkcjonalnej zmiennej kanonicznej U_1, natomiast największy udział w budowie funkcjonalnej zmiennej kanonicznej V_1 przypada wydatkom na artykuły żywnościowe i napoje bezalkoholowe oraz wydatkom na odzież i obuwie.

Słowa kluczowe: analiza kanoniczna, wielowymiarowa funkcjonalna analiza danych, spożycie indywidualne według celu.

JEL: C18, C22, C38

Inspiracją do przygotowania tego opracowania jest artykuł Marleny Piekut (2016). W pracy podjęto się empirycznego zbadania zależności między indywidualnymi wydatkami gospodarstw domowych na napoje alkoholowe i wyroby tytoniowe a pozostałymi artykułami konsumpcyjnymi gospodarstw domowych w wybranych krajach europejskich w latach 2000—2010. Nowością w naszej pracy jest rozpatrywanie okresu badawczego łącznie, a nie każdego roku oddzielnie. Można to było osiągnąć po przekształceniu oryginalnych danych w postaci szeregów czasowych na wektorowe funkcje ciągłe, określone na ustalonym przedziale czasowym, zwane wielowymiarowymi danymi funkcjonalnymi.

W klasycznych metodach statystycznych obiekty podlegające badaniu charakteryzowane są za pomocą cech obserwowanych w ustalonym momencie. Tutaj zakładamy, że poszczególne kraje charakteryzowane są za pomocą zmiennych funkcjonalnych. Czym są zmienne funkcjonalne? Zmienna funkcjonalna X jest zmienną losową przyjmującą wartości w pewnej przestrzeni funkcjonalnej E. Zbiór danych funkcjonalnych jest próbką $\{X_1, X_2, \ldots, X_n\}$ (oznaczaną również przez $\{X_1(t), X_2(t), \ldots, X_n(t)\}$, jeśli jest to wygodne) pobraną z rozkładu zmiennej funkcjonalnej X. Dalej zakładać będziemy, że E jest przestrzenią Hilberta wszystkich funkcji całkowalnych z kwadratem na pewnym przedziale $[\mathrm{a}, \mathrm{b}]$, czyli przestrzenią $L_2([\mathrm{a}, \mathrm{b}])$.

W tym przypadku dane funkcjonalne mogą być przedstawione w postaci:

$$X(t) = \sum_{b=0}^{\infty} c_b \varphi_b(t)$$

gdzie $\varphi_b(t)$ są znanymi, ustalonymi funkcjami ortonormalnymi lub inaczej elementami ortonormalnej bazy $\{\varphi_0, \varphi_1, \ldots\}$. Zauważmy, że reprezentacja funkcji za pomocą nieskończonego szeregu ortonormalnego wymaga znajomości nieskończonej liczby współczynników c_b. Niestety nikt z nas nie potrafi radzić sobie z nieskończoną liczbą współczynników. W związku z tym do aproksymacji funkcji $X(t)$ wykorzystuje się ucięty (skończony) szereg ortonormalny, zwany inaczej sumą częściową, o postaci:

$$X_B(t) := \sum_{b=0}^{B} c_b \varphi_b(t) = \boldsymbol{c}' \varphi(t) = \varphi'(t) \boldsymbol{c}$$

gdzie $\boldsymbol{c} = (c_0, c_1, \ldots, c_B)'$, a $\varphi(t) = (\varphi_0(t), \varphi_1(t), \ldots, \varphi_B(t))'$. Parametr B, będący liczbą naturalną, nazywa się punktem ucięcia. Zazwyczaj tylko niewielka liczba współczynników rozwinięcia jest istotna, a pozostałe są mało znaczące. Prowadzi to do istotnej redukcji danych bez straty informacji w nich zawartej.

Z grubsza rzecz biorąc główny problem statystyczny polega na optymalnym wyborze punktu ucięcia B oraz optymalnym oszacowaniu współczynników c_b. Problem ten będzie opisany w dalszej części artykułu.

W tym miejscu nasuwa się naturalne pytanie, czy w rzeczywistości istnieją dane funkcjonalne? Pytanie to ma istotne znaczenie, gdyż w praktyce wartości obserwowanego procesu losowego $X(t)$ są zawsze rejestrowane w dyskretnych momentach czasu $t_1, t_2, \ldots, t_J$, rzadziej lub gęściej rozmieszczonych w przedziale zmienności argumentu t. Tak więc ostatecznie mamy zawsze do czynienia z szeregiem czasowym $\{x(t_1), x(t_2), \ldots, x(t_J)\}$ lub inaczej — z wysokowymiarowym wektorem obserwacji. Istnieją jednakże liczne powody, by szeregi takie modelować jako elementy przestrzeni funkcjonalnej, ponieważ dane funkcjonalne mają wiele zalet w porównaniu z innymi sposobami reprezentowania szeregów czasowych. Można wyróżnić ich następujące zalety:

— po pierwsze, łatwo radzą sobie z problemem brakujących obserwacji, ponieważ w przypadku danych funkcjonalnych problem ten jest rozwiązany poprzez wyrażenie szeregów czasowych w postaci zbioru krzywych ciągłych;
— po drugie, dane funkcjonalne w sposób naturalny zachowują strukturę obserwacji, tj. zachowują zależność czasową obserwacji i biorą pod uwagę informację o każdym pomiarze;
— po trzecie, momenty obserwacji nie muszą być równomiernie rozmieszczone w poszczególnych szeregach czasowych;
— po czwarte, dane funkcjonalne unikają „przekleństwa" nadmiernej wymiarowości, ponieważ szeregi czasowe zostają zastąpione zbiorem krzywych ciągłych niezależnych od całkowitej liczby punktów czasowych, w których dokonuje się obserwacji.

Chociaż prace dotyczące danych funkcjonalnych pojawiały się już wcześniej, to za symboliczny początek metod statystycznych dla danych funkcjonalnych należy przyjąć ukazanie się monografii Ramsaya i Silvermana (1997). W roku 2005 ukazało się jej drugie wydanie, które pociągnęło za sobą prawdziwy wysyp opracowań związanych z analizą danych funkcjonalnych.

Wśród wielu metod statystycznych skonstruowanych dla danych funkcjonalnych poczesne miejsce zajmują trzy metody określane wspólnym mianem metod redukcji wymiaru. Są to analiza składowych głównych, analiza zmiennych dyskryminacyjnych oraz analiza korelacji i zmiennych kanonicznych. Klasyczne wersje tych metod zakładają, że rozpatrywane obiekty charakteryzowane są wieloma cechami. Tymczasem, w przypadku danych funkcjonalnych, większość dotychczas istniejących prac przyjmuje, że obiekty charakteryzowane są za pomocą jednowymiarowych danych funkcjonalnych. Przykładem jest tu opracowanie He, Mullera i Wanga (2004) poświęcone funkcjonalnej analizie kanonicznej. Pokazuje ono rozbieżność między założeniami metod klasycznych i metod dla danych funkcjonalnych. Pierwszymi autorami, którzy podali kon-

strukcję składowych głównych dla wielowymiarowych danych funkcjonalnych byli Jacques i Preda (2014). Konstrukcja trzech wspomnianych metod redukcji wymiaru dla wielowymiarowych danych funkcjonalnych podana została w pracy Góreckiego, Krzyśki, Waszaka i Wołyńskiego (2016). Zastosowania składowych głównych dla wielowymiarowych danych funkcjonalnych zostały natomiast opisane w pracach Góreckiego, Krzyśki, Waszaka i Wołyńskiego (2014) oraz Krzyśki, Majki i Wołyńskiego (2016). Nasza praca ma na celu przybliżenie Czytelnikowi nowej techniki badawczej, a mianowicie analizę kanoniczną dla wielowymiarowych danych funkcjonalnych oraz jej zastosowanie na przykładzie badania związku między dwoma zbiorami cech charakteryzujących wybrane kraje europejskie w ustalonym przedziale czasowym.

DOBÓR ZMIENNYCH

Dane pochodzą z Eurostatu i dotyczą spożycia indywidualnego według celu w 27 wybranych krajach europejskich we wspomnianym już wcześniej okresie 2000—2010. Wybór krajów i przedziału czasowego podyktowany był dostępnością i kompletnością danych. Klasyfikacja Spożycia Indywidualnego według Celu (*Classification of Individual Consumption by Purpose* — COICOP)[1] obejmuje następujące grupy wydatków:

1) artykuły żywnościowe i napoje bezalkoholowe;
2) napoje alkoholowe i tytoń;
3) odzież i obuwie;
4) mieszkanie, woda, elektryczność, gaz i inne paliwa;
5) wyposażenie wnętrza, sprzęty domowe i bieżące utrzymanie budynku;
6) opieka zdrowotna;
7) transport;
8) łączność;
9) wypoczynek i kultura;
10) szkolnictwo;
11) hotele, kawiarnie i restauracje;
12) różne towary i usługi.

Wydatki uwzględnione w grupie drugiej podzielono na wydatki na napoje alkoholowe $(\text{cecha } Y_1)$ oraz wydatki na wyroby tytoniowe $(\text{cecha } Y_2)$. Jedenaście pozostałych grup wydatków przyjęto jako cechy $X_1, X_2, \ldots, X_{11}$. Interesuje nas związek między cechami Y_1 i Y_2 oraz cechami $X_1, X_2, \ldots, X_{11}$ dla 27 wybranych państw europejskich w latach 2000—2010 rozpatrywanych łącznie.

[1] *Final consumption expenditure of households by consumption purpose — COICOP 3 digit — aggregates at current prices* [nama_co3_c], Eurostat, dostęp 4.05.2016 r.

DANE FUNKCJONALNE

Pokażemy teraz sposób przejścia od szeregu czasowego do funkcji ciągłej.

Niech x_j oznacza zaobserwowaną wartość cechy X w *j*-tym momencie czasowym t_j, gdzie $j = 1, 2, \ldots, J$. Wówczas dane te składają się z J par (t_j, x_j). Takie dane dyskretne można wygładzić za pomocą pewnej funkcji ciągłej $x(t)$, gdzie $t \in I$ oraz I jest zbiorem spójnym, takim że $t_j \in I$, dla $j = 1, 2, \ldots, J$ (Ramsay i Silverman, 2005). Załóżmy, że funkcję $x(t)$ można przedstawić jako kombinację liniową skończonej liczby $B+1$ ortonormalnych funkcji podstawowych w następującej postaci:

$$x(t) = \sum_{b=0}^{B} c_b \varphi_b(t) \quad t \in I \tag{1}$$

gdzie $\{\varphi_b\}$ jest układem ortonormalnych funkcji bazowych, a $c_0, c_1, \ldots, c_B$ są współczynnikami, które podlegają estymacji.

Przypomnijmy, że w przestrzeni Hilberta $L_2(I)$ funkcji całkowalnych z kwadratem zbiór funkcji $\{\varphi_b\}$ jest nazywany ortonormalnym wtedy i tylko wtedy, gdy iloczyn skalarny dowolnych dwóch funkcji z tej przestrzeni jest równy:

$$< \varphi_i(t),\ \varphi_j(t) >= \int_I \varphi_i(t)\varphi_j(t)dt = \delta_{ij}$$

gdzie δ_{ij} jest deltą Kroneckera $(\delta_{ij} = 1$, gdy $i = j$ oraz poza tym $0)$. Najczęściej wybieranymi układami funkcji podstawowych są bazy Fouriera oraz Legendre'a.

Niech $\boldsymbol{x} = (x_1, x_2, \ldots, x_J)'$ będzie wektorem obserwacji, $\boldsymbol{c} = (c_0, c_1, \ldots, c_B)'$ wektorem nieznanych współczynników, natomiast $\boldsymbol{\Phi}(t)$ macierzą wymiaru $J \times (B+1)$, której elementami są wartości ortonormalnych funkcji bazowych $\varphi_b(t_j)$ w kolejnych punktach czasowych t_j, gdzie $b = 0, 1, \ldots, B$, $j = 1, 2, \ldots, J$. Wektor $\boldsymbol{c} = (c_0, c_1, \ldots, c_B)'$ w wyrażeniu (1) jest estymowany metodą najmniejszych kwadratów w ten sposób, aby minimalizować funkcję:

$$S(\boldsymbol{c}) = (\boldsymbol{x} - \boldsymbol{\Phi}(t)\boldsymbol{c})' (\boldsymbol{x} - \boldsymbol{\Phi}(t)\boldsymbol{c})$$

Różniczkując funkcję $S(\boldsymbol{c})$ względem $\boldsymbol{c}$, otrzymujemy estymator najmniejszych kwadratów postaci:

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}'(t)\,\boldsymbol{\Phi}(t))^{-1}\boldsymbol{\Phi}'(t)\boldsymbol{x}$$

Estymacja wektora $\boldsymbol{c}$ metodą najmniejszych kwadratów jest w tej tematyce powszechnie stosowana, poczynając od monografii Ramsaya i Silvermana (2005) i kończąc na pracy przeglądowej Cuevasa (2014).

Załóżmy, że mamy n niezależnych zbiorów $\{(t_{i1,}\, x_{i1}), \ldots, (t_{iJ,}\, x_{iJ})\}$, $i = 1, 2, \ldots, n$. Każdy z tych zbiorów z osobna można wygładzić za pomocą pewnej funkcji ciągłej postaci:

$$x_i(t) = \sum_{b=0}^{B_i} \hat{c}_{ib}\varphi_b(t) \quad i = 1, 2, \ldots, n, \quad t \in I$$

Otrzymujemy wówczas n wartości B_i, $i = 1, 2, \ldots, n$. Optymalna wartość B_i dobierana jest przy użyciu bayesowskiego kryterium informacyjnego wprowadzonego przez Schwarza (1978), a w literaturze anglojęzycznej oznaczanego przez *BIC*:

$$BIC(x_i(t)) = J\ln\left(\frac{\boldsymbol{e}_i'\boldsymbol{e}_i}{J}\right) + (B_i + 1)\left(\frac{\ln J}{J}\right)$$

gdzie $\boldsymbol{e}_j = (e_{i1}, e_{i2}, \ldots, e_{iJ})'$ jest wektorem błędów, takim że $e_j = x_j + -\sum_{b=0}^{B_i} \hat{c}_{ib}\varphi_b(t_j)$, $j = 1, 2, \ldots, J$, $i = 1, 2, \ldots, n$, B_i wyznacza liczbę elementów bazy, natomiast wspólna wartość B jest wartością średnią z poszczególnych wartości B_i. Obszerna dyskusja na temat optymalnego doboru tych wartości, łącznie z badaniami symulacyjnymi, znajduje się w pracy Waszaka (2016).

Dalej zakładać będziemy, że funkcja ciągła $x_i(t)$, wygładzająca zbiór $\{(t_{i1,}\, x_{i1}), \ldots, (t_{iJ,}\, x_{iJ})\}$, ma następującą postać:

$$x_i(t) = \sum_{b=0}^{B} \hat{c}_{ib}\varphi_b(t) \quad i = 1, 2, \ldots, n, \quad t \in I \tag{2}$$

Dotychczas w analizie danych obiekty były scharakteryzowane za pomocą tylko i wyłącznie jednej cechy obserwowanej w wielu momentach czasowych (wielowymiarowość ze względu na czas). Nasze rozważania uogólnimy na przypadek $p \geq 2$ cech. Wówczas dane składają się z n niezależnych funkcji wektorowych $\boldsymbol{x}_i(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))'$, $i = 1, 2, \ldots, n$, przy czym składowe $x_{ib}(t)$ są klasycznymi danymi funkcjonalnymi postaci (2). Zbiór danych

$\{\boldsymbol{x}_i(t),\ \boldsymbol{x}_2(t),\ \ldots,\ \boldsymbol{x}_n(t)\}$ nazywa się zbiorem wielozmiennych danych funkcjonalnych.

Załóżmy, że *d*-ta składowa funkcji wektorowej $\boldsymbol{x}(t)$ może być reprezentowana za pomocą skończonej liczby ortonormalnych funkcji podstawowych φ_b :

$$x_d(t) = \sum_{b=0}^{B_d} \hat{c}_{db}\varphi_b(t) \quad t \in I,\ \ d = 1,\ 2,\ \ldots,\ p$$

gdzie c_{db} są zmiennymi losowymi, takimi że $E(c_{db}) = 0,\ \ \mathrm{Var}(c_{db}) < \infty,$ $d = 1,\ 2,\ \ldots,\ p,\ \ b = 0,\ 1,\ \ldots,\ B_d$. Niech

$$\boldsymbol{c} = (c_{10},\ \ldots,\ c_{1B_1},\ \ldots,\ c_{p0},\ \ldots,\ c_{pB_p})'$$

będzie wektorem współczynników podlegających estymacji metodą najmniejszych kwadratów, natomiast:

$$\boldsymbol{\Phi}(t) = \begin{bmatrix} \boldsymbol{\varphi}'_{B_1}(t) & 0' & \ldots & 0' \\ 0' & \boldsymbol{\varphi}'_{B_2}(t) & \ldots & 0' \\ \ldots & \ldots & \ldots & \ldots \\ 0' & 0' & \ldots & \boldsymbol{\varphi}'_{B_p}(t) \end{bmatrix} \tag{3}$$

macierzą, której elementami są ortonormalne funkcje podstawowe, przy czym $\boldsymbol{\varphi}_{Bd}(t) = (\boldsymbol{\varphi}_0(t),\ \boldsymbol{\varphi}_1(t),\ \ldots,\ \boldsymbol{\varphi}_{Bd}(t))'$ jest wektorem ortonormalnych funkcji podstawowych, odpowiadającym *d*-tej składowej funkcji wektorowej $\boldsymbol{x}(t)$, $d = 1,\ 2,\ \ldots,\ p$. Wówczas funkcję wektorową $\boldsymbol{x}(t)$ możemy zapisać równoważnie w następującej postaci:

$$\boldsymbol{x}(t) = \boldsymbol{\Phi}(t)\boldsymbol{c} \quad t \in I,\ \ E(c) = 0 \tag{4}$$

ZMIENNE KANONICZNE DLA WIELOWYMIAROWYCH DANYCH FUNKCJONALNYCH

Klasyczna analiza kanoniczna wywodząca się z pracy Hotellinga (1936) jest metodą pozwalającą na badanie zależności między zespołem cech zależnych oraz zespołem cech niezależnych. Jeżeli zespół cech zależnych składa się tylko z jednej cechy, to metoda ta jest równoważna regresji wielokrotnej. Obydwa zespoły cech obserwowane są na tych samych jednostkach statystycznych w ustalonym momencie czasu. Jeśli zespoły cech obserwowane są w wielu momentach czasowych, to uzyskane dane reprezentują szeregi czasowe. Szeregi te

mogą być przekształcone do postaci funkcji ciągłych określonych na pewnym przedziale czasowym. Przekształcone dane nazywają się wielowymiarowymi danymi funkcjonalnymi. Przedstawimy teraz konstrukcję zmiennych kanonicznych dla wielowymiarowych danych funkcjonalnych, które noszą nazwę funkcjonalnych zmiennych kanonicznych. Wykorzystując reprezentację danych funkcjonalnych opisaną wcześniej możemy założyć, że składowe $Y_g(t)$ procesu losowego $\boldsymbol{Y}(t)$ oraz składowe $X_h(t)$ procesu $\boldsymbol{X}(t)$ mogą zostać przedstawione odpowiednio za pomocą skończonej liczby ortonormalnych funkcji bazowych $\{\varphi_e\}$ oraz $\{\varphi_f\}$:

$$Y_g(t) = \sum_{e=0}^{E_g} \alpha_{ge}\varphi_e(t) \quad t \in I_1, \quad g = 1, 2, \ldots, p$$

$$X_h(t) = \sum_{f=0}^{F_h} \beta_{hf}\varphi_f(t) \quad t \in I_2, \quad h = 1, 2, \ldots, q$$

Wprowadźmy dodatkowo następującą notację:

$$\boldsymbol{\alpha} = (\alpha_{10}, \ldots, \alpha_{1E_1}, \ldots, \alpha_{p0}, \ldots, \alpha_{pE_p})'$$

$$\boldsymbol{\beta} = (\beta_{10}, \ldots, \beta_{1F_1}, \ldots, \beta_{q0}, \ldots, \beta_{qF_q})'$$

$$\boldsymbol{\Phi}_1(t) = \begin{bmatrix} \boldsymbol{\varphi}'_{E_1}(t) & 0' & \ldots & 0' \\ 0' & \boldsymbol{\varphi}'_{E_2}(t) & \ldots & 0' \\ \ldots & \ldots & \ldots & \ldots \\ 0' & 0' & \ldots & \boldsymbol{\varphi}'_{E_p}(t) \end{bmatrix}$$

$$\boldsymbol{\Phi}_2(t) = \begin{bmatrix} \boldsymbol{\varphi}'_{F_1}(t) & 0' & \ldots & 0' \\ 0' & \boldsymbol{\varphi}'_{F_2}(t) & \ldots & 0' \\ \ldots & \ldots & \ldots & \ldots \\ 0' & 0' & \ldots & \boldsymbol{\varphi}'_{F_q}(t) \end{bmatrix}$$

gdzie $\boldsymbol{\varphi}_{E_1}, \ldots, \boldsymbol{\varphi}_{Ep}$ oraz $\boldsymbol{\varphi}_{F_1}, \ldots, \boldsymbol{\varphi}_{Fg}$ są wektorami, których składowymi są ortonormalne funkcje bazowe odpowiednio przestrzeni $L_2(I_1)$ oraz $L_2(I_2)$. Używając powyższej notacji procesy $\boldsymbol{Y}(t)$ oraz $\boldsymbol{X}(t)$ można zapisać w postaci:

$$\boldsymbol{Y}(t)=\boldsymbol{\Phi}_1(t)\boldsymbol{\alpha} \quad \mathrm{E}(\boldsymbol{\alpha})=0$$

$$\boldsymbol{X}(t)=\boldsymbol{\Phi}_2(t)\boldsymbol{\beta} \quad \mathrm{E}(\boldsymbol{\beta})=0$$

Funkcjonalne zmienne kanoniczne U oraz V dla procesów losowych $\boldsymbol{Y}(t)$ oraz $\boldsymbol{X}(t)$ można zdefiniować w następujący sposób:

$$U =< \boldsymbol{u}(t),\ \boldsymbol{Y}(t) >= \int_{I_1} \boldsymbol{u}'(t)\boldsymbol{Y}(t)dt$$

$$V =< \boldsymbol{v}(t),\ \boldsymbol{X}(t) >= \int_{I_2} \boldsymbol{v}'(t)\boldsymbol{X}(t)dt$$

gdzie funkcje wektorowe $\boldsymbol{u}(t)$ oraz $\boldsymbol{v}(t)$ nazywane są wektorowymi funkcjami wagowymi. Funkcje wagowe $\boldsymbol{u}(t)$ oraz $\boldsymbol{v}(t)$ są dobierane w taki sposób, aby zmaksymalizować współczynnik korelacji

$$\rho = \frac{\mathrm{Cov}(U,V)}{\sqrt{\mathrm{Var}(U)\,\mathrm{Var}(V)}} \tag{5}$$

między funkcjonalnymi zmiennymi kanonicznymi U oraz V, przy dodatkowych warunkach ograniczających

$$\mathrm{Var}(U) = \mathrm{Var}(V) = 1 \tag{6}$$

Identyczne kryterium jest stosowane przy konstrukcji klasycznych zmiennych kanonicznych. Współczynnik ρ nazywany jest współczynnikiem korelacji kanonicznej. Jednakże maksymalizacja tego współczynnika nie daje zadowalających wyników, bowiem w przypadku funkcjonalnym możemy dowolnie wybrać funkcję wektorową $\boldsymbol{u}(t)$, skonstruować zmienną kanoniczną $U =< \boldsymbol{u}(t),\ \boldsymbol{Y}(t) >$, a następnie znaleźć funkcję wektorową $\boldsymbol{v}(t)$ taką, aby współczynnik korelacji ρ dany wzorem (5) był równy 1, gdzie $V =< \boldsymbol{v}(t),\ \boldsymbol{X}(t) >$. Funkcje wagowe $\boldsymbol{u}(t)$ oraz $\boldsymbol{v}(t)$ nie dają zatem użytecznej informacji o sile powiązań między zbiorami cech, co wyraźnie wskazuje na potrzebę zastosowania technik wygładzania. Prostym sposobem wygładzenia jest modyfikacja warunków ograniczających (6) poprzez dodanie pewnego współczynnika kary:

$$\mathrm{U}^{(\mathrm{N})}=\mathrm{Var}(U)+\lambda PEN_2(\boldsymbol{u}(t))=1 \tag{7}$$

$$\mathrm{V}^{(\mathrm{N})}=\mathrm{Var}(V)+\lambda PEN_2(\boldsymbol{v}(t))=1 \tag{8}$$

gdzie współczynnik kary PEN_2 jest scałkowanym kwadratem drugich pochodnych, tj.:

$$\begin{aligned} PEN_2(\boldsymbol{u}(t)) &= \int_{I_1} \left(\frac{\partial^2 \boldsymbol{u}(t)}{\partial t^2}\right)' \frac{\partial^2 \boldsymbol{u}(t)}{\partial t^2} dt = \int_{I_1} \left(\frac{\partial^2 \boldsymbol{\Phi}_1(t)\boldsymbol{u}}{\partial t^2}\right)' \frac{\partial^2 \boldsymbol{\Phi}_1(t)\boldsymbol{u}}{\partial t^2} dt \\ &= \boldsymbol{u}' \int_{I_1} \left(\frac{\partial^2 \boldsymbol{\Phi}_1(t)}{\partial t^2}\right)' \frac{\partial^2 \boldsymbol{\Phi}_1(t)}{\partial t^2} dt\boldsymbol{u} = \boldsymbol{u}'\boldsymbol{R}_1\boldsymbol{u} \end{aligned} \tag{9}$$

gdzie $\boldsymbol{R}_1 = \int_{I_1} \left(\frac{\partial^2 \boldsymbol{\Phi}_1(t)}{\partial t^2} \right)' \frac{\partial^2 \boldsymbol{\Phi}_1(t)}{\partial t^2} dt$

oraz

$$PEN_2(\boldsymbol{v}(t)) = \int_{I_2} \left(\frac{\partial^2 \boldsymbol{v}(t)}{\partial t^2} \right)' \frac{\partial^2 \boldsymbol{v}(t)}{\partial t^2} dt = \int_{I_2} \left(\frac{\partial^2 \boldsymbol{\Phi}_2(t)\boldsymbol{v}}{\partial t^2} \right)' \frac{\partial^2 \boldsymbol{\Phi}_2(t)\boldsymbol{v}}{\partial t^2} dt = \\ = \boldsymbol{v}' \int_{I_2} \left(\frac{\partial^2 \boldsymbol{\Phi}_2(t)}{\partial t^2} \right)' \frac{\partial^2 \boldsymbol{\Phi}_2(t)}{\partial t^2} dt \boldsymbol{v} = \boldsymbol{v}' \boldsymbol{R}_2 \boldsymbol{v} \tag{10}$$

gdzie $\boldsymbol{R}_2 = \int_{I_2} \left(\frac{\partial^2 \boldsymbol{\Phi}_2(t)}{\partial t^2} \right)' \frac{\partial^2 \boldsymbol{\Phi}_2(t)}{\partial t^2} dt$

Współczynnik kary PEN_2 jest uogólnieniem współczynnika wprowadzanego przez Ramsaya i Silvermana (2005, s. 84) na przypadek wielozmienny. Współczynnik PEN_2 służy do oszacowania gładkości funkcji $x(t)$. Kwadrat drugiej pochodnej $\left[D^2 x(t)\right]^2$ tej funkcji w momencie t jest nazywany jej krzywizną w punkcie t. Zauważmy, że linia prosta, z czym się wszyscy zgadzamy, nie ma krzywizny i jej druga pochodna w każdym punkcie jest równa zeru. Im większa wartość drugiej pochodnej funkcji $x(t)$ w punkcie t, tym większa krzywizna tej funkcji w punkcie t. Zatem naturalną miarą krzywizny funkcji jest scałkowany kwadrat drugiej pochodnej tej funkcji:

$$PEN_2 = \int_I [D^2(x(t))]^2 dt$$

Efekt wprowadzenia współczynnika kary jest taki, że bierzemy pod uwagę nie tylko wariancje zmiennych kandydujących na funkcjonalne zmienne kanoniczne, ale także krzywizny i porównujemy ważoną sumę tych dwóch wielkości.

Pierwszy współczynnik korelacji kanonicznej ρ_1 i odpowiadające mu wektorowe funkcje wagowe $\boldsymbol{u}_1(t)$ oraz $\boldsymbol{v}_1(t)$ można zdefiniować następująco:

$$\rho_1 = \sup_{\boldsymbol{u} \in L_2(I_1^p),\, \boldsymbol{v} \in L_2(I_1^q)} \frac{\mathrm{Cov}(<\boldsymbol{u}(t), \boldsymbol{Y}(t)>, <\boldsymbol{v}(t), \boldsymbol{X}(t)>)}{\sqrt{\mathrm{U}^{(\mathrm{N})} \mathrm{V}^{(\mathrm{N})}}} \tag{11}$$

przy dodatkowych warunkach ograniczających

$$\mathrm{U}^{(\mathrm{N})} = \mathrm{V}^{(\mathrm{N})} = 1 \tag{12}$$

W ogólności k-ty współczynnik korelacji kanonicznej ρ_k i odpowiadające mu wektorowe funkcje wagowe $\boldsymbol{u}_k(t)$ oraz $\boldsymbol{v}_k(t)$ są zdefiniowane w następujący sposób:

$$\rho_k = \sup_{\boldsymbol{u}\in L_2(I_1^p),\, \boldsymbol{v}\in L_2(I_1^q)} \mathrm{Cov}(<\boldsymbol{u}(t),\boldsymbol{Y}(t)>,<v(t),\boldsymbol{X}(t)>) =$$
$$= \mathrm{Cov}(<\boldsymbol{u}_k(t),\boldsymbol{Y}(t)>,<\boldsymbol{v}_k(t),\boldsymbol{X}(t)>)$$

gdzie $\boldsymbol{u}_k(t)$ oraz $\boldsymbol{v}_k(t)$ spełniają warunek (12) oraz k-ta para zmiennych kanonicznych (U_k, V_k) nie jest skorelowana z k –1 zmiennymi kanonicznymi, gdzie

$$U_k = <\boldsymbol{u}_k(t),\ \boldsymbol{Y}(t)>$$

$$V_k = <\boldsymbol{v}_k(t),\ \boldsymbol{X}(t)>$$

są funkcjonalnymi zmiennymi kanonicznymi. Taką procedurę nazywa się wygładzoną analizą korelacji kanonicznych. Wyrażenie $(\rho_k,\ \boldsymbol{u}_k(t),\ \boldsymbol{v}_k(t))$ nazywać będziemy k-tym układem kanonicznym pary procesów losowych $\boldsymbol{Y}(t)$ oraz $\boldsymbol{X}(t)$. Niech

$$\sum\nolimits_{11} = \mathrm{Var}(\boldsymbol{\alpha}) = \mathrm{E}(\boldsymbol{\alpha}\boldsymbol{\alpha}')$$
$$\sum\nolimits_{12} = \mathrm{Var}(\boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{\beta}\boldsymbol{\beta}')$$
$$\sum\nolimits_{12} = \mathrm{C}ov(\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{\alpha},\boldsymbol{\beta}')$$

Zdefiniujmy teraz zmienne kanoniczne $U^* = <\boldsymbol{u},\boldsymbol{\alpha}>$ oraz $V^* = <\boldsymbol{v},\boldsymbol{\beta}>$ odpowiednio dla wektorów losowych $\boldsymbol{\alpha}$ oraz $\boldsymbol{\beta}$. k-ta korelacja kanoniczna γ_k i związane z nią wektory wagowe $\boldsymbol{u}_k$ oraz $\boldsymbol{v}_k$ są zdefiniowane następująco:

$$\gamma_k = \sup_{\boldsymbol{u}\in\Re^{E+p},\, \boldsymbol{v}\in\Re^{F+q}} \mathrm{Cov}(<\boldsymbol{u},\boldsymbol{\alpha}>,<\boldsymbol{v},\boldsymbol{\beta}>) = \boldsymbol{u}_k'\boldsymbol{\Sigma}_{12}\boldsymbol{u}_k$$

przy dodatkowych warunkach ograniczających

$$\boldsymbol{u}_k'\left(\sum\nolimits_{11} + \lambda\mathbf{R}_1\right)\boldsymbol{u}_k = 1$$

$$\boldsymbol{v}'_k\left(\sum\nolimits_{22} + \lambda\mathbf{R}_2\right)\boldsymbol{v}_k = 1$$

gdzie macierze $\mathbf{R}_1$ i $\mathbf{R}_2$ są zdefiniowane odpowiednio za pomocą wyrażeń (9) oraz (10), a k-ta para zmiennych kanonicznych (U_k^*, V_k^*) nie jest skorelowana z pierwszymi k –1 parami zmiennych kanonicznych. Wyrażenie $(\gamma_k, \boldsymbol{u}_k, \boldsymbol{v}_k)$ nazywać będziemy k-tym układem kanonicznym wektorów losowych $\boldsymbol{\alpha}$ oraz $\boldsymbol{\beta}$.

Prawdziwe jest następujące twierdzenie (Górecki i in., 2016):

Twierdzenie: *k-ty układ kanoniczny* $(\rho_k, \boldsymbol{u}_k(t), \boldsymbol{v}_k(t))$, *pary procesów losowych* $\boldsymbol{Y}(t)$ *oraz* $\boldsymbol{X}(t)$ *jest zależny od k-tego układu kanonicznego* $(\gamma_k, \boldsymbol{u}_k, \boldsymbol{v}_k)$ *pary wektorów losowych* $\boldsymbol{\alpha}$ *oraz* $\boldsymbol{\beta}$ *poprzez następujące równości:*

$$\rho_k = \gamma_k$$

$$\boldsymbol{u}_k(t) = \boldsymbol{\Phi}_1(t)\boldsymbol{u}_k \quad t \in I_1$$

$$\boldsymbol{v}_k(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{v}_k \quad t \in I_2$$

gdzie $k = 1, 2, \ldots, \min(K_1 + p, K_2 + q)$, $K_1 = E_1 + E_2 + \ldots + E_p$, $K_2 = F_1 + F_2 + \ldots + F_q$.

Z tego twierdzenia wynika, że zmienne kanoniczne U_k i V_k dla pary procesów losowych $\boldsymbol{Y}(t) = \boldsymbol{\Phi}_1(t)\boldsymbol{\alpha}$ oraz $\boldsymbol{X}(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{\beta}$ są postaci $U_k = \boldsymbol{u}_k'\boldsymbol{\alpha}$ oraz $V_k = \boldsymbol{v}_k'\boldsymbol{\beta}$, gdzie $\boldsymbol{u}_k$ oraz $\boldsymbol{v}_k$ są wektorami wagowymi w zmiennych kanonicznych $U_k^* = \boldsymbol{u}_k'\boldsymbol{\alpha}$ oraz $V_k^* = \boldsymbol{v}_k'\boldsymbol{\beta}$ dla pary wektorów losowych $\boldsymbol{\alpha}$ oraz $\boldsymbol{\beta}$, $k = 1, 2, \ldots, \min(K_1 + p, K_2 + q)$.

Analiza korelacji kanonicznych dla wektorów losowych $\boldsymbol{\alpha}$ oraz $\boldsymbol{\beta}$ opiera się na macierzach $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{22}$ i $\boldsymbol{\Sigma}_{12}$, które są nieznane. Estymujemy je na podstawie n niezależnych realizacji $\boldsymbol{y}_1(t)$, $\boldsymbol{y}_2(t)$, …, $\boldsymbol{y}_n(t)$, postaci $\boldsymbol{y}_i(t) = \boldsymbol{\Phi}_1(t)\hat{\boldsymbol{\alpha}}_i$, procesu losowego $\boldsymbol{Y}(t)$ oraz $\boldsymbol{x}_1(t)$, $\boldsymbol{x}_2(t)$, …, $\boldsymbol{x}_n(t)$, postaci $\boldsymbol{x}_i(t) = \boldsymbol{\Phi}_2(t)\hat{\boldsymbol{\beta}}_i$, procesu losowego $\boldsymbol{X}(t)$, $i = 1, 2, \ldots, n$, gdzie:

$$\hat{\boldsymbol{\alpha}}_i = (\hat{\alpha}_{10}^{(i)}, ..., \hat{\alpha}_{1E_1}^{(i)}, ..., \hat{\alpha}_{p0}^{(i)}, ..., \hat{\alpha}_{pE_p}^{(i)})'$$

$$\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{10}^{(i)}, ..., \hat{\beta}_{1F_1}^{(i)}, ..., \hat{\beta}_{q0}^{(i)}, ..., \hat{\beta}_{qF_q}^{(i)})'$$

Niech $\hat{\mathbf{A}} = (\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2, ..., \hat{\boldsymbol{\alpha}}_n)'$ oraz $\hat{\mathbf{B}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, ..., \hat{\boldsymbol{\beta}}_n)'$. Wówczas estymatory macierzy $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{22}$ i $\boldsymbol{\Sigma}_{12}$ są odpowiednio postaci:

$$\hat{\boldsymbol{\Sigma}}_{11} = \frac{1}{n}\hat{\mathbf{A}}'\hat{\mathbf{A}}$$

$$\hat{\boldsymbol{\Sigma}}_{22} = \frac{1}{n}\hat{\mathbf{B}}'\hat{\mathbf{B}}$$

$$\hat{\boldsymbol{\Sigma}}_{12} = \frac{1}{n}\hat{\mathbf{A}}'\hat{\mathbf{B}}$$

Ponadto niech $\hat{\mathbf{C}} = \hat{\mathbf{\Sigma}}_{11}^{-1}\hat{\mathbf{\Sigma}}_{12}$ oraz $\hat{\mathbf{D}} = \hat{\mathbf{\Sigma}}_{22}^{-1}\hat{\mathbf{\Sigma}}_{21}$, gdzie $\hat{\mathbf{\Sigma}}_{21} = \hat{\mathbf{\Sigma}}_{12}$. Macierze $\hat{\mathbf{C}}\hat{\mathbf{D}}$ i $\hat{\mathbf{D}}\hat{\mathbf{C}}$ mają te same niezerowe wartości własne $\hat{\gamma}_k^2$, a odpowiadające im wektory własne $\hat{\boldsymbol{u}}_k$ oraz $\hat{\boldsymbol{v}}_k$ są wyznaczane z równości:

$$(\hat{\mathbf{C}}\hat{\mathbf{D}} - \hat{\gamma}_k^2\mathbf{I}_{K_1+p})\hat{\boldsymbol{u}}_k = 0$$

$$(\hat{\mathbf{D}}\hat{\mathbf{C}} - \hat{\gamma}_k^2\mathbf{I}_{K_2+q})\hat{\boldsymbol{v}}_k = 0$$

gdzie $k = 1, 2, \ldots, \min(K_1 + p, K_2 + q)$.

Wówczas k-ty układ kanoniczny pary procesów losowych $\boldsymbol{Y}(t)$ oraz $\boldsymbol{X}(t)$ wyznaczony z próby ma następującą postać:

$$(\hat{\lambda}_k = \hat{\gamma}_k,\ \hat{\boldsymbol{u}}_k(t) = \boldsymbol{\Phi}_1(t)\,\hat{\boldsymbol{u}}_k,\ \hat{\boldsymbol{v}}_k(t) = \boldsymbol{\Phi}_2(t)\,\hat{\boldsymbol{v}}_k)$$

gdzie $k = 1, 2, \ldots, \min(K_1 + p, K_2 + q)$.

Stąd współczynniki rzutu i-tej realizacji $\boldsymbol{y}_i(t)$ procesu $\boldsymbol{Y}(t)$ na k-tą funkcjonalną zmienną kanoniczną są równe:

$$\hat{U}_{ik} = <\hat{\boldsymbol{u}}_k(t), \boldsymbol{y}_i(t)> = \int_{I_1} \hat{\boldsymbol{u}}_k(t)\boldsymbol{y}_i(t)dt = \hat{\boldsymbol{\alpha}}_i'\hat{\boldsymbol{u}}_k$$

Analogicznie współczynniki rzutu i-tej realizacji $\boldsymbol{x}_i(t)$ procesu $\boldsymbol{X}(t)$ na k-tą funkcjonalną zmienną kanoniczną są równe:

$$\hat{V}_{ik} = \hat{\boldsymbol{\beta}}_i'\hat{\boldsymbol{v}}_k$$

gdzie $i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, \min(K_1 + p, K_2 + q)$.

Ostatecznie współczynniki rzutu i-tej realizacji $(\boldsymbol{x}_i(t), \boldsymbol{y}_i(t))$ procesów losowych $\boldsymbol{X}(t)$ oraz $\boldsymbol{Y}(t)$ na płaszczyznę dwóch pierwszych funkcjonalnych zmiennych kanonicznych z próby są równe $(\hat{\boldsymbol{\beta}}_i'\hat{\boldsymbol{v}}_1,\ \hat{\boldsymbol{\alpha}}_i'\hat{\boldsymbol{u}}_1)$, $i = 1, 2, \ldots, n$.

O wkładzie poszczególnych składowych wektorowych procesów losowych $\boldsymbol{Y}(t)$ oraz $\boldsymbol{X}(t)$ w budowę zmiennych kanonicznych można wnioskować na podstawie wektorowych funkcji wagowych $\boldsymbol{u}(t)$ oraz $\boldsymbol{v}(t)$.

Zmienna kanoniczna U_k wyznaczona jest przez wektorową funkcję wagową $\boldsymbol{u}_k(t)=(u_{k1}(t), u_{k2}(t), ..., u_{kp}(t))'$. Składowa $Y_j(t)$ procesu wektorowego $\boldsymbol{Y}(t)=(Y_1(t), Y_2(t), ..., Y_p(t))'$ ma największy wkład w budowę zmiennej kanonicznej U_k w chwili t, jeżeli:

$$|u_{kj}(t)| = \max_{1\le i\le p}|u_{ki}(t)|$$

Niech P_j będzie polem pod modułem funkcji $u_{kj}(t)$ na przedziale I oraz niech

$$P_j^* = \frac{P_j}{\sum_{i=1}^{p} P_j} \times 100\%, \quad j = 1,\ 2,\ \ldots,\ p \tag{13}$$

Składowa $Y_j(t)$ procesu wektorowego $\boldsymbol{Y}(t) = (Y_1(t),\ Y_2(t),\ \ldots,\ Y_p(t))$ ma największy wkład w budowę zmiennej kanonicznej U_k, dla t zmieniających się w przedziale I, jeżeli:

$$P_j^* = \max_{1 \le i \le p} \left| P_i^* \right|$$

Analogicznie wnioskujemy dla zmiennej kanonicznej V_k.

WYNIKI BADAŃ EMPIRYCZNYCH

Analizą objęto 27 wybranych krajów europejskich $(n = 27)$. Analizowane dane obejmują 11 lat $(J = 11)$. Każdy kraj scharakteryzowano za pomocą dwóch zmiennych zależnych Y_1 i Y_2 oraz 11 zmiennych niezależnych $X_1, \ldots, X_{11}$. Dane pierwotne poddano unitaryzacji zerowanej, a następnie przekształcono na wielowymiarowe dane funkcjonalne. Posłużono się funkcjami bazowymi Fouriera postaci:

$$\varphi_0(t) = \frac{1}{\sqrt{T}}$$

$$\varphi_{2k-1}(t) = \sqrt{\frac{2}{T}} \sin \frac{2\pi k t}{T}$$

$$\varphi_{2k}(t) = \sqrt{\frac{2}{T}} \cos \frac{2\pi k t}{T}$$

gdzie $t \in [0,\ T],\ k = 1,\ 2,\ \ldots$

Górecki i Krzyśko (2012) pokazali, że baza Fouriera prowadzi do minimalnej liczby wyrazów w rozwinięciu danej funkcji w szereg, co jest cechą nadzwyczaj pożądaną, ponieważ współczynniki rozwinięcia pełnią rolę nowych zmiennych w podejściu funkcjonalnym.

Przedziały czasowe $I_1 = I_2 = I = [0,\ 11]$ zostały podzielone na momenty czasowe następująco: $t_1 = 0{,}5\ (2000),\ t_2 = 1{,}5,\ \ldots,\ t_{11} = 10{,}5\ (2010)$. Następnie skonstruowano funkcjonalne zmienne kanoniczne odpowiadające procesom

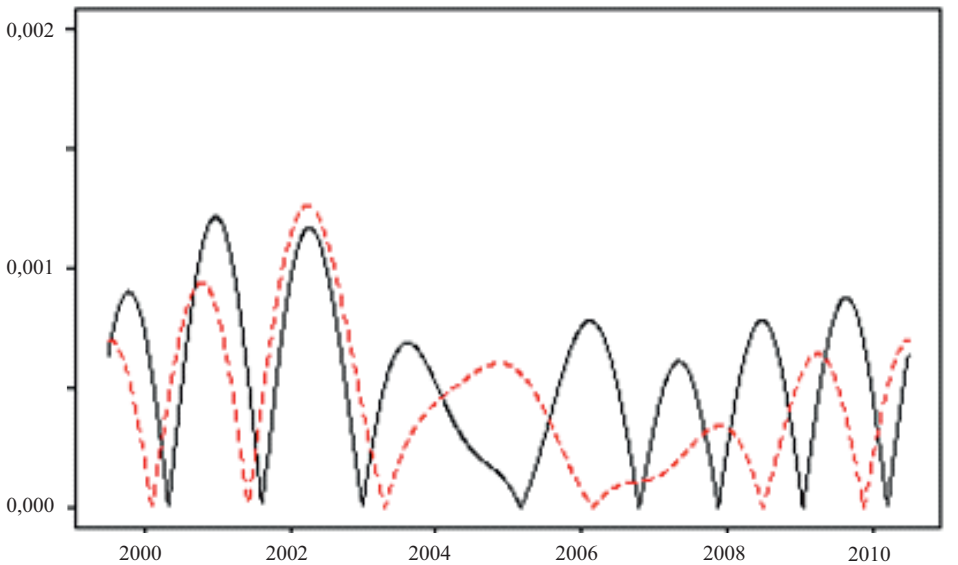
losowym $\boldsymbol{Y}(t)$ i $\boldsymbol{X}(t)$. W celu ich wyznaczenia użyto w obliczeniach pakietu R (R Core Team, 2015). Maksymalny współczynnik korelacji kanonicznej ρ_1, odpowiadający pierwszej parze zmiennych kanonicznych U_1 i V_1, wynosi 0,99. Świadczy to o bardzo silnym związku między wydatkami na napoje alkoholowe i wyroby tytoniowe oraz pozostałe artykuły konsumpcyjne. Obrazy dwóch funkcji wagowych dla procesu $\boldsymbol{Y}(t)$ oraz 11 funkcji wagowych dla procesu $\boldsymbol{X}(t)$ przedstawione są odpowiednio na wykr. 1 i 2. Wskaźniki P_1 oraz P_2 dane wzorem (13), odpowiadające wydatkom na napoje alkoholowe oraz na wyroby tytoniowe, wynoszą 54,7% i 45,3%. Świadczy to o prawie jednakowym wkładzie tych dwóch zmiennych w budowę pierwszej funkcjonalnej zmiennej kanonicznej U_1. Wskaźniki P_i odpowiadające zmiennym X_i przedstawiono w tablicy. Z tych danych wynika, że największy udział w budowie funkcjonalnej zmiennej kanonicznej V_1 mają zmienne X_1 — artykuły żywnościowe i napoje bezalkoholowe (10,3%) oraz X_2 — odzież i obuwie (10,1%). Na wykr. 3 każdy z krajów przedstawiono jako punkt w układzie dwóch pierwszych funkcjonalnych zmiennych kanonicznych (V_1, U_1). Wykr. 4 jest powiększoną lewą dolną ćwiartką wykr. 3. Wysoki stopień skorelowania wydatków na napoje alkoholowe i wyroby tytoniowe oraz pozostałe artykuły konsumpcyjne objawia się wysokim skorelowaniem funkcjonalnych zmiennych kanonicznych U_1 i V_1 $(\rho_1 = 0{,}99)$, dlatego na wykr. 4 punkty przedstawiające poszczególne kraje leżą prawie na linii prostej. Widać, że wraz ze wzrostem indywidualnych wydatków konsumpcyjnych gospodarstw domowych, podzielonych na 11 grup, rosną wydatki na napoje alkoholowe i wyroby tytoniowe. Z jednej strony widzimy kraje o niskich wydatkach na dobra i usługi konsumpcyjne oraz niskich wydatkach na alkohol i tytoń (Malta, Łotwa, Estonia, Luksemburg, Cypr, Litwa, Bułgaria, Słowacja i Słowenia), a z drugiej strony kraje o wysokich wydatkach na dobra i usługi konsumpcyjne oraz na alkohol i tytoń (Niemcy, Wielka Brytania, Francja, Włochy, Hiszpania i Holandia). Pierwsze miejsce zajmują tu bezsprzecznie Niemcy, a Polska zajęła pozycję pośrednią, plasując się między Holandią i Grecją.

WSKAŹNIKI P_j^* ODPOWIADAJĄCE ZMIENNYM X_j

j	P_j^*
1	10,292
2	10,121
3	9,282
4	9,357
5	7,228
6	9,955
7	9,186
8	9,267
9	8,515
10	9,257
11	7,539

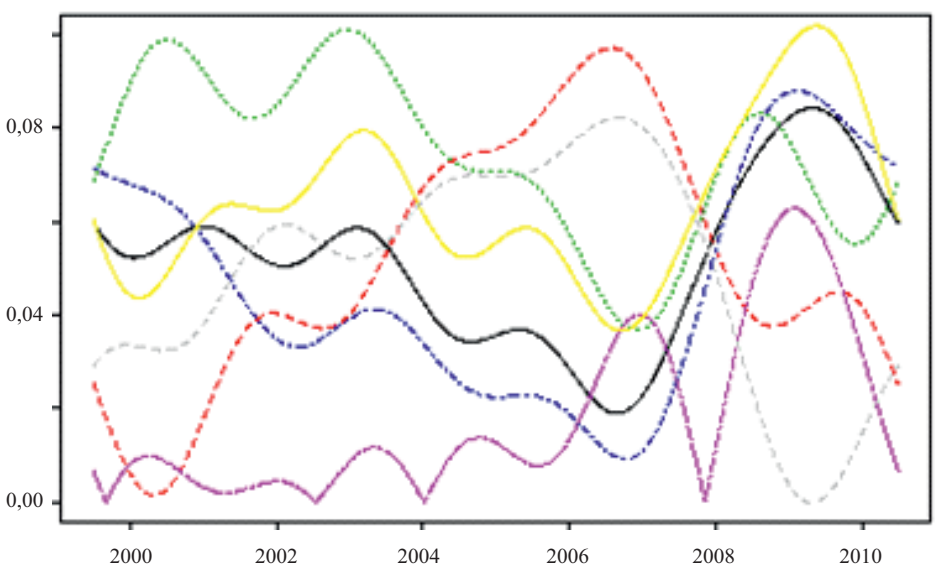
Ź r ó d ł o: opracowanie własne na podstawie danych Eurostatu.

Wykr. 1. WYKRES FUNKCJI WAGOWYCH PIERWSZEJ ZMIENNEJ KANONICZNEJ DLA PROCESU $Y_{(t)}$



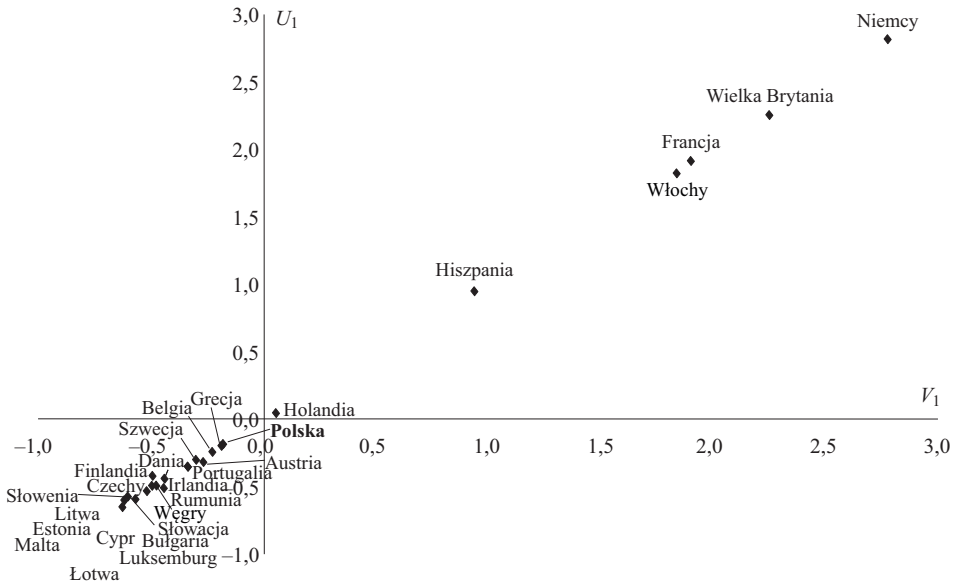
Ź r ó d ł o: opracowanie własne na podstawie danych Eurostatu.

Wykr. 2. WYKRES FUNKCJI WAGOWYCH PIERWSZEJ ZMIENNEJ KANONICZNEJ DLA PROCESU $X_{(t)}$



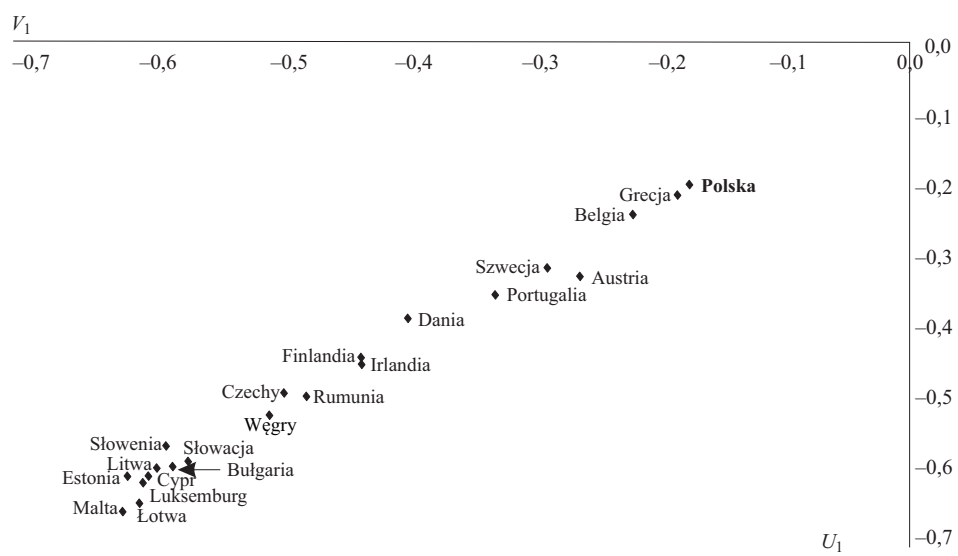
Ź r ó d ł o: jak przy wykr. 1.

Wykr. 3. POŁOŻENIE 27 KRAJÓW W UKŁADZIE PIERWSZYCH ZMIENNYCH KANONICZNYCH (V_1, U_1)



Ź r ó d ł o: jak przy wykr. 1.

Wykr. 4. POWIĘKSZONA LEWA DOLNA ĆWIARTKA WYKR. 3



Ź r ó d ł o: jak przy wykr. 1.

Podsumowanie

Zespół cech zależnych składa się z dwóch cech, a zespół cech niezależnych z 11. Cechy te obserwowano w 27 wybranych państwach europejskich w latach 2000—2010. Klasyczną analizę kanoniczną Hotellinga można by zastosować tylko dla każdego roku oddzielnie, a więc w celu uwzględnienia wszystkich lat z okresu 2000—2010 łącznie zastosowano funkcjonalną analizę kanoniczną podaną w pracy Góreckiego i in. (2016). W latach 2000—2010 wykazano bardzo silny związek między wydatkami na napoje alkoholowe i wyroby tytoniowe oraz pozostałymi wydatkami konsumpcyjnymi. Pierwszy współczynnik korelacji kanonicznej wynosił 0,99. Badanie pokazało, że im większe wydatki konsumpcyjne, tym większe wydatki na napoje alkoholowe i wyroby tytoniowe. Do państw o wysokich wydatkach na dobra i usługi konsumpcyjne oraz na alkohol i tytoń należą: Niemcy, Wielka Brytania, Francja, Włochy, Hiszpania i Holandia, natomiast do państw o niskich wydatkach na dobra i usługi konsumpcyjne oraz na alkohol i tytoń należą: Malta, Łotwa, Estonia, Luksemburg, Cypr, Litwa, Bułgaria, Słowacja i Słowenia. Polska należy do grupy państw o średnim stopniu zależności między tymi zespołami cech i jest najbliższa Holandii z grupy pierwszej oraz Grecji należącej do tej samej grupy co Polska. W budowę pierwszej funkcjonalnej zmiennej kanonicznej dla cech zależnych prawie jednakowy wkład wnosiły wydatki na alkohol (54,7%) i na tytoń (45,3%), natomiast w pierwszej funkcjonalnej zmiennej kanonicznej dla cech niezależnych największy udział miały wydatki na artykuły żywnościowe i napoje bezalkoholowe (10,3%) oraz na odzież i obuwie (10,1%).

dr Karol Deręgowski — *Państwowa Wyższa Szkoła Zawodowa w Kaliszu im. Prezydenta Stanisława Wojciechowskiego*
prof. dr hab. Mirosław Krzyśko — *Uniwersytet im. Adama Mickiewicza w Poznaniu, Państwowa Wyższa Szkoła Zawodowa w Kaliszu im. Prezydenta Stanisława Wojciechowskiego*
dr Łukasz Waszak, dr hab. Waldemar Wołyński — *Uniwersytet im. Adama Mickiewicza w Poznaniu*

LITERATURA

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference, vol. 147*, s. 1—23.

Górecki, T., Krzyśko, M. (2012). *Functional principal components analysis*. W: J. Pociecha, R. Decker (eds.), *Data Analysis and Methods and Its Applications*, s. 71—87. Warszawa: Wydawnictwo C.H. Beck.

Górecki, T., Krzyśko, M., Waszak, Ł., Wołyński, W. (2014). Methods of reducing dimension for functional data. *Statistics in Transition new series, vol. 15, no. 2*, s. 231—242. Warszawa GUS i PTS.

Górecki, T., Krzyśko, M., Waszak, Ł., Wołyński, W. (2016). *Selected statistical methods of data analysis for multivariate functional data*. Statistical Papers, DOI 10.1007/s00362-016-0757-8.

He, G., Muller, H.G., Wang, J.L. (2004). Methods of canonical analysis for functional data. *Journal of Statistical Planning and Inference, vol. 122*, s. 141—159.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika, vol. 28*, s. 321—377.

Jacques, J., Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics&Data Analysis, vol. 71*, s. 92—106.

Krzyśko, M., Majka, A., Wołyński, W. (2016). Ocena zróżnicowania poziomu życia mieszkańców województw w latach 2002—2013 za pomocą składowych głównych dla wielozmiennych danych funkcjonalnych oraz analizy skupień. *Przegląd Statystyczny, nr 63(1)*, s. 81—97.

Piekut, M. (2016). Wydatki na wybrane używki w europejskich gospodarstwach domowych. *Wiadomości Statystyczne, nr 3,* s. 85—100. Warszawa: GUS i PTS.

Ramsay, J.O., Silverman, B.W. (1997). *Functional Data Analysis*. New York: Springer.

Ramsay, J.O., Silverman, B.W. (2005). *Functional Data Analysis*, *Second Edition.* New York: Springer.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Pobrano z https://www.R-project.org.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, vol. 6, no. 2*, s. 461—464.

Waszak, Ł. (2016). *Wybrane wielowymiarowe metody statystyczne dla wielozmiennych danych funkcjonalnych*. Rozprawa doktorska. Poznań: Wydział Matematyki i Informatyki, Uniwersytet im. Adama Mickiewicza.

Summary. *The article aims to examine the relations between expenditure on alcoholic beverages and tobacco and other consumer expenditure of households in 27 European countries within 2000—2010. The choice of countries and time series was determined by the availability and completeness of Eurostat data. The years were analysed collectively not separately, which is a novelty presented in this paper. Such an approach was possible due the transformation of primary data into multivariate functional ones, and then the construction of correlations and canonical variables for transformed data.*

The study shows that expenditure on alcoholic beverages and tobacco is strongly correlated with other consumption expenditure (the canonical correlation coefficient between the two first functional canonical variables is 0.99). The expenditure on alcoholic beverages and tobacco has almost the same contribution to the construction of the functional canonical U_1 variable, while the expenditure on food and non-alcoholic beverages and expenditure on clothing and footwear has the largest impact on the development of the functional canonical V_1 variable.

Keywords: canonical analysis, multivariate functional data analysis (MFDA), individual consumption according to purpose.