PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS **nr 507 • 2018**

Taksonomia 30 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Mirosław Krzyśko*, Waldemar Wołyński, Wojciech Łukaszonek*, Waldemar Ratajczak****

* Państwowa Wyższa Szkoła Zawodowa w Kaliszu
e-mails: mkrzysko@amu.edu.pl; w.lukaszonek@g.pl

** Uniwersytet im. Adama Mickiewicza w Poznaniu
e-mails: wolynski@amu.edu.pl; walrat@amu.edu.pl

ANALIZA SKŁADOWYCH GŁÓWNYCH DLA DANYCH CZASOWO-PRZESTRZENNYCH

PRINCIPAL COMPONENT ANALYSIS FOR TEMPORAL-SPATIAL DATA

DOI: 10.15611/pn.2018.507.11
JEL Classification: C38

Streszczenie: W pracach [Górecki i in. 2014; 2016] przedstawiono konstrukcję składowych głównych dla wielowymiarowych danych zmiennych w czasie (wielowymiarowych danych funkcjonalnych). W pracach [Harris i in. 2011] oraz [Lu i in. 2014] podano konstrukcję składowych głównych dla wielowymiarowych danych przestrzennych. Są to składowe główne lokalne, geograficznie ważone. W pracy tej przedstawiona jest konstrukcja składowych głównych dla wielowymiarowych danych czasowo-przestrzennych, łącząca wyniki przywołanych prac.

Słowa kluczowe: składowe główne, dane czasowo-przestrzenne.

Summary: In Górecki et. al [2014; 2016] the structure of principal components for multidimensional data variables over time (multivariate functional data) are presented. Harris et. al [2011] and Lu et. al [2014] provide the construction of principal components for multivariate spatial data. These are the local principal components, geographically weighted. This paper presents the construction of principal components for multivariate temporal-spatial data, combining the results of the mentioned papers.

Keywords: principal components, temporal-spatial data.

1. Wstęp

Inspiracją do powstania tego artykułu były dwie prace [Harris i in. 2011] oraz [Lu i in. 2014]. Prace te dotyczą analizy składowych głównych w przypadku obserwacji wektorowych w ustalonym momencie czasu. Standardową analizę składowych

głównych autorzy nazywają analizą globalną. Obok analizy globalnej wprowadzili pojęcie analizy lokalnej. Zakładają, że znane są współrzędne geograficzne miejsca obserwacji wartości p-cechowego wektora losowego. Dla każdego z n miejsc obserwacji wykonywana jest oddzielnie lokalna analiza składowych głównych polegająca na tym, że zamiast klasycznej macierzy kowariancji budowana jest ważona macierz kowariancji. Wagami są wartości pewnych funkcji odległości między obserwowanymi miejscami.

Wyniki uzyskane przez wspomnianych autorów zostaną uogólnione na przypadek wielowymiarowych obserwacji dokonywanych w pewnym przedziale czasowym, tj. na przypadek wielowymiarowych danych funkcjonalnych. Globalna analiza składowych głównych dla takich danych została opisana w pracach [Górecki i in. 2014; 2016].

Autorzy niniejszej pracy uważają, że globalna oraz lokalne analizy składowych głównych nie są spójne, ponieważ w analizie lokalnej dane są ważone, natomiast nie są ważone dane w analizie globalnej. Stąd proponujemy, by również w analizie globalnej stosować procedurę ważenia. W pracy wagi wyznaczane są na bazie współczynników dostępności komunikacyjnej [Górniak 2015].

2. Analiza globalna

Załóżmy, że obserwujemy p-wymiarowy proces stochastyczny $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$. Dalej załóżmy, że $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{0}$ i $\boldsymbol{X} \in L_2^p(I)$, gdzie $L_2(I)$ jest przestrzenią Hilberta funkcji całkowalnych z kwadratem na przedziale I z iloczynem skalarnym postaci:

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

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

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

gdzie c_{kb} są zmiennymi losowymi takimi, że $\mathrm{E}(c_{kb}) = 0,\ \mathrm{Var}(c_{kb}) < \infty$.

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

$$\boldsymbol{\Phi}(t) = \begin{bmatrix} \boldsymbol{\varphi}_1'(t) & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\varphi}_2'(t) & \ldots & \boldsymbol{0} \\ \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{\varphi}_p'(t) \end{bmatrix}, \tag{1}$$

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

Używając notacji macierzowej, proces $\boldsymbol{X}$ ma następującą reprezentację:

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

Globalna analiza składowych głównych bazuje na macierzy $\boldsymbol{\Sigma_c}$. W praktyce macierz ta nie jest znana. Możemy ją oszacować na podstawie n niezależnych realizacji $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ procesu losowego $\boldsymbol{X}$.

Załóżmy, że n niezależnych realizacji $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ może być przedstawionych w postaci $\boldsymbol{x}_i(t) = \boldsymbol{\Phi}(t)\hat{\boldsymbol{c}}_i$, gdzie $\boldsymbol{\Phi}(t)$ dane jest wzorem (1) oraz że wektory $\hat{\boldsymbol{c}}_i$, $i = 1, 2, \ldots, n$ są scentrowane.

Oznaczmy $\hat{\boldsymbol{C}} = (\hat{\boldsymbol{c}}_1, \hat{\boldsymbol{c}}_2, \ldots, \hat{\boldsymbol{c}}_n)$. Wtedy $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{c}} = \frac{1}{n}\hat{\boldsymbol{C}}\hat{\boldsymbol{C}}'$.

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

Wówczas k-ta funkcjonalna składowa główna procesu $\boldsymbol{X}$ ma postać

$$U_k = < \hat{\boldsymbol{u}}_k, \boldsymbol{X} >,$$

gdzie funkcje wagowe $\hat{\boldsymbol{u}}_k$ dane są wzorem

$$\hat{\boldsymbol{u}}_k(t) = \boldsymbol{\Phi}(t)\hat{\boldsymbol{d}}_k,\ k = 1, \ldots, s.$$

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

$$\hat{U}_{ik} = < \hat{\boldsymbol{u}}_k, \boldsymbol{x}_i > = \hat{\boldsymbol{d}}_k'\hat{\boldsymbol{c}}_i,$$

dla $i = 1, 2, \ldots, n, k = 1, 2, \ldots, s$.

W szczególności współrzędne rzutu i-tej realizacji $\boldsymbol{x}_i$ procesu $\boldsymbol{X}$ na płaszczyznę wyznaczoną przez dwie pierwsze funkcjonalne składowe główne są równe:

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

Jeżeli każdy z wektorów $\hat{\boldsymbol{c}}_i$ pomnożymy przez odpowiednią wagę wyznaczoną na postawie współczynników dostępności komunikacyjnej i dalej postąpimy tak, jak poprzednio, to otrzymamy ważone globalne składowe główne.

Współczynnik dostępności komunikacyjnej d_j przekształcamy w wagi w_j według wzoru:

$$w_j = \frac{d_j}{\sum_{i=1}^{n} d_i},\ j = 1, \ldots, n.$$

3. Lokalna analiza składowych głównych

Niech $\hat{\boldsymbol{C}} = (\hat{\boldsymbol{c}}_1, \hat{\boldsymbol{c}}_2, \ldots, \hat{\boldsymbol{c}}_n)$ będzie macierzą danych. Ustalmy miejsce obserwacji, powiedzmy k. Niech d_{kj} będzie odległością między k-tym oraz j-tym miejscem obserwacji, $j = 1, 2, \ldots, n$.

Odległość d_{kj} ustalamy według następującej zasady sąsiedztwa. Niech d_k będzie promieniem okręgu, którego środek znajduje się w k-tym wyróżnionym punkcie (np. w mieście wojewódzkim). Promień d_k zmienia się w zależności od tego – czy obok-ległe obszary (np. województwa) sąsiadują bezpośrednio, pośrednio przez jedną krawędź, pośrednio przez dwie krawędzie itd., z obszarem, na którym leży wyróż-niony punkt k. Oznacza to, że wszystkie punkty j, leżące na powierzchni koła o promieniu d_k, będą posiadały tę samą odległość d_{kj} od wyróżnionego punktu k.

Odległość d_{kj} przekształcamy w wagę w_{kj} związaną z k-tym miejscem obserwacji, według wzoru:

$$w_{kj} = \exp(-\lambda_k d_{kj}^2),\ \lambda_k > 0.$$

Przyjmujemy

$$\lambda_k = \frac{1}{\left(\frac{1}{n-1}\sum_{j=1}^{n} d_{kj}^2\right)}.$$

Wektory $\hat{\boldsymbol{c}}_i$ mnożymy przez $\sqrt{w_{ki}}$, a następnie centrujemy. Przez $\boldsymbol{C}_k^{\star}$ oznaczmy macierz złożoną z otrzymanych wektorów. Wówczas estymator macierzy kowariancji dla miejsca k ma postać:

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{c}k} = \frac{1}{n}\boldsymbol{C}_k^{\star}\boldsymbol{C}_k^{\star\prime} = \frac{1}{n}\hat{\boldsymbol{C}}\boldsymbol{W}_k\hat{\boldsymbol{C}}',$$

gdzie $\boldsymbol{W}_k = \mathrm{diag}(w_{k1}, w_{k2}, \ldots, w_{kn}), k = 1, 2, \ldots, n.$

Dla każdego miejsca k macierz $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{c}k}$ jest podstawą konstrukcji lokalnych składowych głównych. Jeżeli wagi w_{kj} zastąpimy wagami wyznaczonymi na podstawie współczynników dostępności komunikacyjnej poszczególnych miejsc obserwacji, to otrzymamy jedno rozwiązanie globalne ważone.

4. Przykład

Badanie obejmowało 15 lat (2002-2016) oraz 7 zmiennych charakteryzujących stan szkolnictwa wyższego na poziomie województw. Uwzględnione cechy to:

- X1 – liczba uczelni,
- X2 – liczba studentów,
- X3 – liczba absolwentów,
- X4 – liczba nauczycieli akademickich,
- X5 – liczba nauczycieli akademickich z tytułem profesora,
- X6 – liczba studentów studiów podyplomowych,
- X7 – liczba studentów studiów doktoranckich.

Dane pochodzą z Banku Danych Lokalnych i dotyczą n = 16 polskich województw. W tabeli 3 oraz na rysunkach województwom przypisano oznaczenia jak w tab. 1.

Tabela 1. Współczynniki dostępności komunikacyjnej

1	Dolnośląskie	285	9	Podkarpackie	131
2	Kujawsko-pomorskie	327	10	Podlaskie	144
3	Lubelskie	108	11	Pomorskie	221
4	Lubuskie	88	12	Śląskie	418
5	Łódzkie	254	13	Świętokrzyskie	189
6	Małopolskie	345	14	Warmińsko-mazurskie	168
7	Mazowieckie	367	15	Wielkopolskie	238
8	Opolskie	91	16	Zachodniopomorskie	93

Źródło: opracowanie własne na podstawie [Górniak 2015].

Tabela 2. Wagi w_j wyznaczone na podstawie współczynników dostępności komunikacyjnej

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
0,0822	0,0943	0,0321	0,0254	0,0733	0,0995	0,1059	0,0262
w_9	w_{10}	w_{11}	w_{12}	w_{13}	w_{14}	w_{15}	w_{16}
0,0378	0,0415	0,0637	0,1206	0,0545	0,0485	0,0686	0,0268

Źródło: opracowanie własne.

Tabela 3. Odległości pomiędzy województwami wyznaczone według zasady sąsiedztwa

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0	213	391	96	213	298	298	96	391	391	298	213	298	298	96	298
2	288	0	288	288	114	419	114	288	419	288	114	288	288	114	114	288
3	359	256	0	479	256	256	128	359	128	128	359	256	128	256	359	479
4	135	231	451	0	231	451	327	231	451	451	231	327	327	327	135	135
5	263	99	263	263	0	263	99	99	263	263	263	99	99	263	99	263
6	305	305	185	447	185	0	185	185	85	305	447	85	85	305	305	447
7	366	139	139	366	139	256	0	256	256	139	256	256	139	139	256	366
8	89	234	351	234	89	234	234	0	351	351	351	89	234	351	89	351
9	497	348	96	497	245	96	245	348	0	245	497	245	96	348	348	497
10	483	298	142	483	298	398	142	398	298	0	298	398	298	142	398	398
11	359	153	426	359	359	515	359	359	515	359	0	426	426	153	153	153
12	227	227	227	355	96	96	227	96	227	355	355	0	96	355	227	355
13	387	252	82	387	82	82	82	252	82	252	387	82	0	252	252	387
14	373	144	273	373	273	373	144	373	373	144	144	373	273	0	273	273
15	160	160	394	160	160	394	277	160	394	394	160	277	277	277	0	160
16	305	305	593	174	305	593	483	305	593	483	174	483	483	305	174	0

Źródło: opracowanie własne.

W pierwszej kolejności dane zostały poddane unitaryzacji, następnie przedstawione w postaci ciągłej (wykorzystano bazę Fouriera). Oszacowane wektory przemnożono przez wagi. Dla analizy globalnej był to wektor wag wyznaczony na podstawie współczynników dostępności komunikacyjnej (tab. 2), dla analiz lokalnych wektor wag bazujący na odległościach ustalonych według zasady sąsiedztwa (tab. 4).

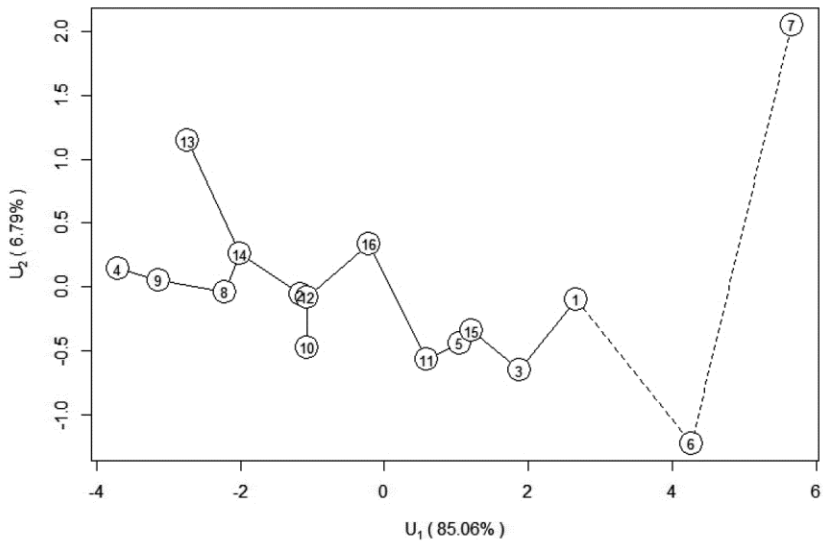
Tabela 4. Wagi w_{kj} związane z województwami małopolskim i wielkopolskim

Województwo małopolskie							
w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
0,3196	0,3196	0,6572	0,0863	0,6572	1,0000	0,6572	0,6572
w_9	w_{10}	w_{11}	w_{12}	w_{13}	w_{14}	w_{15}	w_{16}
0,9152	0,3196	0,0863	0,9152	0,9152	0,3196	0,3196	0,0863
Województwo wielkopolskie							
w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
0,7069	0,7069	0,1220	0,7069	0,7069	0,1220	0,3536	0,7069
w_9	w_{10}	w_{11}	w_{12}	w_{13}	w_{14}	w_{15}	w_{16}
0,1220	0,1220	0,7069	0,3536	0,3536	0,3536	1,0000	0,7069

Źródło: opracowanie własne.

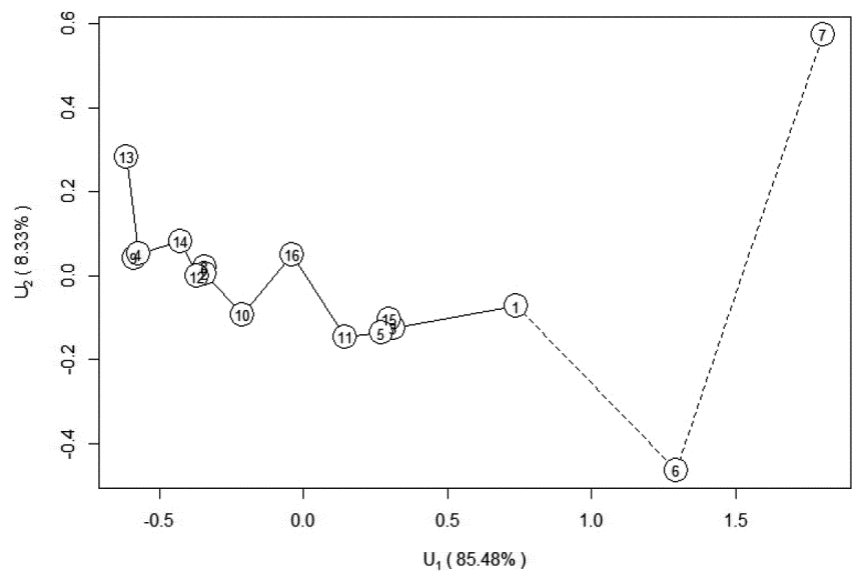
Wyniki prezentowane są na płaszczyźnie dwóch pierwszych funkcjonalnych składowych głównych. Na punktach reprezentujących poszczególne województwa rozpięty został dendryt oraz dokonano jego podziału, zaznaczając linią przerywaną krawędzie dłuższe od wartości krytycznej [Krzyśko i in. 2016, Sekcja 4]. Ze względów technicznych prezentowane są cztery z nich: analiza globalna, analiza globalna ważona dostępnością komunikacyjną oraz dwie analizy lokalne, odpowiednio z Małopolską i Wielkopolską jako punktami centralnymi.

Na rysunku 1 widoczne są 2 skupienia jednoelementowe, izolowane: województwo mazowieckie i województwo małopolskie. Pozostałych 14 województw tworzy jedno skupienie. W tym skupieniu najbliżej województwa małopolskiego położone są województwa dolnośląskie oraz lubelskie. Podobna sytuacja ma miejsce dla analizy globalnej ważonej dostępnością komunikacyjną (rys. 2). Najbliżej województwa małopolskiego położone są województwa lubelskie i dolnośląskie. Jeżeli za punkt odniesienia przyjmiemy województwo małopolskie (rys. 3), to nadal mamy 2 skupienia izolowane: województwo mazowieckie i województwo małopolskie, a pozostałe województwa tworzą jedno skupienie. Wśród nich najbliżej województwa małopolskiego położone jest województwo dolnośląskie. Jeżeli natomiast za punkt odniesienia przyjmiemy województwo wielkopolskie (rys. 4), to województwo mazowieckie stanowi skupienie izolowane. Na przeciwstawnym krańcu położone jest drugie skupienie izolowane – województwo lubuskie. Pozostałe województwa tworzą jedno skupienie. Najbliżej województwa mazowieckiego położone są województwa dolnośląskie, małopolskie i wielkopolskie.



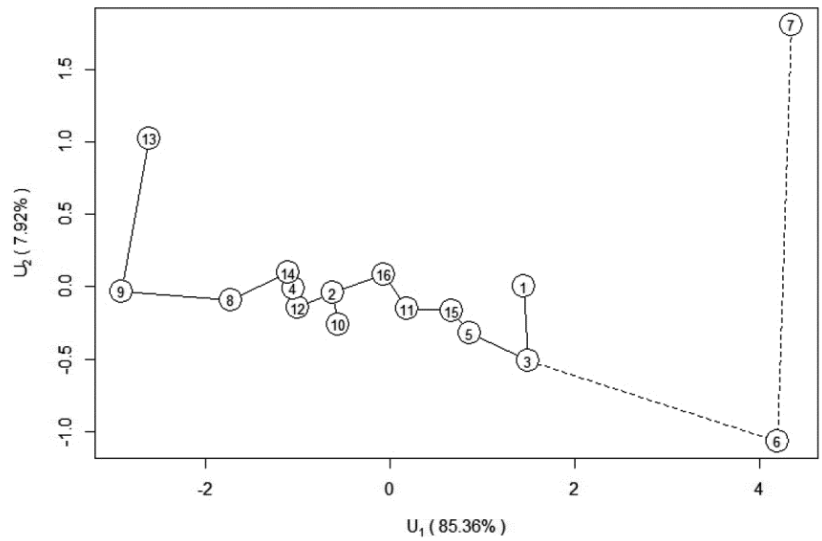
Rys. 1. Analiza globalna klasyczna

Źródło: opracowanie własne.



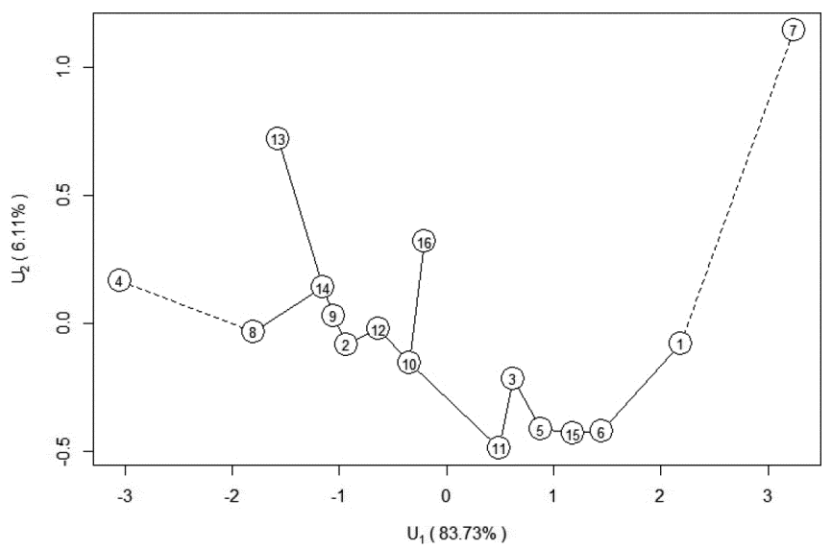
Rys. 2. Analiza globalna ważona dostępnością komunikacyjną

Źródło: opracowanie własne.



Rys. 3. Analiza lokalna (punkt odniesienia – województwo małopolskie)

Źródło: opracowanie własne.



Rys. 4. Analiza lokalna (punkt odniesienia – województwo wielkopolskie)

Źródło: opracowanie własne.

5. Podsumowanie

Przywołane na wstępie pracy modele: analiza składowych głównych dla danych geograficznie ważonych oraz analiza składowych głównych dla danych funkcjonalnych, samodzielnie stanowią rozwinięcie dotychczas stosowanych metod statystycznych.

Połączenie ich w jednym prezentowanym modelu jest kolejnym krokiem umożliwiającym analizę danych w bardziej kompleksowy sposób, uwzględniający trzy wymiary, tj. cechowy (wiele zmiennych), czasowy oraz przestrzenny (dane ważone).

Literatura

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.

Górecki T., Krzyśko M., Waszak Ł., Wołyński W., 2016, *Selected statistical methods of data analysis for multivariate functional data*, Statistical Papers, Publish online: 23 February 2016.

Górniak J., 2015, *Identyfikacja dostępności komunikacyjnej miast na podstawie wskaźników wyposażenia infrastrukturalnego w Polsce*, Studia Ekonomiczne, Zeszyty Naukowe UE w Katowicach, nr 249, s. 282–285.

Harris P., Brunsdon C., Charlton M., 2011, *Geographically weighted principal component analysis*, International Journal of Geographical Information Science, vol. 25, no. 10, s. 1717–1736.

Krzyśko M., Majka A., Wołyński W., 2016, *Ocena zróżnicowania poziomu życia mieszkańców województw w latach 2003-2013 za pomocą składowych głównych dla wielowymiarowych danych funkcjonalnych oraz analizy skupień*, Przegląd Statystyczny, vol. 63, no. 1, s. 81–97.

Lu B., Harris P., Brunsdon C., Charlton M., 2014, *The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models*, Geo-spatial Information Science, vol. 17, no. 2, s. 85–101.