DOI: 10.5604/01.3001.0014.1152

PRZEGLĄD STATYSTYCZNY
R. LXIII – ZESZYT 1 – 2016

MIROSŁAW KRZYŚKO[1,3], AGNIESZKA MAJKA[2], WALDEMAR WOŁYŃSKI[3]

# OCENA ZRÓŻNICOWANIA POZIOMU ŻYCIA MIESZKAŃCÓW WOJEWÓDZTW W LATACH 2003–2013 ZA POMOCĄ SKŁADOWYCH GŁÓWNYCH DLA WIELOWYMIAROWYCH DANYCH FUNKCJONALNYCH ORAZ ANALIZY SKUPIEŃ

## 1. WSTĘP

Badanie poziomu życia i jego zróżnicowania nabiera szczególnego znaczenia w kontekście analizy stopnia przemian gospodarczych, porównania rozwoju wybranych obszarów czy wskazania dysproporcji życia społeczeństwa zamieszkującego dany region. Dzięki takim ocenom można wskazać dystans dzielący poszczególne regiony, wyodrębnić grupy o zbliżonym poziomie życia, uchwycić podobieństwa i różnice występujące pomiędzy poziomem życia w poszczególnych jednostkach administracyjnych czy określić zagrożenia danego regionu.

Wstępując do Unii Europejskiej (UE) Polska włączyła się w realizację polityki spójności mającej na celu promowanie harmonijnego rozwoju całego terytorium UE poprzez działania prowadzące do zmniejszenia zróżnicowania w rozwoju jej regionów, a tym samym do wzmocnienia spójności gospodarczej, społecznej i terytorialnej Wspólnoty. Efektem tej polityki powinno być wyrównywanie dysproporcji w poziomie życia mieszkańców poszczególnych regionów UE. W ramach polityki spójności w latach 2007–2013 Polska otrzymała łącznie 67 mld euro, czyli 20% całego budżetu UE przeznaczonego na ten cel.

Biorąc pod uwagę, że wszystko co dzieje się w społeczeństwie i w gospodarce zmienia się w miarę upływu czasu, zasadne wydaje się spojrzenie na zmiany, jakie nastąpiły w zróżnicowaniu poziomu życia mieszkańców poszczególnych regionów Polski. Celem artykułu jest ocena zróżnicowania poziomu życia mieszkańców województw w latach 2003–2013.

[1] Państwowa Wyższa Szkoła Zawodowa im. Prezydenta Stanisława Wojciechowskiego, Wydział Zarządzania, ul. Nowy Świat 4, 62-800 Kalisz, Polska, autor prowadzący korespondencję, e-mail: mkrzysko @amu.edu.pl.

[2] Uniwersytet Rzeszowski, Wydział Ekonomii, Katedra Metod Ilościowych i Informatyki Gospodarczej, ul. Ćwiklińskiej 2, 35-601 Rzeszów, Polska.

[3] Uniwersytet im. Adama Mickiewicza w Poznaniu, Wydział Matematyki i Informatyki, ul. Umultowska 87, 61-614 Poznań, Polska.

Nowością tej pracy jest fakt rozpatrywania lat 2003–2013 łącznie, a nie oddzielnie każdego roku. Można to było osiągnąć po przekształceniu danych oryginalnych w postaci szeregów czasowych dla każdej cechy oddzielnie na wektorowe funkcje ciągłe określone na ustalonym przedziale czasowym zwane wielowymiarowymi danymi funkcjonalnymi (patrz Jacques, Preda, 2014 oraz Górecki i inni, 2014). Do oceny przestrzennego zróżnicowania poziomu życia zastosowano analizę składowych głównych dla wielowymiarowych danych funkcjonalnych (patrz sekcja 3) oraz dendrytową metodę analizy skupień (patrz sekcja 4). Metody te pozwoliły na wyodrębnienie grup województw o zbliżonym poziomie wartości rozpatrywanych cech dla całego rozpatrywanego okresu łącznie. Wszystkie obliczenia wykonane zostały przy użyciu programu R.

## 2. DOBÓR ZMIENNYCH ORAZ ICH UNITARYZACJA

Na podstawie przesłanek merytorycznych ustalono zestaw zmiennych diagnostycznych dotyczących wielu aspektów życia, takich jak: wynagrodzenia, rynek pracy, opieka zdrowotna i społeczna, komunikacja i infrastruktura gospodarcza, sytuacja mieszkaniowa, oświata i kultura, środowisko, bezpieczeństwo. Przy doborze zmiennych kierowano się koniecznością w miarę wszechstronnego opisu poziomu życia, dostępnością i kompletnością danych statystycznych. Wybrane zmienne miały charakter wskaźnikowy. Na liście zmiennych diagnostycznych znalazły się:

**wynagrodzenia i rynek pracy:**

$x_1$ — przeciętne miesięczne wynagrodzenie brutto w zł (w podmiotach gospodarczych o liczbie pracujących powyżej 9 osób),
$x_2$ — przeciętny miesięczny dochód rozporządzalny na 1 osobę,
$x_3$ — stopa bezrobocia rejestrowanego w % (d),
$x_4$ — liczba pracujących na 1000 mieszkańców,
$x_5$ — odsetek pracujących w rolnictwie (d),
$x_6$ — odsetek pracujących w usługach,
$x_7$ — bezrobotni z wykształceniem wyższym w ogólnej liczbie ludności w wieku produkcyjnym (d),
$x_8$ — bezrobotni pozostający bez pracy powyżej 24 miesięcy w ogólnej liczbie bezrobotnych (d),
$x_9$ — ludność w wieku produkcyjnym na 1 tys. osób w wieku nieprodukcyjnym,
$x_{10}$ — ludność w wieku poprodukcyjnym na 1 tys. osób w wieku produkcyjnym (d);

**opieka zdrowotna i społeczna:**

$x_{11}$ — wydatki budżetowe w dziale ochrona zdrowia na 1 mieszkańca w zł,
$x_{12}$ — liczba lekarzy na 1 tys. mieszkańców,
$x_{13}$ — liczba mieszkańców przypadających na 1 aptekę ogólnodostępną (d),
$x_{14}$ — placówki stacjonarnej opieki społecznej na 1 tys. ludności;

**infrastruktura gospodarcza:**
$x_{15}$ — rozdzielcza sieć wodociągowa w km na 100 km$^2$,
$x_{16}$ — rozdzielcza sieć kanalizacyjna w km na 100 km$^2$,
$x_{17}$ — rozdzielcza sieć gazowa w km na 100 km$^2$,
$x_{18}$ — gęstość dróg (drogi o twardej nawierzchni w km na 100 km$^2$),
$x_{19}$ — liczba ludności przypadająca na 1 placówkę pocztową (d);

**zasoby mieszkaniowe:**
$x_{20}$ — wydatki budżetowe w dziale gospodarka mieszkaniowa na 1 mieszkańca w zł,
$x_{21}$ — przeciętna powierzchnia mieszkań w przeliczeniu na 1 osobę,
$x_{22}$ — liczba mieszkań na 1 tys. ludności,
$x_{23}$ — odsetek mieszkań wyposażonych w wodociąg,
$x_{24}$ — odsetek mieszkań wyposażonych w łazienkę,
$x_{25}$ — odsetek mieszkań wyposażonych w gaz sieciowy;

**oświata, kultura i rekreacja:**
$x_{26}$ — odsetek dzieci w wieku 3–6 lat objętych wychowaniem przedszkolnym,
$x_{27}$ — studenci szkół wyższych na 10 tys. ludności,
$x_{28}$ — liczba uczniów szkół podstawowych przypadających na 1 komputer z dostępem do Internetu (d),
$x_{29}$ — liczba uczniów szkół ponadgimnazjalnych przypadających na 1 komputer z dostępem do Internetu (d),
$x_{30}$ — wydatki budżetowe w dziale kultura i sport na 1 mieszkańca w zł,
$x_{31}$ — liczba klubów sportowych na 1 tys. ludności,
$x_{32}$ — księgozbiór bibliotek na 1 tys. ludności,
$x_{33}$ — domy i ośrodki kultury, kluby i świetlice na 10 tys. mieszkańców;

**bezpieczeństwo i środowisko:**
$x_{34}$ — przestępstwa stwierdzone w zakończonych postępowaniach przygotowawczych w przeliczeniu na 10 tys. mieszkańców (d),
$x_{35}$ — nakłady na środki trwałe służące ochronie środowiska na 1 mieszkańca,
$x_{36}$ — odpady wytworzone na 1 km$^2$ (poza odpadami komunalnymi) (d),
$x_{37}$ — lesistość w %.
Literą (d) oznaczono destymulanty. Pozostałe cechy są stymulantami.

W celu ujednolicenia wartości rozpatrywanych cech, które są wyrażone w różnych jednostkach pomiarowych i mają różne przedziały zmienności, przeprowadzono ich unitaryzację zerowaną (por. np. Walesiak, 2014).

Niech $X_{kij}$ będzie wartością cechy $X_k$ zaobserwowaną w $i$-tym województwie oraz $j$-tym roku, gdzie $k = 1,\ldots,37$, $i = 1,\ldots,16$, $j = 1,\ldots,11$.

Wówczas zunitaryzowana wartość $z_{kij}$ wartości $X_{kij}$ ma postać:

$$z_{kij} = \begin{cases} \dfrac{x_{kij} - \min_i x_{kij}}{r_{kj}}, & \text{gdy cecha } X_k \text{ jest stymulantą,} \\ \dfrac{\max_i x_{kij} - x_{kij}}{r_{kj}}, & \text{gdy cecha } X_k \text{ jest destymulantą,} \end{cases}$$

gdzie

$$r_{kj} = \max_i x_{kij} - \min_i x_{kij}$$

jest rozstępem $k$-tej cechy w $j$-tym roku.

Równoważnie, zunitaryzowane wartości $z_{kij}$ można zapisać w postaci:

$$z_{kij} = b_{kj} x_{kij} + a_{kj},$$

gdzie

$$b_{kj} = \frac{1}{r_{kj}},\ a_{kj} = -\frac{\min_i x_{kij}}{r_{kj}}, \tag{1}$$

w przypadku stymulant oraz

$$b_{kj} = \frac{1}{r_{kj}},\ a_{kj} = -\frac{\max_i x_{kij}}{r_{kj}}, \tag{2}$$

w przypadku destymulant.

Niech $\bar{x}_{kj}$, $s_{kj}^2$ oraz $r_{kj}$ będą odpowiednio: wartością średnią, wariancją oraz rozstępem cechy $X_k$ w $j$-tym roku. Wówczas zunitaryzowane wartości tych wielkości są równe:

$$\bar{z}_{kij} = \begin{cases} \dfrac{\bar{x}_{kij} - \min_i x_{kij}}{r_{kj}}, & \text{gdy cecha } X_k \text{ jest stymulantą,} \\ \dfrac{\max_i x_{kij} - \bar{x}_{kij}}{r_{kj}}, & \text{gdy cecha } X_k \text{ jest destymulantą,} \end{cases}$$

$$\tilde{s}_{kj}^2 = \frac{s_{kj}^2}{r_{kj}^2},$$
$$\tilde{r}_{kj} = 1,$$

dla $k = 1,\ldots,37$, $j = 1,\ldots,11$.

Jeśli $s_{klj}$ jest kowariancją między cechami $X_k$ oraz $X_j$ w $j$-tym roku, gdzie $k \neq l$, to kowariancja danych zunitaryzowanych jest równa:

$$\overline{s}_{klj} = b_{kj} b_{lj} s_{klj},$$

gdzie $b_{kj}$ oraz $b_{lj}$ dane są wzorem (1) dla stymulant oraz wzorem (2) dla destymulant.

Widzimy, że po unitaryzacji zerowanej rozstęp wszystkich cech we wszystkich latach jest stały i równy 1, natomiast wariancje i kowariancje w danym roku są proporcjonalne do wariancji i kowariancji cech bez unitaryzacji zerowanej w tym roku.

## 3. WIELOWYMIAROWE FUNKCJONALNE SKŁADOWE GŁÓWNE

W rozdziale tym omówiona zastanie analiza składowych głównych dla wielowymiarowych danych funkcjonalnych (MFPCA) (por. Jacques, Preda, 2014; Górecki i inni, 2014). Klasycznymi pozycjami z zakresu metod statystycznych dla jednowymiarowych danych funkcjonalnych są monografie Ramsaya, Silvermana (2005) oraz Horvátha, Kokoszki (2012).

Załóżmy, że obserwujemy $p$-wymiarowy proces stochastyczny $\boldsymbol{X}(t) = (X_1(t), X_2(t),\ldots,X_p(t))'$ z ciągłym parametrem $t \in I$. Dalej załóżmy, że $\mathrm{E}(\boldsymbol{X}(\mathrm{t})) = \mathbf{0}$ i $\boldsymbol{X}(\mathrm{t}) \in L_2^p(I)$, gdzie $L_2(I)$ jest przestrzenią Hilberta funkcji całkowalnych z kwadratem na przedziale $I$ z iloczynem skalarnym postaci:

$$< \boldsymbol{u}(t), \boldsymbol{v}(t) >= \int_I \boldsymbol{u}'(t)\boldsymbol{v}(t)dt.$$

Ponadto załóżmy, że $k$-ta składowa procesu $\boldsymbol{X}(t)$ może być reprezentowana przez skończoną liczbę ortonormalnych funkcji bazowych $\{\varphi_b\}$

$$X_k(t) = \sum_{b=0}^{B_k} c_{kb}\varphi_b(t),\ t \in I,\ k = 1,2,\ldots,p,$$

gdzie $c_{kb}$ są zmiennymi losowymi takimi, że $E(c_{kb}) = 0$, $\mathrm{Var}(c_{kb}) < \infty$, $k = 1,2,\ldots,p$, $b = 0,\ldots,B_k$.

Niech

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

$$\boldsymbol{\Phi}(t) = \begin{bmatrix} \varphi_1'(t) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \varphi_2'(t) & \ldots & \mathbf{0} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{0} & \mathbf{0} & \ldots & \varphi_p'(t) \end{bmatrix}, \tag{3}$$

gdzie $\boldsymbol{\varphi}_k(t) = \left(\varphi_0(t),\ldots,\varphi_{B_k}(t)\right)'$, $k = 1,2,\ldots,p$.

Używając notacji macierzowej proces $\boldsymbol{X}(t)$ ma następującą reprezentację

$$\boldsymbol{X}(t) = \boldsymbol{\Phi}(t)\boldsymbol{c},\ t \in I,\ \mathrm{E}(\boldsymbol{c}) = 0,\ \mathrm{Var}(\boldsymbol{c}) = \Sigma_{\boldsymbol{c}}.$$

Poszukujemy zmiennej losowej

$$U = <\boldsymbol{u}(t), X(t)> = \int_I \boldsymbol{u}'(t)X(t)dt$$

mającej maksymalną wariancję dla wszystkich $\boldsymbol{u}(t) \in L_2^p(I)$ takich, że $<\boldsymbol{u}(t), \boldsymbol{u}(t)> = 1$. Możemy założyć, że wektor funkcji wagowych $\boldsymbol{u}(t)$ oraz proces $\boldsymbol{X}(t)$ należą do tej samej przestrzeni, tzn. funkcja $\boldsymbol{u}(t)$ może być przedstawiona w następującej postaci:

$$\boldsymbol{u}(t) = \boldsymbol{\Phi}(t)\boldsymbol{u},$$

gdzie $\boldsymbol{u} \in \mathbb{R}^{K+p}$, $K = B_1 + \ldots + B_p$. Wtedy

$$<\boldsymbol{u}(t), \boldsymbol{X}(t)> = <\boldsymbol{\Phi}(t)\boldsymbol{u}, \boldsymbol{\Phi}(t)\boldsymbol{c}> = \boldsymbol{u}' <\boldsymbol{\Phi}(t), \boldsymbol{\Phi}(t)> \boldsymbol{c} = \boldsymbol{u}'\boldsymbol{c}$$

oraz

$$\mathrm{E}(<\boldsymbol{u}(t), \boldsymbol{X}(t)>) = \boldsymbol{u}'E(\boldsymbol{c}) = \boldsymbol{u}'\boldsymbol{0} = 0,$$

$$\mathrm{Var}(<\boldsymbol{u}(t), \boldsymbol{X}(t)>) = \boldsymbol{u}'E(\boldsymbol{c}\boldsymbol{c}')\boldsymbol{u} = \boldsymbol{u}'\Sigma_{\boldsymbol{c}}\boldsymbol{u}.$$

Niech

$$\lambda_1 = \sup_{\boldsymbol{u}(t) \in L_2^p(I)} \mathrm{Var}(<\boldsymbol{u}(t), X(t)>) = \mathrm{Var}(<\boldsymbol{u}_1(t), X(t)>) = \boldsymbol{u}_1'\Sigma_c\boldsymbol{u}_1,$$

gdzie $<\boldsymbol{u}_1(t), \boldsymbol{u}_1(t)> = \boldsymbol{u}'_1\boldsymbol{u}_1 = 1$.

Zmienną losową $U_1 = <\boldsymbol{u}_1(t), \boldsymbol{X}(t)> = \boldsymbol{u}'_1\boldsymbol{c}$ nazywać będziemy pierwszą funkcjonalną składową główną, a funkcję wektorową $\boldsymbol{u}_1(t)$ pierwszym wektorem funkcji wagowych. Następnie szukamy drugiej funkcjonalnej składowej głównej $U_2 = <\boldsymbol{u}_2(t), \boldsymbol{X}(t)> = \boldsymbol{u}'_2\boldsymbol{c}$, maksymalizującej $\mathrm{Var}(<\boldsymbol{u}(t), \boldsymbol{X}(t)>) = \boldsymbol{u}'\Sigma_{\boldsymbol{c}}\boldsymbol{u}$ takiej, że $<\boldsymbol{u}_2(t), \boldsymbol{u}_2(t)> = \boldsymbol{u}'_2\boldsymbol{u}_2 = 1$ oraz nieskorelowanej z pierwszą funkcjonalną składową główną $U_1$, tzn. spełniającą warunek $<\boldsymbol{u}_1(t), \boldsymbol{u}_2(t)> = \boldsymbol{u}'_1\boldsymbol{u}_2 = 0$.

Ogólnie $k$-ta funkcjonalna składowa główna $U_k = <\boldsymbol{u}_k(t), \boldsymbol{X}(t)> = \boldsymbol{u}'_k\boldsymbol{c}$ spełnia warunki:

$$\lambda_k = \sup_{\boldsymbol{u}(t) \in L_2^p(I)} \mathrm{Var}(<\boldsymbol{u}(t), X(t)>) = \mathrm{Var}(<\boldsymbol{u}_k(t), X(t)>) = \boldsymbol{u}'_k\Sigma_c\boldsymbol{u}_k,$$

$$<\boldsymbol{u}_{\kappa_1}(t), \boldsymbol{u}_{\kappa_2}(t)> = \delta_{\kappa_1\kappa_2};\ \kappa_1,\kappa_2 = 1,\ldots,k.$$

Parę $(\lambda_k, \boldsymbol{u}_k(t))$ będziemy nazywać $k$-tym układem głównym procesu $\boldsymbol{X}(t)$.

Rozważmy teraz zagadnienie składowych głównych dla wektora losowego $\boldsymbol{c}$. $k$-ta składowa główna $U_k^* = < \boldsymbol{u}_k, \boldsymbol{c} >$ tego wektora spełnia warunki:

$$\gamma_k = \sup_{\boldsymbol{u} \in \mathbb{R}^{K+p}} Var(< \boldsymbol{u}, \boldsymbol{c} >) = \sup_{\boldsymbol{u} \in \mathbb{R}^{K+p}} \boldsymbol{u}' Var(\boldsymbol{c}) \boldsymbol{u} = \sup_{\boldsymbol{u} \in \mathbb{R}^{K+p}} \boldsymbol{u}' \Sigma_c \boldsymbol{u} = \boldsymbol{u}_k' \boldsymbol{\Sigma}_c \boldsymbol{u}_k,$$

$$\boldsymbol{u}_{\kappa_1}' \boldsymbol{u}_{\kappa_2} = \delta_{\kappa_1 \kappa_2},$$

gdzie $\kappa_1, \kappa_2 = 1,\ldots,k$, $K = B_1 + \ldots + B_p$. Parę $(\gamma_k, \boldsymbol{u}_k)$ nazywać będziemy $k$-tym układem głównym wektora $\boldsymbol{c}$.

Wyznaczenie $k$-tego układu głównego wektora $\boldsymbol{c}$ jest równoważne z wyznaczeniem wartości własnych oraz odpowiadających im wektorów własnych macierzy kowariancji $\Sigma_{\boldsymbol{c}}$ spełniających warunek $\boldsymbol{u}_{k_1}' \boldsymbol{u}_{k_2} = \delta_{k_1 k_2}$.

Z powyższych rozważań wynika następujące twierdzenie.

**Twierdzenie** *$k$-ty układ główny $(\lambda_k, \boldsymbol{u}_k(t))$ procesu stochastycznego $\boldsymbol{X}(t)$ jest związany z $k$-tym układem głównym $(\gamma_k, \boldsymbol{u}_k)$ wektora losowego $\boldsymbol{c}$ następującymi zależnościami:*

$$\lambda_k = \gamma_k, \ \boldsymbol{u}_k(t) = \boldsymbol{\Phi}(t) \boldsymbol{u}_k, \ t \in I,$$

gdzie $k = 1,\ldots, K + p$, $K = B_1 + B_2 + \ldots + B_p$.

Analiza składowych głównych dla wektora losowego $\boldsymbol{c}$ bazuje na macierzy $\Sigma_{\boldsymbol{c}}$. W praktyce macierz ta nie jest znana. Możemy ją oszacować na podstawie $n$ niezależnych realizacji $\boldsymbol{x}_1(t), \boldsymbol{x}_2(t),\ldots, \boldsymbol{x}_n(t)$ procesu losowego $\boldsymbol{X}(t)$.

W typowych sytuacjach dane pochodzą z obserwacji zmiennych w dyskretnych momentach czasowych. Proces transformacji takich danych dyskretnych do danych funkcjonalnych wykonujemy osobno dla każdej zmiennej $X_1, X_2,\ldots, X_p$.

Niech $x_{kj}$ oznacza obserwowaną wartość zmiennej $X_k$, $k = 1,2,\ldots,p$ w $j$-tym momencie czasowym $t_j$, gdzie $j = 1,2,\ldots,J$. Zatem dane składają się z $pJ$ par $(t_j, x_{kj})$. Te dane dyskretne wygładzamy za pomocą funkcji ciągłych $x_k(t)$, gdzie $t \in I$ (por. Ramsay, Silverman, 2005). Niech $I$ będzie zbiorem zwartym takim, że $t_j \in I$, dla $j = 1,\ldots,J$. Załóżmy, że funkcja $x_k(t)$ ma następującą reprezentację

$$x_k(t) = \sum_{b=0}^{B_k} c_{kb} \varphi_b(t), \ t \in I, \ k = 1,\ldots,p, \tag{4}$$

gdzie $\{\varphi_b\}$ są ortonormalnymi funkcjami bazowymi, a $c_{k0}, c_{k1},\ldots, c_{kB_k}$ są współczynnikami.

Niech $\boldsymbol{x}_k = (x_{k1}, x_{k2},\ldots, x_{kJ})'$, $\boldsymbol{c}_k = (c_{k0}, c_{k1},\ldots, c_{kB_k})'$ oraz $\boldsymbol{\Phi}_k(t)$ będzie macierzą wymiaru $J \times (B_k + 1)$ zawierającą wartości $\varphi_b(t_j)$, $b = 0,1,\ldots,B$, $j = 1,2,\ldots,J$, $k = 1,\ldots,p$. Współczynnik $\boldsymbol{c}_k$ we wzorze (4) jest oszacowany metodą najmniejszych kwadratów tak, aby minimalizował funkcję:

$$S(\boldsymbol{c}_k) = (\boldsymbol{x}_k - \boldsymbol{\Phi}_k(t)\boldsymbol{c}_k)'\,(\boldsymbol{x}_k - \boldsymbol{\Phi}_k(t)\boldsymbol{c}_k),\ k = 1,\ldots,p.$$

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

$$\hat{\boldsymbol{c}}_k = \left(\boldsymbol{\Phi}_k'(t)\boldsymbol{\Phi}_k(t)\right)^{-1}\boldsymbol{\Phi}_k'(t)\boldsymbol{x}_k,\ k = 1,\ldots,p.$$

Stopień gładkości funkcji $x_k(t)$ zależy od wartości $B_k$ (mała wartość $B_k$ oznacza większe wygładzenie krzywej). Optymalną wartość $B_k$ możemy ustalić przy pomocy bayesowskiego kryterium informacyjnego BIC (por. Schwarz, 1978; Shmueli, 2010).

Załóżmy, że dysponujemy $n$ niezależnymi parami wartości $(t_j, x_{kij})$, $k = 1,\ldots,p$, $i = 1,\ldots,n$, $j = 1,\ldots,J$. Dane te wygładzamy za pomocą funkcji ciągłych postaci:

$$x_{ki}(t) = \sum_{b=0}^{B_{ki}} \hat{c}_{kib}\varphi_b(t),\ k = 1,\ldots,p,\ i = 1,\ldots,n,\ t \in I.$$

Spośród wszystkich wartości $B_{k1}, B_{k2},\ldots, B_{kn}$ wybieramy jedną wspólną wartość $B_k$ jako modę z wartości $B_{k1}, B_{k2},\ldots, B_{kn}$ oraz zakładamy, że funkcje $x_{ki}(t)$ mają postać

$$x_{ki}(t) = \sum_{b=0}^{B_k} \hat{c}_{kib}\varphi_b(t),\ k = 1,\ldots,p,\ i = 1,\ldots,n,\ t \in I.$$

Dane postaci $\{x_{k1}(t),\ldots, x_{kn}(t)\}$ noszą nazwę danych funkcjonalnych (por. Ramsay, Silverman, 2005).

Ogólnie załóżmy, że $n$ niezależnych realizacji $\boldsymbol{x}_1(t), \boldsymbol{x}_2(t),\ldots, \boldsymbol{x}_n(t)$ może być przedstawionych w postaci $\boldsymbol{x}_i(t) = \boldsymbol{\Phi}(t)\hat{\boldsymbol{c}}_i$ gdzie $\boldsymbol{\Phi}(t)$ dane jest wzorem (3) oraz że wektory $\hat{\boldsymbol{c}}_i = (\hat{c}_{10},\ldots, \hat{c}_{1B_1},\ldots,\hat{c}_{p0},\ldots, \hat{c}_{pB_p})'$ są scentrowane, $i = 1,2,\ldots,n$.

Oznaczmy $\hat{\boldsymbol{C}} = (\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2,\ldots,\hat{\boldsymbol{c}}_n)$. Wtedy

$$\hat{\boldsymbol{\Sigma}}_c = \frac{1}{n}\hat{C}\hat{C}'.$$

Niech $\hat{\gamma}_1 \geqslant \hat{\gamma}_2 \geqslant \ldots \geqslant \hat{\gamma}_s$ będą niezerowymi wartościami własnymi macierzy $\hat{\boldsymbol{\Sigma}}_c$, a $\hat{\boldsymbol{u}}_1,\hat{\boldsymbol{u}}_2,\ldots,\hat{\boldsymbol{u}}_s$ odpowiadającymi im wektorami własnymi, gdzie s = rank($\hat{\boldsymbol{\Sigma}}_c$).

Ponadto $k$-ty układ główny procesu losowego $\boldsymbol{X}(t)$ wyznaczony na podstawie próby ma następującą postać:

$$\left(\hat{\lambda}_k = \hat{\gamma}_k, \hat{\boldsymbol{u}}_k(t) = \boldsymbol{\Phi}(t)\hat{\boldsymbol{u}}_k\right),\ k = 1,\ldots,s.$$

Współrzędne rzutu $i$-tej realizacji $\boldsymbol{x}_i(t)$ procesu $\boldsymbol{X}(t)$ na kierunek wyznaczony przez $k$-tą funkcjonalna składową główną są równe:

$$\hat{U}_{ik} = < \hat{\boldsymbol{u}}_k(t), \boldsymbol{x}_i(t) > = < \boldsymbol{\Phi}(t)\hat{\boldsymbol{u}}_k, \boldsymbol{\Phi}(t)\hat{\boldsymbol{c}}_i > = \hat{\boldsymbol{u}}'_k < \boldsymbol{\Phi}(t), \boldsymbol{\Phi}(t) > \hat{\boldsymbol{c}}_i = \hat{\boldsymbol{u}}'_k \hat{\boldsymbol{c}}_i,$$

dla $i = 1,2,\ldots,n$, $k = 1,2,\ldots,s$. Ogólnie współrzędne rzutu $i$-tej realizacji $\boldsymbol{x}_i(t)$ procesu $\boldsymbol{X}(t)$ na płaszczyznę wyznaczoną przez dwie pierwsze funkcjonalne składowe główne są równe:

$$(\hat{\boldsymbol{u}}'_1 \hat{\boldsymbol{c}}_i, \hat{\boldsymbol{u}}'_2 \hat{\boldsymbol{c}}_i),\ i = 1,2,\ldots,n.$$

## 4. DENDRYTOWA ANALIZA SKUPIEŃ

Omówimy teraz krótko metodę analizy skupień bazującą na dendrycie. Z punktu widzenia teorii grafów dendryt jest synonimem drzewa. Dendryt lub drzewo jest grafem spinającym (każde dwa wierzchołki łączy jakaś droga), który nie zawiera cykli. Minimalny dendryt, to taki dendryt, w którym suma wag przy krawędziach jest minimalna. Najczęściej wagami są odległości. Konstrukcja dendrytu podana została przez grupę matematyków wrocławskich w pracy Florek i inni (1951a). Metoda ta znana pod nazwą *taksonomia wrocławska* została spopularyzowana w pracach Florek inni (1951b) oraz Perkal (1953). Istnieją dwie niezwykle proste procedury postępowania pozwalające skonstruować minimalny dendryt. Pierwszą z nich jest algorytm Kruskala:

1. Wybieramy krawędź o najmniejszej długości.
2. Z pozostałych krawędzi wybieramy tę o najmniejszej długości, która nie prowadzi do cyklu (z połączeń o jednakowych długościach wybieramy dowolne).
3. Powtarzamy poprzedni krok do zakończenia budowy minimalnego dendrytu.

Algorytm Prima rozpoczyna od drzewa składającego się z jednego wierzchołka, ciągle dodając najkrótszą krawędź drzewa.

Najkrótszy dendryt można wykorzystać w analizie skupień. Idea polega na usunięciu z minimalnego dendrytu wszystkich krawędzi, których długość jest większa od wspólnej wartości krytycznej $d$. Wyliczamy wartość średnią $\bar{x}$ oraz odchylenie standardowe $s$ z długości wszystkich krawędzi najkrótszego dendrytu i przyjmujemy $d = \bar{x} + s$. Wierzchołki, które pozostają połączone w minimalnym dendrycie tworzą skupienie. Zaprezentowana tutaj metoda jest szerzej omówiona w sekcji 10.3 podręcznika Dziechciarza (2003).

## 5. WYNIKI BADAŃ EMPIRYCZNYCH

Analizą objęto 16 województw Polski ($n = 16$). Na prezentowanych dalej wykresach poszczególne województwa są oznaczone numerami przedstawionymi w tabeli 1.

Analizowane dane obejmują okres 11 lat, od 2003 do 2013 roku ($J = 11$). Każde województwo scharakteryzowano za pomocą 37 cech, zgrupowanych w 6 podzbiorach obrazujących różnorodne aspekty życia mieszkańców danego regionu, jak:

– wynagrodzenia i rynek pracy;

– opieka zdrowotna i społeczna;
– infrastruktura gospodarcza;
– zasoby mieszkaniowe;
– oświata, kultura i rekreacja oraz
– bezpieczeństwo i środowisko.

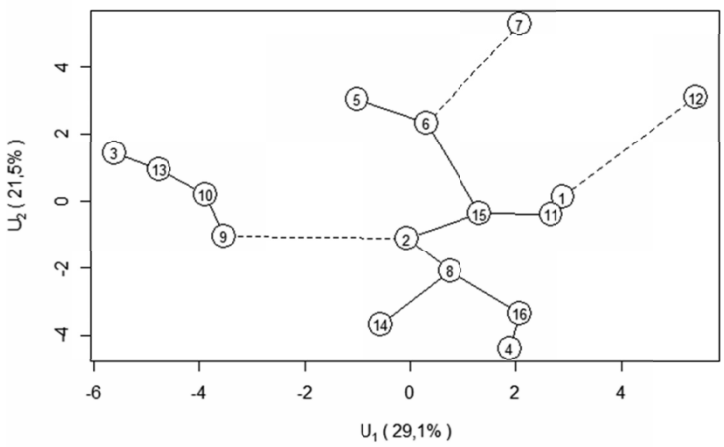
Tabela 1.

Numeryczne oznaczenia województw

| Nr | Województwo | Nr | Województwo |
|---|---|---|---|
| 1 | dolnośląskie | 9 | podkarpackie |
| 2 | kujawsko-pomorskie | 10 | podlaskie |
| 3 | lubelskie | 11 | pomorskie |
| 4 | lubuskie | 12 | śląskie |
| 5 | łódzkie | 13 | świętokrzyskie |
| 6 | małopolskie | 14 | warmińsko-mazurskie |
| 7 | mazowieckie | 15 | wielkopolskie |
| 8 | opolskie | 16 | zachodniopomorskie |

Źródło: opracowanie własne.

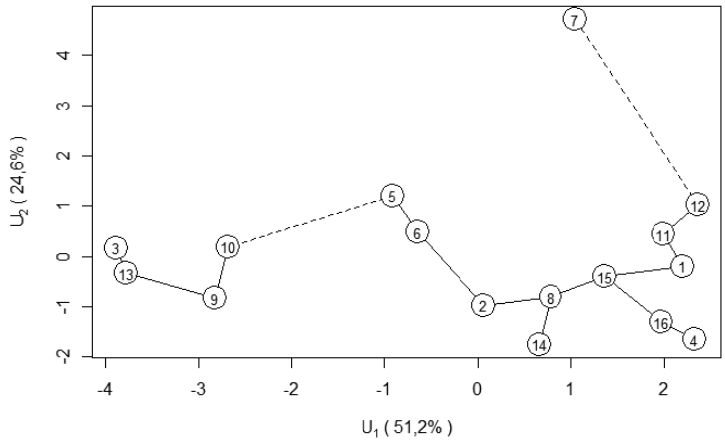
Dane pierwotne zostały poddane unitaryzacji zerowanej (patrz sekcja 2), a następnie przekształcone do wielowymiarowych danych funkcjonalnych. Posłużono się funkcjami bazowymi Fouriera. Przedział czasowy $I = [0,11]$ został podzielony na momenty czasowe następująco: $t_1 = 0{,}5$ (2003), $t_2 = 1{,}5$ (2004),…, $t_{11} = 10{,}5$ (2013). Następnie dla wszystkich 37 cech łącznie oraz oddzielnie dla każdej z 6 grup cech zostały skonstruowane funkcjonalne składowe główne. Każde z 16 województw zostało przedstawione jako punkt w układzie dwóch pierwszych funkcjonalnych składowych głównych. Kolejno dla wszystkich grup cech zostały zbudowane dendryty (por. Florek i inni, 1951) na podstawie tablicy wzajemnych odległości euklidesowych między województwami w przestrzeni wszystkich funkcjonalnych składowych głównych. Odległości te są równe odległościom euklidesowym w oryginalnej przestrzeni danych funkcjonalnych. Dendryty te rozpięto na punktach reprezentujących poszczególne województwa i wykorzystano do analizy skupień. W każdym z dendrytów policzono wartość średnią $\bar{x}$ długości krawędzi oraz odchylenie standardowe $s$ tych długości. Krawędzie, których długość była większa od $d = \bar{x} + s$ zaznaczone są na rysunkach linią przerywaną. W ten sposób uzyskano podział województw na względnie jednorodne skupienia.




Rysunek 1. Województwa w układzie dwóch pierwszych funkcjonalnych składowych głównych – grupowanie na podstawie wszystkich 37 cech
Źródło: opracowanie własne.

Analiza przeprowadzona w oparciu o wszystkie 37 cech diagnostycznych pozwoliła wyodrębnić cztery grupy województw względnie jednorodnych pod względem poziomu życia mieszkańców. Dwie z nich to grupy jednoelementowe: woj. śląskie i woj. mazowieckie. Kolejną grupę utworzyły województwa: dolnośląskie, pomorskie, wielkopolskie, zachodniopomorskie, lubuskie, opolskie, małopolskie, kujawsko-pomorskie, łódzkie i warmińsko-mazurskie. Grupę czwartą utworzyły cztery województwa ściany wschodniej, często określane mianem „wschodniej ściany płaczu": podkarpackie, podlaskie, świętokrzyskie i lubelskie. Województwa te są najuboższymi regionami Polski, a do momentu wejścia w struktury UE Bułgarii i Rumunii, były też uznawane za najuboższe w całej Wspólnocie. Niepokojącym wydaje się także fakt, iż w świetle badań prowadzonych przez OECD we współpracy z Ministerstwem Rozwoju Regionalnego a także danych z ostatniego Spisu Powszechnego, różnice w rozwoju gospodarczym i społecznym polskich regionów pogłębiają się. Z raportu „Przegląd Regionalny Polski 2012" opracowanego przez Ministerstwo Rozwoju Regionalnego wynika, iż mimo długoletnich działań oraz liczonej w miliardach euro pomocy z Unii Europejskiej przepaść między Polską wschodnią a resztą kraju niebezpiecznie rośnie. „Najwolniej dystans do średniej unijnej nadrabiały województwa o najniższym poziomie PKB per capita (województwa Polski Wschodniej i woj. zachodniopomorskie). Pięć województw Polski Wschodniej nadal znajdowało się w grupie 20 europejskich regionów NUTS 2 o najniższym poziomie PKB per capita, z poziomem tego wskaźnika w relacji do średniej unijnej od 42% (lubelskie) do 47% (świętokrzyskie)".

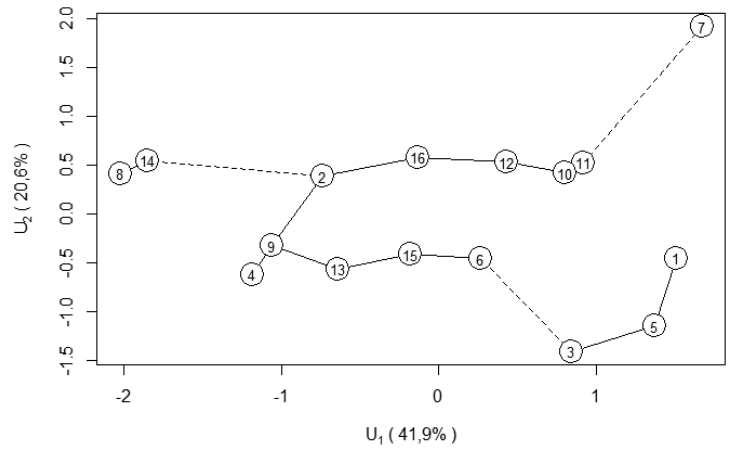
Wykorzystując podzbiory cech diagnostycznych wyodrębniono grupy województw względnie jednorodnych pod względem poszczególnych aspektów kształtujących ogólny poziom życia mieszkańców. Uzyskane wyniki zaprezentowano na rysunkach 2–7.




Rysunek 2. Województwa w układzie dwóch pierwszych funkcjonalnych składowych głównych – grupowanie na podstawie cech opisujących „wynagrodzenie i rynek pracy"
Źródło: opracowanie własne.

Cechy charakteryzujące wynagrodzenia i rynek pracy podzieliły województwa Polski na trzy grupy. Jedną z nich tworzą województwa: lubelskie, świętokrzyskie, podkarpackie i podlaskie. Województwa te, będące typowo rolniczym regionem naszego kraju, charakteryzuje relatywnie wysoki odsetek zatrudnionych w rolnictwie, mała konkurencyjność gospodarki, niski poziom bezpośrednich inwestycji zagranicznych w skali kraju – co niewątpliwie wpływa na bardzo niski poziom dochodów ludności oraz relatywnie wysokie bezrobocie. Drugą grupę, względnie jednorodną pod względem poziomu wynagrodzeń i sytuacji na rynku pracy, utworzyło 11 województw. Trzecią grupą stanowi województwo mazowieckie, w którym przeciętne miesięczne wynagrodzenia brutto w analizowanych 11 latach wahały się pomiędzy 129,8% (w roku 2003) a 123,1% (w roku 2013) średniej krajowej. Woj. mazowieckie jest jednym z dwóch województw Polski, w których przeciętne miesięczne wynagrodzenia przekraczają średnią krajową, przy czym dystans wynagrodzeń w drugim, tj. w woj. śląskim do średniej krajowej jest zdecydowanie mniejszy. W mazowieckim obserwuje się też relatywnie niski poziom bezrobocia.

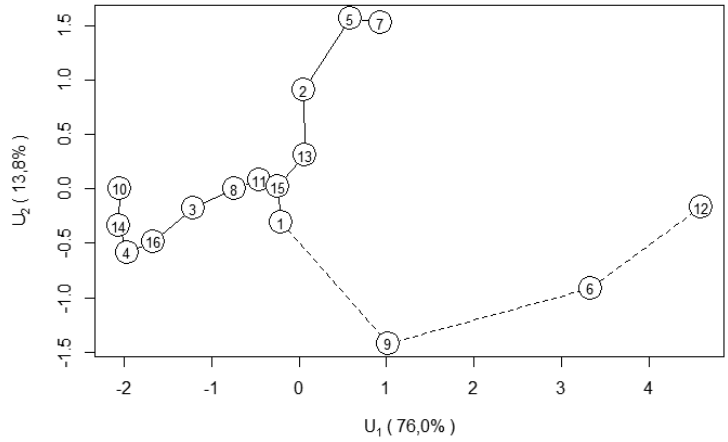
Pod względem „opieki zdrowotnej i społecznej" województwo mazowieckie – po raz kolejny – nie weszło w skład żadnej z grup względnie jednorodnych utworzonych z pozostałych województw. Wiele „przewag" województwa mazowieckiego wynika niewątpliwie z faktu, iż jest to region ze stolicą kraju, mamy tu zlokalizowaną dużą bazę akademicką z największą uczelnią medyczną w Polsce, dużą liczbę szpitali i lekarzy (przykładowo w roku 2010 na terenie tego województwa działało 14% ogółu szpitali publicznych i 12% ogółu szpitali niepublicznych, pracowało tu blisko 30 tys. (tj. 17%) spośród 172 tys. pracujących lekarzy w Polsce).




Rysunek 3. Województwa w układzie dwóch pierwszych funkcjonalnych składowych głównych – grupowanie na podstawie cech opisujących „opiekę zdrowotną i społeczną"

Źródło: opracowanie własne.

Grupy względnie jednorodne pod względem poziomu opieki zdrowotnej i społecznej utworzyły województwa: łódzkie z dolnośląskim i lubelskim, opolskie z warmińsko-mazurskim i grupę trzecią – dziesięć pozostałych województw.

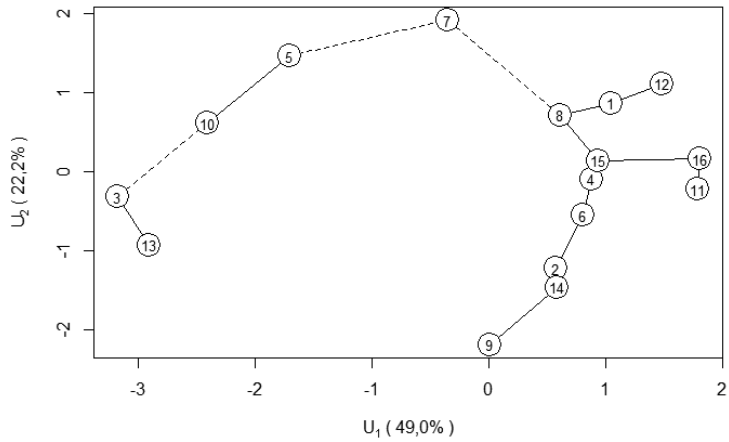



Rysunek 4. Województwa w układzie dwóch pierwszych funkcjonalnych składowych głównych – grupowanie na podstawie cech opisujących „infrastrukturę gospodarczą"

Źródło: opracowanie własne.

Aż 13. spośród szesnastu województw utworzyło grupę względnie jednorodną pod względem wyposażenia i stanu infrastruktury gospodarczej. Poza wspomnianą grupą znalazły się – postrzegane jako najzasobniejsze pod względem infrastruktury technicznej – województwa śląskie i małopolskie (usytuowane wzdłuż międzynarodo-

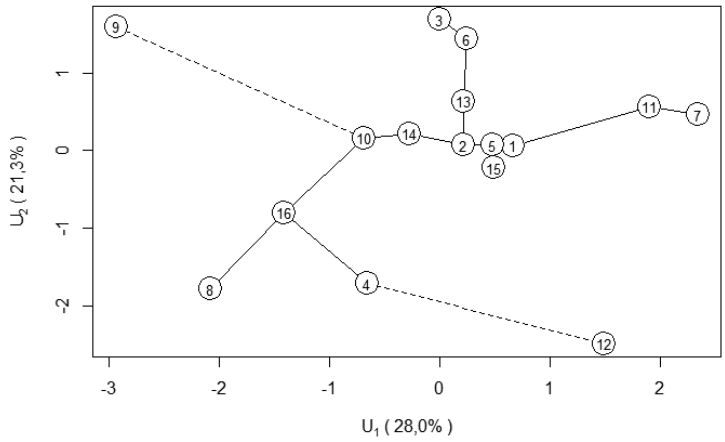
wych szlaków komunikacyjnych i transportowych, z największym w kraju zagęszczeniem sieci drogowej i kolejowej oraz wodno-kanalizacyjnej) oraz podkarpackie, które cechuje słaba dostępność komunikacyjna, wynikająca z niedostatecznej i niskiej jakości sieci drogowej i kolejowej, niskie tempo rozwoju infrastruktury teleinformatycznej (zwłaszcza szerokopasmowego dostępu do Internetu) oraz wymagający rozbudowy i modernizacji stan infrastruktury komunalnej i energetycznej.



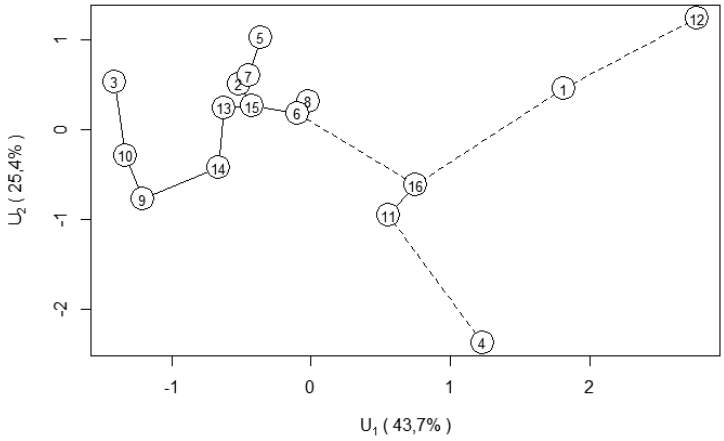

Rysunek 5. Województwa w układzie dwóch pierwszych funkcjonalnych składowych głównych – grupowanie na podstawie cech opisujących „zasoby mieszkaniowe”
Źródło: opracowanie własne.

Analiza przeprowadzona w oparciu o cechy obrazujące stan zasobów mieszkaniowych w latach 2003–2013 pozwoliła na wyodrębnienie czterech grup województw o zbliżonym poziomie rozpatrywanych cech. Jedną z tych grup, tworzy (po raz kolejny „jednoosobowo”) województwo mazowieckie, w którym notuje się największą liczbę mieszkań na 1 tys. mieszkańców i największą powierzchnię użytkową mieszkania w przeliczeniu na 1 osobę.

W wyniku analizy przeprowadzonej w oparciu o cechy opisujące wyposażenie i stan „oświaty, kultury i rekreacji” w poszczególnych województwach wyodrębniono grupę względnie jednorodną w skład, której weszło aż 14 województw Polski. Poza nią znalazły się jedynie województwa śląskie i podkarpackie. „Oddalenie” województwa podkarpackiego od pozostałych wynika z faktu, iż mamy tu do czynienia z relatywnie największą liczbą klubów sportowych w przeliczeniu na 10 tys. mieszkańców ale też z jednym z najniższych – w skali kraju – odsetkiem dzieci objętych wychowaniem przedszkolnym. W śląskim, z kolei, mamy najsłabsze wyposażenie szkół w komputery z dostępem do Internetu oraz relatywnie małą liczbę klubów sportowych oraz domów, klubów i ośrodków kultury w przeliczeniu na liczbę mieszkańców.




Rysunek 6. Województwa w układzie dwóch pierwszych funkcjonalnych składowych głównych – grupowanie na podstawie cech opisujących „oświatę, kulturę i rekreację"
Źródło: opracowanie własne.




Rysunek 7. Województwa w układzie dwóch pierwszych funkcjonalnych składowych głównych – grupowanie na podstawie cech opisujących „bezpieczeństwo i środowisko"
Źródło: opracowanie własne.

Grupowanie w oparciu o cechy opisujące „bezpieczeństwo i środowisko" doprowadziło do wyodrębnienia dwóch grup względnie jednorodnych, z których jedna liczyła jedenaście, a druga dwa województwa. Poziom rozpatrywanych cech w pozostałych województwach: śląskim, dolnośląskim i lubuskim, wyłączył je poza grupy względnie jednorodne. W każdym z tych trzech województw notuje się relatywnie wysoką przestępczość, w śląskim i dolnośląskim – najwyższe w skali kraju ilości wytwarzanych odpadów na 1 km$^2$ (lubuskie pod tym względem plasuje się na ostatnich pozycjach w rankingu województw). „Odmienność" lubuskiego jest także wynikiem

dużego zalesienia tego województwa – udział lasów w całkowitej powierzchni jest tu najwyższy w Polsce i przekracza 50% (podczas, gdy w kolejnym pod tym względem, woj. podkarpackim wynosi ok. 38%).

Podsumowując, warto zauważyć, że pozycję województwa śląskiego, w którym poziom życia mieszkańców w latach 2003–2013 był relatywnie najwyższy (rys. 1) determinowały przede wszystkim: wyposażenie w infrastrukturę gospodarczą, wysoki poziom wynagrodzeń, dobra sytuacja na rynku pracy oraz posiadane zasoby mieszkaniowe. O położeniu woj. mazowieckiego zdecydowały najwyższy w Polsce poziom wynagrodzeń, dobra sytuacja na rynku pracy, sytuacja w zakresie opieki zdrowotnej i społecznej oraz oświaty, kultury i rekreacji. Z kolei o położeniu czwartej, wyodrębnionej w oparciu o 37 cech diagnostycznych, grupy województw o relatywnie najniższym poziomie życia mieszkańców, tj. grupy woj. ściany wschodniej, zdecydowały: najniższe z obserwowanych w kraju wynagrodzenia, trudna sytuacja na lokalnych rynkach pracy oraz słabe zasoby mieszkaniowe.

## LITERATURA

Diechciarz J., (red.), (2003), *Ekonometria. Metody, przykłady, zadania*, Wydanie 2 poprawione, Wydawnictwo AE we Wrocławiu.

Florek K., Łukaszewicz J., Perkal J., Steinhaus H., Zubrzycki S., (1951a), Sur la Liaison et la Division des Points d'un Ensemble Fini, *Colloquium Mathematicum*, 2, 282–285.

Florek K., Łukaszewicz J., Perkal J., Steinhaus H., Zubrzycki S., (1951b), Taksonomia wrocławska, *Przegląd Antropologiczny*, 17, 193–211.

Górecki T., Krzyśko M., Waszak Ł., Wołyński W., (2014), Methods of Reducing Dimension for Functional Data, *Statistics in Transition – New Series*, 15 (2), 231–242.

Horváth L., Kokoszka P., (2012), *Inference for Functional Data with Applications*, Springer.

Jacques J., Preda C., (2014), Model-Based Clustering for Multivariate Functional Data, *Computational Statistics & Data Analysis*, 71, 92–106.

Kruskal J. B., (1956), On the Shortest Spanning Subtree of a Graph and the Travelling Salesman Problem, *Proceedings of the* American Mathematical Society, 7, 48–50.

Perkal J., (1953), Taksonomia wrocławska, *Przegląd Antropologiczny*, 19, 209–221.

Prim R. C., (1957), Shortest Connection Networks and Some Generalizations, *Bell System Technical Journal*, 36, 1389–1401.

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

Schwarz G., (1978), Estimating the Dimension of a Model, *Annals of Statistics*, 6, 461–464.

Shmueli G., (2010), To Explain or to Predict? *Statistical Science*, 25 (3), 289–310.

Walesiak M., (2014), Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej, *Przegląd Statystyczny*, 61 (4), 363–372.

## OCENA ZRÓŻNICOWANIA POZIOMU ŻYCIA MIESZKAŃCÓW WOJEWÓDZTW W LATACH 2003–2013 ZA POMOCĄ SKŁADOWYCH GŁÓWNYCH DLA WIELOWYMIAROWYCH DANYCH FUNKCJONALNYCH ORAZ ANALIZY SKUPIEŃ

Streszczenie

W artykule przedstawiono ocenę zróżnicowania poziomu życia mieszkańców województw w latach 2003–2013. Do oceny zastosowano analizę składowych głównych dla wielowymiarowych danych funkcjonalnych oraz dendrytową analizę skupień. Metody te pozwoliły na wyodrębnienie względnie jednorodnych grup województw o zbliżonym poziomie rozpatrywanych cech dla całego rozpatrywanego okresu łącznie.

**Słowa kluczowe:** wielowymiarowe dane funkcjonalne, funkcjonalna analiza danych, analiza składowych głównych

## ESTIMATION OF DIVERSITY OF LIVING STANDARDS IN POLISH VOIVODSHIPS IN 2003–2013 USING PRINCIPAL COMPONENTS FOR MULTIDIMENSIONAL FUNCTIONAL DATA AND CLUSTER ANALYSIS

Abstract

The paper presents an estimation of life standard diversity for residents of Polish voivodships in 2003–2013. The principal component analysis was applied for multidimensional functional data and the dendrite method was used for cluster analysis. These methods made it possible to isolate relatively homogeneous groups of voivodships that had similar values of characteristics under consideration, for the whole period at issue.

**Keywords:** multivariate functional data, functional data analysis, principal components analysis