PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 385

Taksonomia 25
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2015

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronie internetowej Wydawnictwa
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 385 • 2015
Taksonomia 25 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Mirosława Sztemberg-Lewandowska
Uniwersytet Ekonomiczny we Wrocławiu
e-mail: mirosława.sztemberg-lewandowska@ue.wroc.pl

PROBLEMY DECYZYJNE W FUNKCJONALNEJ ANALIZIE GŁÓWNYCH SKŁADOWYCH

Streszczenie: Analiza funkcjonalna bazuje na danych funkcjonalnych, tzn. na krzywych i trajektoriach, czyli ciągu indywidualnych obserwacji, a nie jak w przypadku danych wielowymiarowych na pojedynczej obserwacji. Funkcjonalna analiza głównych składowych polega na transformacji funkcjonalnych zmiennych pierwotnych w zbiór nowych wzajemnie ortogonalnych zmiennych, zwanych głównymi składowymi. Zastosowanie metody dla danych funkcjonalnych umożliwia analizę danych o charakterze dynamicznym. Celem artykułu jest scharakteryzowanie etapów funkcjonalnej analizy głównych składowych ze szczególnym omówieniem kroków, które nie występują w klasycznej analizie głównych składowych.

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

DOI: 10.15611/pn.2015.385.29

1. Wstęp

Analiza funkcjonalna zajmuje się analizą danych o charakterze funkcjonalnym. Danymi funkcjonalnymi są krzywe i trajektorie, czyli ciąg indywidualnych obserwacji, a nie 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 bada naturę danych, kształt trajektorii w czasie [Ingrassia i Costanzo 2005].

Dane funkcjonalne mają realizacje dyskretne. Dane te przekształca się za pomocą procedur wygładzających, np. za pomocą liniowych kombinacji znanych funkcji bazowych, na odpowiednią funkcję $x_i(t)$, która jest właściwą postacią funkcjonalną danych [Daniele 2006; Hall, Müller, 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).

Prekursorami zastosowania analizy głównych składowych dla danych funkcjonalnych byli:

- Besse i Ramsay [1986],
- Ramsay i Dalzell [1991],
- Rice i Silverman [1991],
- Silverman [1995; 1996].

Zarówno klasyczna (PCA), jak i funkcjonalna (FPCA) analiza głównych składowych pozwalają wykonać rzut wielowymiarowych danych na przestrzeń o dużo mniejszym wymiarze, jednocześnie zachowując maksymalnie dużo informacji (w tym przypadku zmienności danych). Podstawową różnicą tych dwóch metod jest rodzaj danych: PCA bazuje na danych wielowymiarowych, natomiast FPCA na danych funkcjonalnych [Ramsay i Silverman 2005; Ramsay, Dalzell 1991].

Celem artykułu jest scharakteryzowanie etapów funkcjonalnej analizy głównych składowych ze szczególnym omówieniem przekształcenia danych do postaci funkcjonalnej. W artykule przedstawiono przykład empiryczny, w którym dokonano analizy liczby studentów szkół wyższych w Polsce. Celem badania było porównanie sytuacji różnych typów szkół wyższych na przełomie lat 2000-2013.

2. Przekształcanie danych na dane funkcjonalne

Dana jest zmienna $\mathbf{y}_i$. Niech $\mathbf{y}_i = \left(y_i(t_1), y_i(t_2), \ldots, y_i(t_p)\right)$ będzie próbkowym pomiarem zmiennej Y w czasie $t_1, t_2, \ldots, t_p$ dla *i*-tej jednostki $\left(i = 1,2,\ldots,n\right)$. Dane y_i nazywane są surowymi danymi funkcjonalnymi (*raw functional data*). Realizacje dyskretne przekształca się w funkcję ciągłą $x_i(t)$. Zbiór $\mathbf{X}_t = \left(x_1(t), x_2(t), \ldots, x_n(t)\right)$ nazywany jest funkcjonalnym zbiorem danych (*functional dataset*) [Daniele 2006; Hall i in. 2006].

Funkcjonalna analiza głównych składowych polega na znalezieniu składowych głównych wyjaśniających najwięcej zmienności wspólnej wszystkich zmiennych. Problemem jest wyznaczenie funkcji $x_i(t)$, dla której możliwe jest wyznaczenie głównych składowych.

W przypadku gdy nie można przedstawić Y w prostej funkcyjnej postaci, Ramsay i Silverman (1997) zaproponowali trzy podejścia do FPCA:

1. Dyskretyzacja danych – przeprowadza się PCA dla danych dyskretnych (równo odległe punkty pomiaru), a następnie otrzymane wektory własne przekształca się do funkcjonalnej postaci.

2. Numeryczny schemat obliczeniowy wyznaczenia funkcjonalnych wektorów własnych z równania własnego (dopuszczalne nieregularne punkty pomiaru).

3. Funkcję $x_i(t)$ przedstawia się jako kombinację liniową funkcji bazowych:

$$x_i(t) = \sum_g c_{ij}\phi_j(t), \tag{1}$$

gdzie: c_{ij} – współczynniki kombinacji liniowej,

$\phi_j(t)$ – funkcje tworzące bazę ortonormalną przestrzeni $L_2(I)$ oraz $x_i(t) \in L_2(I)$,

$L_2(I)$ – przestrzeń Hilberta funkcji całkowalnych z kwadratem na przedziale I

wyposażona w iloczyn skalarny $\langle u, v\rangle = \int_I u(t)v(t)dt$.

Najczęściej wykorzystywane są następujące funkcje bazowe:

- jednomiany 1, t, t^2, t^3, . . . , t^k, ...
- funkcje Fouriera (dla danych cyklicznych) 1, sin(ωt), cos(ωt), sin(2ωt), cos(2ωt), sin(3ωt), cos(3ωt), . . . , sin(kωt), cos(kωt), . . .
- funkcje *B*-spline, które posiadają następujące własności:
  - każda funkcja bazowa jest „sklejeniem" funkcji rzędu m w punktach nazywanych węzłami,
  - suma, różnica i kombinacja liniowa tych funkcji bazowych jest nadal funkcją typu B-spline.

Jako kryterium dopasowania dla każdej krzywej przyjmuje się całkę z kwadratu błędu:

$$\left\|x_i - \hat{x}_i\right\|^2 = \int\left[x(s) - \hat{x}(s)\right]^2 ds, \tag{2}$$

gdzie x_i *i* $\hat{x}_i$ to odpowiednio obserwowane i dopasowane krzywe.

Globalna miara aproksymacji dana jest wzorem:

$$SSE = \sum_{i=1}^{n}\left\|x_i - \hat{x}_i\right\|^2 . \tag{3}$$

Dane funkcjonalne powinny być **wygładzone**, wszelkie chropowatości funkcji traktowane są jako szum, który powinien być całkowicie usunięty. Miarą chropowatości funkcji jest jej pochodna drugiego rzędu, funkcje wygładzone powinny przyjmować małe wartości tej pochodnej.

Stosuje się dwa podejścia do wygładzania danych:

1) funkcje wygładza się w trakcie wyodrębniania głównych składowych,
2) dane wygładza się przed zastosowaniem procedury FPCA.

Ramsay i Silverman (1997) zaproponowali **wygładzanie danych podczas wyodrębniania głównych składowych** poprzez maksymalizację następującej funkcji przez funkcje wagowe:

$$\frac{\frac{1}{n-1}\sum_i\left[\int \xi_j(t)x_i(t)dt\right]^2}{\int \xi_j{}^2(t)dt+\alpha\int\left(D^2\xi_j(t)\right)^2dt} \tag{4}$$

przy warunku $\int_T \xi_j^2(t)dt=1$, gdzie α jest parametrem. $D^2\xi_j(s)$ jest pochodną II rzędu funkcji $\xi_j(s)$. Dla $\alpha=0$ funkcja jest równa wariancji próbkowej, zatem otrzymujemy niewygładzoną FPCA. Rozwiązanie tej optymalizacji sprowadza się do rozwiązania równania własnego:

$$\left\langle v(s),\xi_j\right\rangle=\lambda_j\left(1+\alpha D^4\xi_j(s)\right)\xi_j(s), \tag{5}$$

gdzie $D^4\xi_j(s)$ jest pochodną IV rzędu funkcji $\xi_j(s)$.

W drugim podejściu stosuje się **wygładzanie danych przed zastosowaniem procedury FPCA**. Jako kryterium dopasowania krzywej do danych obserwowalnych przyjmuje się minimalizację kwadratu błędu uzupełnioną o „karę" dla funkcji niewygładzonych:

$$\left\|x_i-\hat{x}_i\right\|^2+\alpha\left\|\hat{x}_i\right\|^2. \tag{6}$$

Obie normy niekoniecznie są takie same. Druga norma powinna być powiązana z pochodną drugiego rzędu funkcji $\hat{x}_i(t)$, która jest miarą chropowatości funkcji.

3. Funkcjonalna analiza głównych składowych

Klasyczna analiza głównych składowych (PCA) służy do eksploracji zmienności w wielowymiarowym zbiorze danych. Wykorzystując wartości własne macierzy wariancji dla danych, PCA wyznacza składowe, które wyjaśniają zmienność w obserwowanym zbiorze danych. Dla każdej składowej głównej określa się ładunki czynnikowe na wszystkich zmiennych określające wariancję wyjaśnioną przez daną składową.

W przypadku funkcjonalnej analizy głównych składowych (FPCA) 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; Hall, Hosseini-Nasab 2006]. Funkcja własna maksymalizuje wariancję funkcji głównych składowych:

$$v(t,s) \overset{def}{=} \frac{1}{n-1}\sum_{i=1}^{n}\{x_i(t)-\bar{x}(t)\}\{x_i(s)-\bar{x}(s)\}. \tag{7}$$

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

$$v(t,s)=\sum_j \lambda_j \xi_j(t)\xi_j(s), \tag{8}$$

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

$$\langle v(s), \xi_j \rangle = \lambda_j \xi_j(s) \tag{9}$$

oraz wartości własne są dodatnie i niemalejące:

$$\lambda_j \overset{def}{=} \int_T \xi_j(t) v(t,s) \xi_j(s) dtds\,. \tag{10}$$

Funkcje własne spełniają warunek:

$$\int_T \xi_j^2(t)dt = 1 \text{ oraz } \int_T \xi_j(t)\xi_i(t)\,dt = 0 \quad (i<j). \tag{11}$$

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

$$w_i^{(j)} \overset{def}{=} \langle x_i, \xi_j \rangle = \int_T \xi_j(t) x_i(t)\,dt\,. \tag{12}$$

Funkcje własne określają główne składowe zmienności między próbkowymi funkcjami x_i [Ingrassia i Costanzo 2005; Hall i in. 2006; Krzyśko i in. 2012].

Problemy decyzyjne w funkcjonalnej analizie głównych składowych występują na każdym etapie jej procedury.

Przed przystąpieniem do wyodrębniania składowych należy **określić liczbę składowych**, wykorzystując do tego celu np. wykres osypiska, wartości własne lub wskaźnik CV (*cross validation*). Następnie **wyodrębnia się funkcje składowe**. Po wyodrębnieniu składowych, podobnie jak w klasycznej analizie głównych składowych, należy je **zinterpretować**. W przypadku danych funkcjonalnych interpretacja jest trudniejsza.

Praktyczne wyjaśnienie funkcjonalnych głównych składowych ułatwiają wykresy odchylenia każdej ze składowych od średniej.

W celu łatwiejszej interpretacji składowych można przeprowadzić **rotację układu**.

Funkcjonalna analiza czynnikowa umożliwia przedstawienie **obiektów w przestrzeni funkcji głównych składowych**. Taka wizualizacja danych umożliwia porównanie badanych obiektów.

4. Przykład empiryczny

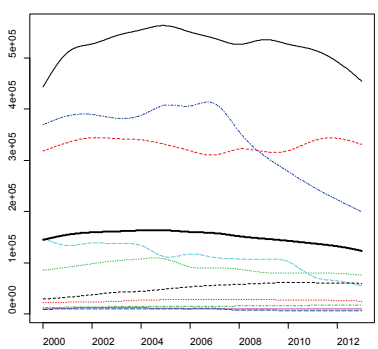
Spadek liczby ludności, starzenie się społeczeństw skutkuje wieloma niekorzystnymi zmianami o charakterze ekonomicznym i społecznym. Demograficzne tsunami wpływa również na sytuację szkolnictwa wyższego. Od kilku lat w szkołach wyższych liczba studentów spada, co znacząco wpływa na ograniczenie możliwości rozwoju szkolnictwa wyższego. Pojawia się pytanie, w jakich typach szkół wyższych sytuacja jest najgorsza.

W tabeli 1 zamieszczono liczbę studentów poszczególnych uczelni wyższych w Polsce na przełomie lat 2000-2013.

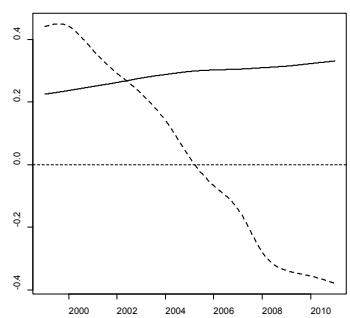
Tabela 1. Liczba studentów w szkołach wyższych (2000-2013)

Rok	Uniwersytety	Wyższe szkoły techniczne	Wyższe szkoły rolnicze	Wyższe szkoły ekonomiczne	Wyższe szkoły pedagogiczne	Wyższe szkoły morskie	Uniwersytety medyczne	Akademie wychowania fizycznego	Wyższe szkoły artystyczne	Wyższe szkoły teologiczne
2000	443 291	318 377	85 539	369 498	148 293	10 135	29 487	22 171	12 793	9 283
2001	510 145	334 511	91 339	386 596	134 089	11 688	32 824	23 010	13 314	9 848
2002	527 248	344 317	98 147	389 537	138 871	12 401	37 669	23 724	14 129	10 033
2003	543 368	342 407	104 077	382 319	137 204	12 216	42 360	24 893	14 563	10 200
2004	554 878	340 219	107 645	387 878	133 800	12 111	44 460	26 951	15 101	10 438
2005	563 062	331 052	107 696	407 755	111 846	11 476	48 842	28 157	15 391	10 422
2006	550 494	318 905	91 997	406 171	117 409	10 500	53 060	29 048	14 932	10 652
2007	538 208	310 555	89 735	410 810	110 334	9 921	56 114	28 713	15 377	10 980
2008	526 381	322 111	87 556	356 561	107 668	10 103	58 015	28 184	15 736	7 392
2009	535 576	317 468	81 245	309 991	106 822	9 977	59 922	28 206	16 132	7 480
2010	526 796	318 738	80 494	278 425	102 540	10 402	61 957	27 574	16 444	6 784
2011	516 237	337 828	80 430	248 642	73 585	10 566	61 210	27 231	16 970	7 000
2012	492 939	343 083	79 403	223 467	64 956	10 398	60 595	26 459	17 134	6 106
2013	454 225	331 099	76 064	199 409	54 921	10 064	59 665	25 335	17 065	6 064

Źródło: opracowanie własne na podstawie danych BDR [http://stat.gov.pl/bdl/app/wybrane_cechy.display?p_id=235850&p_token=0.945545744150877].



Rys. 1. Liczba studentów

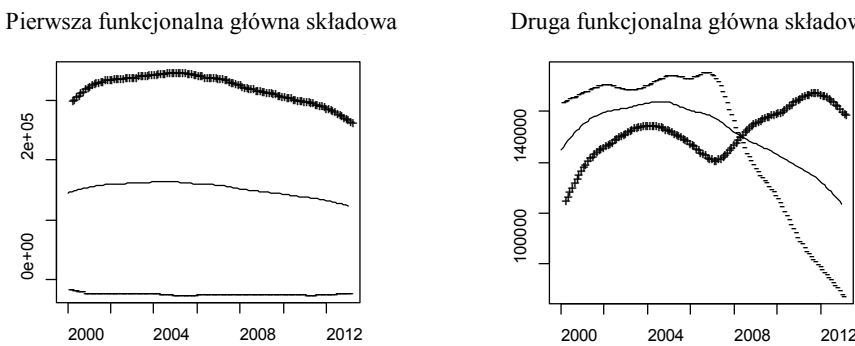


Rys. 2. Funkcjonalne główne składowe

Źródło: opracowanie własne z wykorzystaniem programu R.

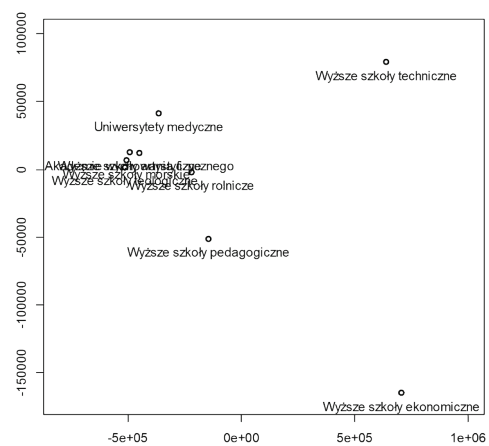
Empiryczne wartości liczby studentów w szkołach wyższych w latach 2000--2013 przedstawia rys. 1. Pogrubiona krzywa oznacza średnią dla badanych typów uczelni. Dane wielowymiarowe przekształcono na dane funkcjonalne metodą B-spline, a następnie za pomocą funkcjonalnej analizy głównych składowych wyodrębniono dwie główne składowe (rys. 2).

Pierwsza funkcjonalna główna składowa, oznaczona linią ciągłą, wyjaśnia 75,1% zmienności wspólnej, natomiast druga, oznaczona linią przerywaną, 24,7%. Praktyczne wyjaśnienie funkcjonalnych głównych składowych ułatwiają wykresy odchylenia każdej ze składowych od średniej (rys. 3). Linia ciągła przedstawia średnią liczbę studentów, linia zbudowana ze znaków „+" oznacza średnią powiększoną o tę część składowej, która wyjaśnia zmienność wspólną. Linia zbudowana ze znaków „–" oznacza średnią pomniejszoną o tą wartość.



Rys. 3. Odchylenia funkcjonalnych głównych składowych od średniej

Źródło: opracowanie własne z wykorzystaniem programu R.



Rys. 4. Rzut obiektów na płaszczyznę wyznaczoną przez dwie funkcje składowe

Źródło: opracowanie własne z wykorzystaniem programu R.

Pierwsza składowa odpowiada za ogólną tendencję. Dodatni ładunek na tej składowej oznacza, że krzywa opisująca liczbę studentów danego typu uczelni leży powyżej średniej. Druga składowa pokazuje tendencje w pierwszych latach w odniesieniu do ostatnich („początek kontra koniec") – porównuje okres do roku 2008 z okresem po 2008 roku. Dodatni ładunek na drugiej składowej oznacza, że liczba studentów na początku badanego okresu była mniejsza od średniej, natomiast na końcu liczba była większa od średniej.

Funkcjonalna analiza czynnikowa pozwala na wizualizację danych umożliwiającą porównanie badanych obiektów. Rysunek 4 zawiera rzut danych na płaszczyznę wyznaczoną przez dwie funkcjonalne główne składowe.

5. Zakończenie

Podjęcie decyzji w kolejnych krokach funkcjonalnej analizy głównych składowych zależy od natury danych, celu badania, wiedzy badacza na temat badanego zjawiska oraz jego doświadczenia. Wybór sposobu wyznaczania funkcji $x_i(t)$ oraz funkcji bazowej w dużej mierze zależy od natury danych. W przypadku $x_i(t)$ najczęściej wykorzystuje się kombinację liniową funkcji bazowych, natomiast funkcje bazowe często są funkcjami B-spline. Wybór podejścia do wygładzania danych najczęściej zależy od wykorzystywanego programu, np. w programie R wygładzanie przeprowadza się przed FPCA.

Takie decyzje podjęto w przedstawionym przykładzie empirycznym. Badanie miało na celu porównanie liczby studentów w różnych typach szkół wyższych w Polsce w latach 2000-2013. Na podstawie rysunku 4 można zauważyć, że wyższe szkoły techniczne mają dodatnie ładunki na obu składowych głównych, co oznacza, że liczba studentów jest wyższa od średniej oraz sytuacja w późniejszych latach jest lepsza niż na początku badanego okresu. Liczba studentów wyższych szkół ekonomicznych jest większa od średniej, jednak sytuacja w początkowych latach (do 2008) była lepsza niż na końcu badanego okresu. Uniwersytety medyczne mają liczbę studentów poniżej średniej, jednak sytuacja po 2008 roku jest lepsza niż przed tym rokiem.

Literatura

Besse P., Ramsay J.O. [1986], *Principal components analysis of sampled functions,* Psychometrika, 51, 285-311.

Daniele M. [2006], *Functional principal components analysis to study environmental data*, artykuł dostępny pod adresem http://www.sis-statistica.it/files/pdf/atti/Spontanee%202006_677-680.pdf.

Hall P., Hosseini-Nasab M. [2006], *On properties of functional principal components analysis*, Journal of the Royal Statistical Society, Series B (Statistical Methodology), vol. 68, no. 1, s. 109-126.

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, no. 3, s. 1493-1517.

Ingrassia S., Costanzo G.D. [2005], *Functional principal component analysis of financial time series*, [w:] Vichi M., Monari P., Mignani S., Montanari A. (red.), *New Developments in Classification and Data Analysis*, Springer-Verlag, Berlin, s. 351-358.

Krzyśko M., Górecki T., Deręgowski K. [2012], *Jądrowa i funkcjonalna analiza składowych głównych*, spotkanie PTS o. w Poznaniu, prezentacja dostępna na stroniehttp://stat.gov.pl/cps/rde/xbcr/pts/Krzysko_wyklad_7_11_12.pdf (data dostępu 1.03.2015).

Ramsay J.O., Dalzell C. [1991], *Some tools for functional data analysis (with discussion)*, Journal of the Royal Statistical Society 53: 539-72.

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

Ramsay J.O., Silverman B.W. [1997], *Functional Data Analysis,* Springer, New York.

Rice J.A., Silverman B.W. [1991], *Estimating the mean and covariance structure nonparametrically when the data are curves*, J. Roy. Statist. Soc. Ser. B, 53 233-243.

Shang H.L. [2011], *A survey of functional principal component analysis*, Working Paper 06/11, Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia.

Silverman B.W. [1995], *Incorporating parametric effects into functional principal components analysis*, Journal of the Royal Statistical Society, Series B, 57, 673-689.

Silverman B.W. [1996], *Smoothed functional principal components analysis by choice of norm*, Ann. Statist. 24 1-24.

DECISION PROBLEMS IN FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS

Summary: The functional principal components analysis combines advantages of the standard principal components analysis and enables analyzing data with dynamic nature. The main difference in both methods is the type of data: the PCA is based on multivariate data, whereas the FPCA on the functional data including curves and trajectories, i.e. a series of individual observations, not a single observation, as usual. The purpose of this article is to characterize the functional stages of principal component analysis with a special discussion of the steps that are not present in the classical principal component analysis.

Keywords: functional data, functional principal components analysis, longitudinal data.