Zeszyty Naukowe Uniwersytet Ekonomiczny w Krakowie
Metody analizy danych

916

ISSN 1898-6447
Zesz. Nauk. UEK, 2013; 916: 71–81

Małgorzata Sej-Kolasa
Mirosława Sztemberg-Lewandowska
Katedra Ekonometrii i Informatyki
Uniwersytet Ekonomiczny we Wrocławiu

Funkcjonalna analiza głównych składowych w badaniu zmian liczby studentów w czasie w krajach europejskich

Streszczenie

Analiza głównych składowych (PCA) polega na transformacji zmiennych pierwotnych w zbiór nowych wzajemnie ortogonalnych zmiennych, zwanych głównymi składowymi. Funkcjonalna analiza głównych (FPCA) składowych ma zalety klasycznej analizy głównych składowych, dodatkowo umożliwia analizę danych o charakterze dynamicznym. Podstawową różnicą między tymi dwiema metodami jest rodzaj danych: PCA bazuje na danych wielowymiarowych, natomiast FPCA na danych funkcjonalnych. Danymi funkcjonalnymi są krzywe i trajektorie, czyli ciąg indywidualnych obserwacji, a nie pojedyncza obserwacja. Celem artykułu jest pokazanie możliwości wykorzystania funkcjonalnej analizy głównych składowych do badania zjawisk opisanych danymi wzdłużnymi (*longitudinal data*). Przykład wykorzystania tej metody omówiony w artykule opiera się na analizie zmiany liczby studentów w czasie w wybranych krajach europejskich. Możliwości wizualizacyjne metody pozwalają na porównanie krajów i wyodrębnienie obserwacji odstających.

Słowa kluczowe: dane funkcjonalne, dane wzdłużne, funkcjonalna analiza głównych składowych, szkolnictwo wyższe.

1. Wprowadzenie

Analiza głównych składowych (PCA) polega na transformacji zmiennych pierwotnych w zbiór nowych wzajemnie ortogonalnych zmiennych, zwanych głównymi składowymi. Metodę tę wykorzystuje się do konstrukcji map percepcji, stosuje się ją także na etapie doboru zmiennych jako metodę redukcji danych. PCA umożliwia ponadto opis zjawisk z punktu widzenia nowych kategorii społeczno-ekonomicznych zdefiniowanych przez czynniki główne (zob. [Harman 1975, s. 154–160; Hair *et al*. 1998]).

Klasyczna analiza głównych składowych nie jest jednak odpowiednia do analizy danych o charakterze dynamicznym. Na analizę takich danych pozwala m.in. dynamiczna analiza głównych składowych lub funkcjonalna analiza głównych składowych.

Dynamiczne modele czynnikowe umożliwiają uzyskanie syntetycznej informacji o kształtowaniu się zmienności dużego zbioru danych, wykorzystywane są do konstruowania prognoz (głównie krótkookresowych) i monitorowania zjawisk. Istota metody sprowadza się do agregacji dużej liczby potencjalnych zmiennych objaśniających do kilku wzajemnie niezależnych czynników, które następnie wykorzystywane są do prognozowania wybranej zmiennej. Równanie prognostyczne opisujące zależności pomiędzy zmiennymi prognozowanymi a czynnikami ma zwykle postać liniową. Oprócz czynników w równaniu tym mogą wystąpić ich opóźnienia, a także składniki o charakterze autoregresyjnym.

Funkcjonalna analiza głównych składowych (FPCA) pozwala określić naturę danych, kształt trajektorii w czasie. Zarówno klasyczna, jak i funkcjonalna analiza głównych składowych pozwalają dokonać rzutu wielowymiarowych danych na przestrzeń o dużo mniejszym wymiarze przy jednoczesnym zachowaniu maksimum informacji (w tym przypadku zmienności danych). Analizowana metoda polega na znalezieniu takich składowych, których iloczyn skalarny z danymi daje maksymalną zmienność. Pierwsza składowa wyjaśnia najwięcej zmienności, druga jest prostopadła do pierwszej i wyjaśnia maksymalnie dużo z pozostałej części zmienności danych. Podstawową różnicą między tymi dwiema metodami jest rodzaj danych: PCA bazuje na danych wielowymiarowych, natomiast FPCA na danych funkcjonalnych [Ramsay i Silverman 2005, Ramsay, Hooker i Graves 2009].

Celem artykułu jest analiza zmiany liczby studentów w czasie w krajach europejskich. Badanie obejmuje nie tylko tendencję, ale także tempo zmian liczby studentów w czasie. Zastosowano funkcjonalną analizę głównych składowych, której możliwości wizualizacyjne pozwalają na porównanie krajów i wyodrębnienie obserwacji odstających.

Obliczenia zostały wykonane w programie R (v.2.15.0, pakiet fda i fda.usc).

2. Metoda badawcza

Danymi funkcjonalnymi są krzywe i trajektorie, czyli ciąg indywidualnych obserwacji, a nie jak zwykle pojedyncza obserwacja. Chociaż dane funkcjonalne często są wyrażone w czasie (zależą od czasu), to ich zakres i cel są zupełnie inne niż szeregów czasowych. Analiza szeregów czasowych ma na celu modelowanie lub prognozowanie danych, natomiast funkcjonalna analiza danych pozwala badać naturę danych, kształt trajektorii w czasie [Ingrassia i Costanzo 2005].

Dane funkcjonalne posiadają realizacje dyskretne. Niech $\mathbf{y}_i = \big((y_i(t_1), y_i(t_2), \ldots, y_i(t_p)\big)$ będzie próbkowym pomiarem zmiennej Y w czasie $t_1, t_2, \ldots, t_p$ dla i-tej jednostki ($i = 1, 2, \ldots, n$). Dane y_i nazywane są surowymi danymi funkcjonalnymi (*raw functional data*). Dane te przekształca się zgodnie z procedurami wygładzającymi, np. za pomocą liniowych kombinacji K znanych funkcji bazowych, na odpowiednią funkcję $x_i(t)$, która jest właściwą postacią funkcjonalną danych. Zbiór $\mathbf{X}_t = \big((x_1(t), x_2(t), \ldots, x_n(t)\big)$ nazywany jest funkcjonalnym zbiorem danych (*functional dataset*) [Daniele 2006, Hall, Müller i Wang 2006].

Techniki statystyczne dla funkcjonalnych danych zakładają, że funkcje opisujące dane należą do przestrzeni Hilberta: są funkcjami rzeczywistymi określonymi na przedziale domkniętym, całka kwadratów tych funkcji jest skończona (tzn. norma funkcji jest skończona).

Klasyczna analiza głównych składowych służy do eksploracji zmienności w wielowymiarowym zbiorze danych. Wykorzystując wartości własne macierzy wariancji dla danych PCA, za jej pomocą wyznacza się składowe, które wyjaśniają zmienność w obserwowanym zbiorze danych. Dla każdej składowej głównej ustala się ładunki czynnikowe na wszystkich zmiennych, określające wariancję wyjaśnioną przez daną składową [Harman 1975].

W przypadku funkcjonalnej analizy głównych składowych każda główna składowa wyrażona jest przez funkcję wagową głównych składowych (*principal component weight function*), inaczej nazwaną funkcją własną (*eigenfunction*) $\xi_j(t)$ zależną od czasu [Daniele 2006]. Funkcja własna maksymalizuje wariancję funkcji głównych składowych:

$$v(t,s) \overset{def}{=} \frac{1}{n-1}\sum_{i=1}^{n}\big\{x_i(t)-\overline{x}(t)\big\}\big\{x_i(s)-\overline{x}(s)\big\}. \tag{1}$$

Analogicznie do klasycznej PCA problemem w funkcjonalnej analizie głównych składowych jest rozkład wariancji funkcji:

$$v(t,s) = \sum_j \lambda_j \xi_j(t)\xi_j(s), \tag{2}$$

gdzie λ_j, $\xi_j(t)$ spełniają równanie własne:

$$\left\langle v(s), \xi_j \right\rangle = \lambda_j \xi_j(s), \tag{3}$$

a wartości własne są dodatnie i niemalejące:

$$\lambda_j \overset{def}{=} \int_T \xi_j(t) v(t,s) \xi_j(s) dtds. \tag{4}$$

Funkcje własne spełniają warunek:

$$\int_T \xi_j^2(t) dt = 1 \quad \text{oraz} \quad \int_T \xi_j(t) \xi_i(t) dt = 0 \quad (i < j). \tag{5}$$

Wyniki głównych składowych dla *i*-tego obiektu w zbiorze danych są zdefiniowane następująco:

$$w_i^{(j)} \overset{def}{=} \left\langle x_i, \xi_j \right\rangle = \int_T \xi(t) x_i(t) dt. \tag{6}$$

Funkcje własne określają główne składowe zmienności między próbkowymi funkcjami x_i [Ingrassia i Costanzo 2005, Hall, Müller i Wang 2006].

3. Przebieg badania

Spadek liczby ludności oraz starzenie się społeczeństw skutkują wieloma niekorzystnymi zmianami o charakterze ekonomicznym i społecznym. Demograficzne tsunami wpływa również na sytuację szkolnictwa wyższego. Od kilku lat w szkołach wyższych w większości krajów Europy liczba studentów spada, co znacząco wpływa na ograniczenie możliwości rozwoju szkolnictwa wyższego.

W celu uniknięcia bezpośredniego porównywania liczby studentów w przeprowadzonych analizach wykorzystano zmienną będącą stosunkiem liczby studentów w danym roku do liczby studentów w pierwszym badanym roku.

Badaniem objęto kraje europejskie, dane pochodziły z baz Eurostatu. Początkowo zakładano uwzględnienie wszystkich krajów Europy, jednak brak danych dla niektórych państw spowodował konieczność zawężenia zakresu czasowego (lata 2000–2009) oraz przestrzennego (28 państw). Z tego powodu wśród państw poddanych analizie nie znalazła się m.in. Francja (tabela 1).

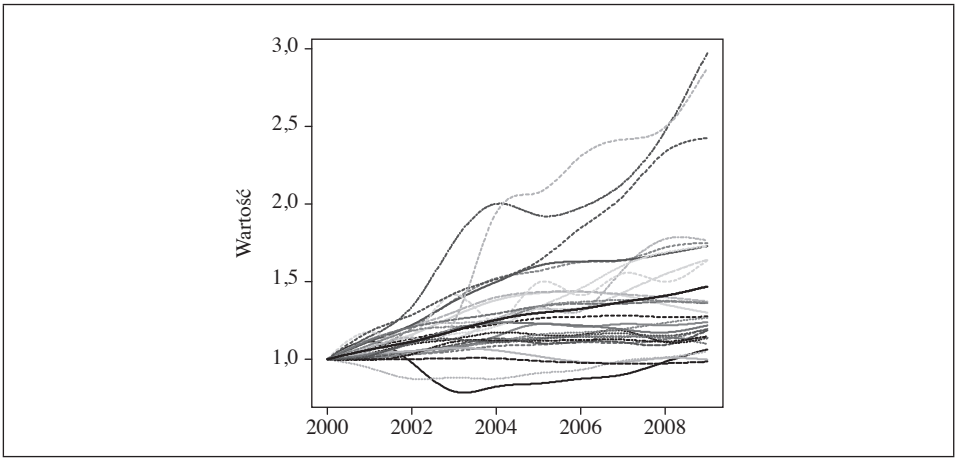
Tabela 1. Zmiana liczby studentów w czasie

Kraj	2000[a]	2001	2002	2003	2004	2005	2006	2007	2008	2009
Austria	1,00	1,12	0,98	0,79	0,82	0,84	0,87	0,90	0,98	1,06
Belgia	1,00	1,01	1,03	1,05	1,08	1,09	1,11	1,11	1,13	1,19
Bułgaria	1,00	0,95	0,87	0,88	0,87	0,91	0,93	0,99	1,01	1,05
Cypr	1,00	1,15	1,34	1,75	2,00	1,93	1,98	2,13	2,47	2,98
Czechy	1,00	1,03	1,12	1,13	1,26	1,33	1,33	1,43	1,55	1,64
Dania	1,00	1,01	1,03	1,07	1,15	1,23	1,21	1,23	1,22	1,24
Estonia	1,00	1,08	1,13	1,19	1,22	1,26	1,27	1,28	1,27	1,28
Finlandia	1,00	1,03	1,05	1,08	1,11	1,13	1,14	1,14	1,15	1,10
Hiszpania	1,00	1,00	1,00	1,01	1,01	0,99	0,98	0,97	0,97	0,98
Holandia	1,00	1,03	1,05	1,08	1,11	1,16	1,17	1,19	1,24	1,27
Irlandia	1,00	1,04	1,10	1,13	1,17	1,16	1,16	1,19	1,11	1,14
Islandia	1,00	1,05	1,20	1,38	1,52	1,57	1,63	1,64	1,72	1,75
Litwa	1,00	1,12	1,22	1,37	1,50	1,60	1,63	1,64	1,68	1,73
Łotwa	1,00	1,13	1,21	1,30	1,40	1,43	1,44	1,42	1,40	1,37
Macedonia	1,00	1,09	1,21	1,24	1,26	1,34	1,31	1,58	1,77	1,77
Malta	1,00	1,18	1,15	1,42	1,25	1,50	1,41	1,55	1,50	1,64
Niemcy	1,00	1,01	1,05	1,09	1,13	1,10	1,11	1,11	1,09	1,19
Norwegia	1,00	1,00	1,03	1,11	1,12	1,12	1,12	1,13	1,11	1,15
Polska	1,00	1,12	1,21	1,26	1,29	1,34	1,36	1,36	1,37	1,36
Portugalia	1,00	1,04	1,05	1,07	1,06	1,02	0,98	0,98	1,01	1,00
Rumunia	1,00	1,18	1,29	1,42	1,51	1,63	1,84	2,05	2,33	2,43
Słowacja	1,00	1,06	1,12	1,16	1,21	1,33	1,46	1,60	1,69	1,73
Słowenia	1,00	1,09	1,18	1,21	1,25	1,34	1,37	1,38	1,38	1,36
Szwecja	1,00	1,03	1,10	1,20	1,24	1,23	1,22	1,19	1,17	1,22
Turcja	1,00	1,08	1,14	1,24	1,94	2,07	2,31	2,42	2,49	2,88
Węgry	1,00	1,08	1,16	1,28	1,38	1,43	1,44	1,41	1,35	1,30
Wielka Brytania	1,00	1,02	1,11	1,13	1,11	1,13	1,15	1,17	1,15	1,19
Włochy	1,00	1,02	1,05	1,08	1,12	1,14	1,15	1,15	1,14	1,14

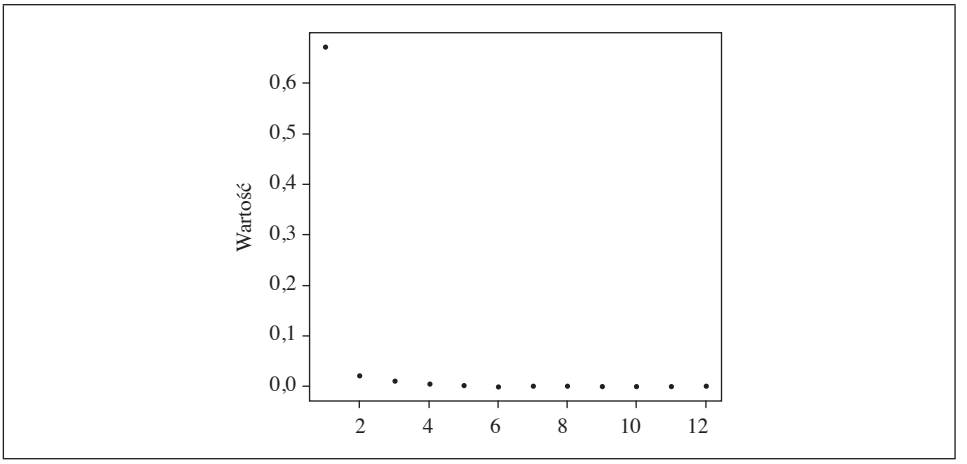
[a] 2000 r. – rok bazowy.

Źródło: opracowanie własne na podstawie [*Eurostat Statistics*… 2012].

Wielkości zmian liczby studentów w badanym okresie dla wybranych państw europejskich przedstawia rys. 1. Pogrubiona krzywa oznacza średnią dla badanych państw. Rok 2000 był rokiem bazowym, w związku z czym dla tego roku wskaźnik dla wszystkich państw przyjmuje wartość 1.

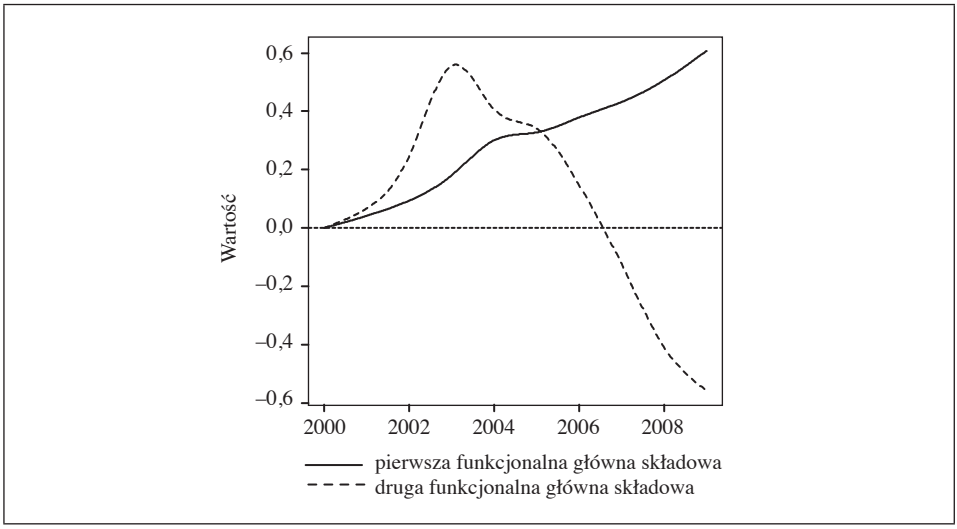


Rys. 1. Wielkości zmian liczby studentów w czasie
Źródło: opracowanie własne z wykorzystaniem programu R.



Rys. 2. Wykres osypiska
Źródło: opracowanie własne z wykorzystaniem programu R.

Wykres pokazuje, jak zmienia się liczba studentów w czasie. Średnia liczba studentów w badanym okresie rośnie, jednak trudno jest wskazać, w których krajach tempo zmian jest podobne. W tym celu przeprowadzono funkcjonalną analizę głównych składowych.



Rys. 3. Wyodrębnione funkcjonalne główne składowe
Źródło: opracowanie własne z wykorzystaniem programu R.

Na podstawie wykresu osypiska (rys. 2) ustalono liczbę funkcji składowych, a następnie wyodrębniono te funkcje (rys. 3).

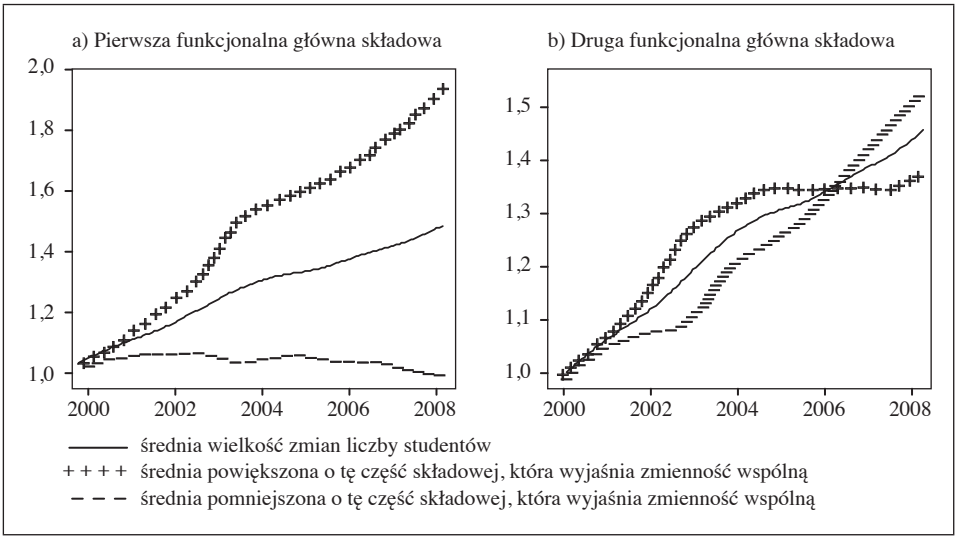
Pierwsza funkcjonalna główna składowa, na rys. 3 oznaczona linią ciągłą, wyjaśnia 85% zmienności wspólnej, natomiast druga, oznaczona linią przerywaną – 7%.

Podobnie jak w klasycznej analizie czynnikowej główne składowe należy zinterpretować, jednak w przypadku danych funkcjonalnych interpretacja jest trudniejsza. Praktyczne wyjaśnienie funkcjonalnych głównych składowych ułatwiają wykresy odchylenia każdej ze składowych od średniej (rys. 4).

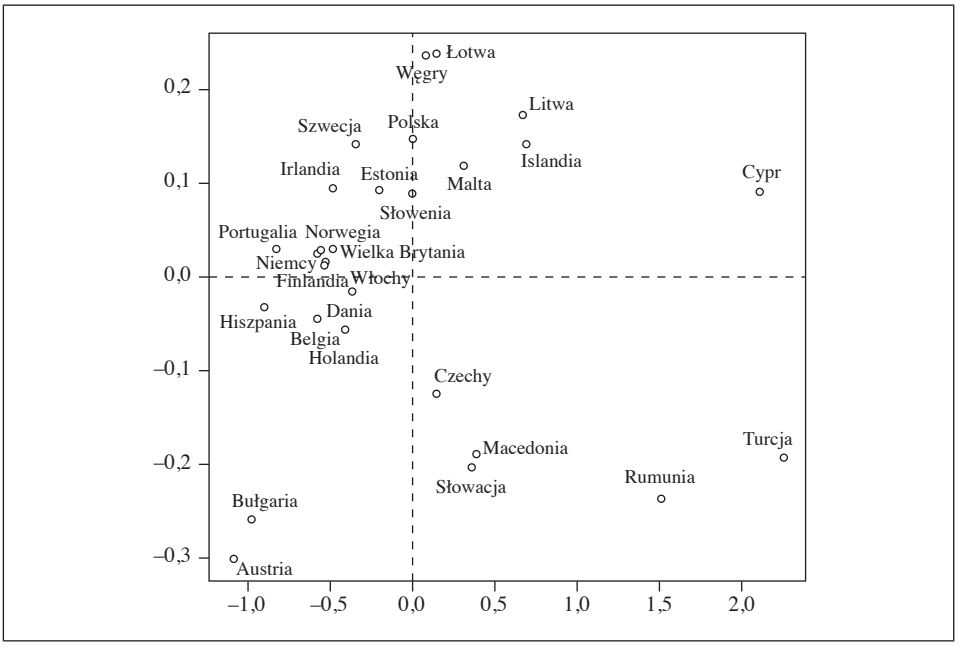
Z rys. 4a wynika, że pierwsza składowa odpowiada za ogólną tendencję. Przyjmuje wartości dodatnie (rys. 3), zatem dla danego kraju dodatni ładunek na tej składowej oznacza, że krzywa opisująca wielkość zmian liczby studentów leży powyżej średniej.

Druga składowa pokazuje tendencje w pierwszych latach w odniesieniu do lat ostatnich („początek kontra koniec") – porównuje zatem okres do 2007 r. z okresem po 2007 r. (rys. 4b). Dodatni ładunek na drugiej składowej oznacza, że tempo zmian na początku badanego okresu było większe od średniej, natomiast na końcu tego okresu tempo było mniejsze od średniej.

Funkcjonalna analiza czynnikowa pozwala na wizualizację danych umożliwiającą porównanie badanych obiektów. Rys. 5 zawiera rzut danych na płaszczyznę wyznaczoną przez dwie funkcjonalne główne składowe.



Rys. 4. Odchylenia funkcjonalnych głównych składowych od średniej
Źródło: opracowanie własne z wykorzystaniem programu R.



Rys. 5. Rzut obiektów na płaszczyznę wyznaczoną przez dwie funkcje składowe
Źródło: opracowanie własne z wykorzystaniem programu R.

Wśród krajów znacząco odbiegających od pozostałych pod względem tempa i kierunku zmian liczby studentów znalazły się m.in.: Turcja, Cypr, Rumunia, Bułgaria i Austria.

Najbardziej od pozostałych państw różni się Turcja, która ma wysoki ładunek na pierwszej składowej (tempo zmian powyżej średniej) i niski (ujemny) na drugiej, co oznacza, że tempo zmian w końcowym okresie było większe niż w początkowym. Szybkie tempo zmian liczby studentów można tłumaczyć odmienną niż w pozostałych krajach sytuacją demograficzną – w Turcji systematycznie wzrasta liczba osób młodych, co jest ewenementem na skalę Europy.

Cypr – podobnie jak Turcja – wykazuje tempo zmian powyżej średniej (wysoki ładunek na pierwszej składowej), ale tempo zmian w początkowym okresie było wyższe niż w końcowym (dodatni ładunek na drugiej składowej). Tempo zmian powyżej średniej częściowo wyjaśnia duża liczba studiujących tam obcokrajowców – jak wynika z danych Eurostatu, wskaźnik umiędzynarodowienia uczelni wynosi na Cyprze ponad 28% i jest jednym z najwyższych w Europie.

W Rumunii zmiany liczby studentów mają podobny charakter jak w Turcji: odnotowuje się tempo zmian powyżej średniej, wyższe w końcowym okresie. Wynika to z sytuacji demograficznej – w badanym okresie osoby w wieku „studenckim" zgodnie z danymi Eurostatu stanowiły ponad 20% społeczeństwa Rumunii.

Bułgaria i Austria są w podobnej sytuacji pod względem omawianych składowych. Ujemny ładunek na pierwszej składowej oznacza, że tempo zmian liczby studentów jest poniżej średniej, większe tempo zmian miało miejsce na końcu badanego okresu (ujemny ładunek na drugiej składowej).

W Polsce tempo zmian liczby studentów jest średnie, w początkowym okresie zaobserwowano większe zmiany niż w końcowym.

4. Wnioski

Na podstawie przeprowadzonych analiz można wysnuć wnioski zarówno o charakterze teoretycznym, dotyczące metody, jak i o charakterze aplikacyjnym.

Podobnie jak klasyczna analiza głównych składowych FPCA pozwala na wizualizację zjawiska (co znacznie ułatwia analizę), wykazując nietrywialnie zależności, które trudno wykryć w inny sposób. Istotną zaletą obu metod jest redukcja danych przy zachowaniu maksimum informacji.

Funkcjonalna analiza głównych składowych, wzbogacając możliwości klasycznej analizy głównych składowych, pozwala na analizę danych o charakterze dynamicznym – pokazuje nie tylko tendencję, ale i tempo zmian w czasie.

Zastosowana metoda, określając tempo zmiany liczby studentów, umożliwia wskazanie krajów podobnych oraz znalezienie krajów różniących się od pozostałych ze względu na przyjęte kryterium analiz.

Braki danych, dotyczące głównie zakresu przestrzennego, mogą wpływać na otrzymane wyniki badań. Badaniem należałoby objąć wszystkie kraje. Uwzględnienie w badaniu pominiętych państw (np. Francji) mogłoby spowodować konieczność zmiany interpretacji funkcjonalnych składowych głównych, a tym samym zmianę wniosków na temat sytuacji poszczególnych państw.

Analiza otrzymanych wyników w celu wyjaśniania przyczyn sytuacji określonych państw nie była przedmiotem zainteresowania i powinna być przedmiotem odrębnych badań.

Literatura

Daniele M. [2006], *Functional Principal Components Analysis to Study Environmental Data*, http://www.sis-statistica.it/files/pdf/atti/Spontanee%202006_677-680.pdf (dostęp: 5.12.2013).

Eurostat Statistics [2012], http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database?_piref458_1209540_458_211810_211810.node_code=educ_enrl8, (dostęp: 30.10.2012).

Hair J.F. *et al*. [1998], *Multivariate Data Analysis with Readings*, Prentice-Hall, New York.

Hall P., Müller H.G., Wang J.L. [2006], *Properties of Principal Component Methods for Functional and Longitudinal Data Analysis*, „The Annals of Statistics", vol. 34, nr 3.

Harman H. [1975], *Modern Factor Analysis*, The University of Chicago Press, Chicago.

Ingrassia S., Costanzo G.D. [2005], *Functional Principal Component Analysis of Financial Time Series* [w:] *New Developments in Classification and Data Analysis*, red. M. Vichi *et al*., Springer, Berlin.

Ramsay J.O., Hooker G., Graves S. [2009], *Functional Data Analysis with R and MATLAB*, Springer, New York.

Ramsay J.O., Silverman B.W. [2005], *Functional Data Analysis*, Springer, New York.

The Analysis of Changes over Time in the Number of Students Using Functional Principal Component Analysis in European Countries

Principal component analysis (PCA) transforms an original set of variables into a new orthogonal set called principal components. Functional principal component analysis (FPCA) has the same advantages as classical principal component analysis while also enabling the analysis of dynamic data. The main difference between them is that PCA is based on multidimensional data and FPCA is based on functional data. The functional data are curves, surfaces or anything else varying over a continuum. They are not a single observation. The main aim of the paper is to show the usefulness of applying functional principal component analysis in order to analyse longitudinal data. The paper presents an

example of how this method has been used based on the analysis of changes in the number of students (over time) in chosen European countries. Visualisation of the results makes it possible to compare countries and detect outliers.

Keywords: functional data, longitudinal data, functional principal component analysis, higher education.