Tomasz GÓRECKI, Ewa ŁAŹNIEWSKA

Funkcjonalna analiza składowych głównych PKB

Coraz bardziej złożone zagadnienie analiz regionalnych (związane z koniecznością uwzględniania większych terytoriów oraz większej liczby jednostek statystycznych) przyczynia się do poszukiwania nowszych metod analizy, a także wizualizacji zjawisk — również takich, które pokazują ich dynamikę. Zaprezentowana w artykule metoda może być pomocna w określaniu konkurencyjności regionalnej, polegającej na porównywaniu charakterystyki uwzględniającej cechy brane pod uwagę w *benchmarkingu* regionów.

Postęp technologiczny, zwłaszcza w zakresie techniki komputerowej, znacznie ułatwił analizę danych wysokowymiarowych lub danych z powtarzanymi pomiarami. Jeśli informacje takie zbierane są dostatecznie często w jakimś przedziale czasu I, np. przez urządzenie pomiarowe, nazywa się je danymi funkcyjnymi, gdzie dla każdego przypadku mamy jedną krzywą. Jest to wykorzystywane w przypadku, gdy zjawisko obserwujemy z błędem, ponieważ wygładzanie może znacznie zredukować efekt występowania szumu. W takich przypadkach możemy całą krzywą traktować jako wykres funkcji $X_i(t)$, którą traktujemy jak funkcję ciągłą, pomimo że czas jest dyskretny. Statystyczna analiza n takich krzywych jest powszechnie nazywana **funkcjonalną analizą danych (FDA[1])**[2]. FDA jest zatem metodą analizy, która pozwala odpowiedzieć na pytania typu: *jak różnią się dwie krzywe między sobą?* Unikalną cechą tego typu analizy jest możliwość wykorzystania informacji o szybkości zmian funkcji i jej pochodnych. Możemy wykorzystać wszystkie informacje, z których otrzymujemy gładkie funkcje, a których nie mamy używając jedynie dyskretnych danych. Tego typu analiza, której głównym celem jest dostarczenie narzędzi służących do opisu i modelowania zbioru krzywych jest obecnie celem coraz większego zainteresowania społeczności statystycznej.

W książkach Ramsaya i Silvermana (2002, 2005) możemy znaleźć opis procedur radzenia sobie z danymi funkcyjnymi. Podane tam metody znalazły szerokie zastosowanie m.in. w chemometrii, klimatologii, biologii oraz ekonomii. Statystycy w pierwszym kroku zazwyczaj próbują tak zredukować dane, aby zachować maksymalnie dużo informacji przy minimalnej liczbie zmiennych. Optymalne byłoby uzyskanie dwóch lub trzech zmiennych, które można w prosty sposób wizualizować.

Z kolei za pomocą **funkcjonalnej analizy składowych głównych (FPCA[3])** możemy dokonać rzutu wysokowymiarowych danych na przestrzeń o dużo

[1] Ang. *Functional Data Analysis*.
[2] Ramsay J. O., Dalzell C. J. (1991), s. 539—572.
[3] Ang. *Functional Principal Components Analysis*.

mniejszym wymiarze, jednocześnie zachowując maksymalnie dużo informacji (w tym przypadku zmienności danych). Główna idea metody polega na znalezieniu takich funkcji, których iloczyn skalarny[4] z danymi pozwala otrzymać maksymalną zmienność. Pierwsza składowa wyjaśnia najwięcej zmienności, druga jest prostopadła do pierwszej i wyjaśnia maksymalnie dużo z tego, co zostało itd.

Wkład każdej kolejnej składowej w ogólną zmienność początkowych zmiennych jest nie tylko znany, ale i coraz mniejszy, od pewnego momentu składowe będą wyjaśniać znikomy zakres zmienności. Z punktu widzenia badacza nie będą już zatem wnosiły istotnych informacji i mogą być pominięte w dalszej analizie. Rozpatrywane dane można w ten sposób zredukować, przy jednoczesnej pełnej kontroli badacza nad stopniem utraty informacji. Większość zmienności zachowana jest wtedy przez kilka pierwszych funkcjonalnych składowych głównych. Metoda konstrukcji takich składowych opisana przez Ramsaya i Silvermana (2005) znalazła wiele zastosowań. Ze względu na sposób konstrukcji funkcjonalnych składowych głównych są one wzajemnie nieskorelowane, a zatem można je również wykorzystać w dalszej analizie z użyciem tych metod, w których wymagane jest założenie o braku współliniowości (jak np. analiza dyskryminacyjna, analiza regresji wielorakiej). Jednym z takich zastosowań jest estymacja losowych trajektorii z dostępnych danych[5].

METODA KONSTRUKCJI FUNKCJONALNYCH SKŁADOWYCH GŁÓWNYCH

Załóżmy, że obserwujemy proces losowy $X(t)_{t\in I} \in L^2(I)$, gdzie $L^2(I)$ jest przestrzenią Hilberta[6] funkcji całkowalnych z kwadratem w przedziale czasu I, wyposażoną w iloczyn skalarny. Dodatkowo załóżmy, że $\forall_{t\in I} EX(t) = 0$ (tzn. proces jest scentrowany) oraz ma skończoną wariancję (zmienność).

Możemy założyć, że taki proces losowy $X(t)$ jest reprezentowany przez skończoną liczbę funkcji ortonormalnych. Wyrazimy to wzorem[7]:

$$X(t) = \sum_{k=0}^{K} c_k \phi_k(t) = c'\phi(t) \quad t \in I \tag{1}$$

[4] W przypadku przestrzeni $L^2(I)$ iloczyn skalarny funkcji f oraz g wyraża się wzorem:

$$\langle f, g \rangle = \int_I f(t)\overline{g(t)}\, dt$$

[5] Ingrassia S., Constanzo G. D. (2005), s. 351—358.

[6] Zauważmy, że jeśli skupimy się na funkcjach mających ograniczone normy, to wystarcza założenie, aby $L^2(I)$ była jedynie przestrzenią Hilberta. Jednakże w wielu przypadkach jest to niewystarczające. W takiej sytuacji potrzebna jest nam przestrzeń Hilberta z jądrem reprodukującym (Wahba, 1990).

[7] Pierwszy znak w równości jest formalnie niepoprawny — w praktyce aproksymujemy $X(t)$. Jednakże dla zachowania prostoty zapisu będziemy pomijać różnicę pomiędzy prawdziwym $X(t)$ i jego przybliżeniem.

gdzie:

$\{\phi_k\}$ — $(K+1)$ pierwsze elementy bazy ortonormalnej w przestrzeni $L^2(I)$,

$\{c_k\}$ — zmienne losowe z wartością oczekiwaną równą zero ($E(c)=0$) oraz skończoną, niezerową wariancją ($\mathrm{Var}(c)=\mathbf{\Sigma}$).

Macierz $\mathbf{\Sigma}$ jest nieznana, dlatego użyjemy estymatora $\hat{\mathbf{\Sigma}}$ opartego na N niezależnych realizacjach wektora losowego c:

$$\hat{C} = \begin{bmatrix} \hat{c}_{10} & \cdots & \hat{c}_{1K} \\ \vdots & \ddots & \vdots \\ \hat{c}_{N0} & \cdots & \hat{c}_{NK} \end{bmatrix} \tag{2}$$

gdzie $\hat{c}_{ik}$ — oszacowania parametrów c_{ik} uzyskane metodą najmniejszych kwadratów w reprezentacji:

$$x_i(t) = \sum_{k=0}^{K} c_{ik}\phi_k(t) \quad t \in I, \quad i = 1, 2 ..., N \tag{3}$$

procesu losowego $X(t)$.

Wygładzenie liniowe otrzymujemy, jeśli wyznaczymy współczynniki rozwinięcia c_k poprzez minimalizację kryterium najmniejszych kwadratów. W taki sposób otrzymujemy następujące rozwiązanie (Ramsay, Silverman, 2005):

$$\hat{c} = (\phi'\phi)^{-1}\phi' x \tag{4}$$

Estymator $\hat{\Sigma}$ nieznanej macierzy $\mathbf{\Sigma}$ ma postać:

$$\hat{\Sigma} = \frac{1}{N}\hat{C}'\hat{C} \tag{5}$$

Teraz możemy już wyznaczyć wartości własne $\hat{\lambda}_k$ (wszystkie są niezerowe i różne) oraz odpowiadające im wektory własne $\hat{u}_k$ macierzy $\hat{\Sigma}$. Następnie możemy znaleźć funkcje wagowe:

$$\hat{u}_k(t) = \hat{u}_k'\phi(t) \quad t \in I \tag{6}$$

Ostatecznie współrzędna rzutu i-tej realizacji procesu $X(t)$ na j-tą funkcjonalną składową główną ma postać:

$$\hat{U}_{ij} = \sum_{k=0}^{K} \hat{c}_{ik}\hat{u}_{jk} = \hat{c}_i'\hat{u}_j \quad i = 1, 2, ..., N, \quad j = 1, 2, ... \tag{7}$$

W tej formule wykorzystujemy klasyczny iloczyn skalarny wektorów. Poziom (siła) wygładzenia zależy od K, stąd małe (duże) wartości K dają bardziej (mniej) wygładzone krzywe. Optymalna liczba elementów bazy K jest wyznaczana za pomocą bayesowskiego kryterium informacyjnego[8] (BIC) dla każdej funkcji $x_i(t)$ osobno, następnie najczęściej występująca wartość (dominanta) wyznacza optymalną wartość K dla całego zbioru danych. Przyjęto, że K nie może przekraczać wartości 50[9]. Jako baza wybrana została ortonormalna baza wielomianów Legendre'a[10] w przestrzeni $L^2([-1,1])$, która ma postać:

$$\widetilde{P}_k(x) = \sqrt{\frac{2k+1}{2}} P_k(x) \tag{8}$$

gdzie:

$$P_{k+1}(x) = \frac{1}{k+1}[(2k+1)xP_k(x) - kP_{k-1}(x)], \quad k \geq 1,$$
$$P_0(x)=1,$$
$$P_1(x)=x.$$

Najczęściej rozpatrywane obiekty przedstawiane są w układzie dwóch pierwszych funkcjonalnych składowych głównych. W tym przypadku kryterium jakości skonstruowanych funkcjonalnych składowych głównych jest wyrażenie (Hastie i in., 2009):

$$\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{\sum \hat{\lambda}_i} 100\% \tag{9}$$

gdzie $\hat{\lambda}_i \geq \hat{\lambda}_{i+1}, \quad i = 1, 2, \ldots$ — niezerowe wartości własne macierzy $\hat{\Sigma}$.

Im wyrażenie to jest większe, tym większą zmienność wykazują dwie pierwsze funkcjonalne składowe główne, a co za tym idzie, jakość reprezentacji jest lepsza.

---

[8] Jest to kryterium wyboru modelu spośród skończonego zbioru modeli, który oparty jest na funkcji największej wiarygodności i jest ściśle związane z kryterium informacyjnym Akaike (AIC). Podczas dopasowywania modeli można podnieść wartość funkcji największej wiarygodności poprzez zwiększenie liczby parametrów, powodując często przeuczenie modelu (model jest idealnie dopasowany do danych uczących, ale nie ma własności generalizacji, zatem nie sprawdza się na innych danych tego samego typu). Kryteria informacyjne rozwiązują ten problem poprzez wprowadzenie kary zależnej od liczby użytych w modelu parametrów (tzw. kary za złożoność), przy czym w kryterium BIC ta kara jest wyższa (McQuarrie, Tsai, 1998).

[9] Liczba funkcjonalnych składowych głównych jest mniejsza od N (lub N–1, gdy dane zostały scentrowane). Jeśli jednak $K<N$, to liczba funkcjonalnych składowych głównych jest równa K.

[10] Wybrano bazę wielomianów Legendre'a ponieważ uzyskuje się dla niej najlepsze wyniki (Górecki, Krzyśko, 2012, s. 71—87) — porównywane były bazy ortonormalne: Fouriera, cosinusów, sinusów oraz baza wielomianów Legendre'a.

ZASTOSOWANIE FPCA W ANALIZACH REGIONALNYCH

Rozpatrywane były dane dotyczące PKB *per capita* w latach 1999—2008 w Polsce. Dane te pokazano (dla skrajnych lat) na wykr. 1 i 2. Kody na wykresach oznaczają regiony Polski, rozpoczynają się dwuliterowym kodem naszego kraju, kolejne poziomy podziału oznaczane są cyframi. Jeśli istnieje więcej niż 9 jednostek numeracja jest kontynuowana z użyciem wielkich liter poczynając od A.

W takiej sytuacji mamy $N = 66$ szeregów czasowych o długości $T = 10$. Na wykr. 3 przedstawiono trajektorie tych danych, czerwoną pogrubioną linią zaznaczono trajektorię średniej w badanym okresie zaprezentowanego zagadnienia. Możemy zatem zaobserwować, jak zachowuje się badany wskaźnik w ciągu dekady. Niestety dość trudne jest określenie, które regiony zmieniały się podobnie w rozważanym okresie pod względem PKB *per capita*.

Kolejny wykres (4), to wykres piargowy[11] (osypisko) dla FPCA[12]. Widzimy, że pierwsza funkcjonalna składowa główna wyjaśnia 93,31% zmienności, podczas gdy druga 5,33%. W sumie dwie pierwsze funkcjonalne składowe główne wyjaśniają aż 98,64% zmienności[13]. W takiej sytuacji rzut na płaszczyznę skonstruowaną za pomocą dwóch pierwszych funkcjonalnych składowych głównych wydaje się być bardzo dobrym rozwiązaniem. Algorytm wybrał K=11 elementów ortonormalnej bazy wielomianów Legendre'a w przestrzeni $L^2([-1,1])$.

Wykres 5 przedstawia dwie pierwsze funkcje wagowe $u_k(t)$, k=1,2, dla K=11 elementów ortonormalnej bazy wielomianów Legendre'a w przestrzeni $L^2([-1,1])$. Pokazują one, w jaki sposób taki zestaw danych funkcjonalnych oscyluje wokół średniej, czyli w zasadzie potwierdzają zaobserwowane wcześniej trendy. Zauważmy jednak, że obie funkcje (linie ciągłe) przyjmują wartości ujemne (i są malejące), natomiast wyjściowe dane przyjmowały jedynie wartości dodatnie (i były rosnące). Wyjaśnienie tego fenomenu leży w pewnej, niewielkiej, niejednoznaczności wyznaczonych składowych głównych. Mianowicie, jak już wspomniano, składowe wyznaczone są jednoznacznie z dokładnością do znaku. Gdybyśmy zmienili znaki wyznaczonych składowych, otrzymalibyśmy funkcje (linie przerywane) cały czas dodatnie (i rosnące), które dobrze wpasowują się w trend pokazany na wykr. 3. Widzimy więc, że pierwsza funkcja wagowa odpowiada za trend, natomiast druga za odchylenia okresowe od średniej.

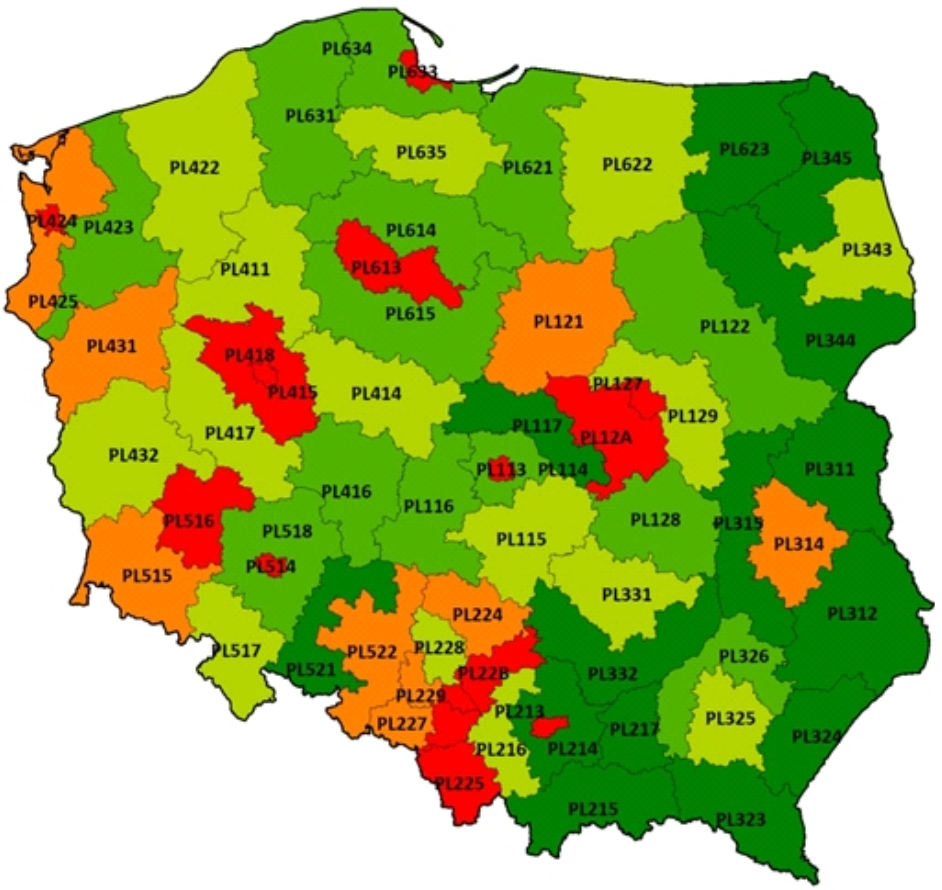
---

[11] Wykres pokazujący kolejne wartości własne. Kolejne funkcjonalne składowe główne wyodrębniają coraz mniejszą część zmienności, to i spadki wartości na wykresie są coraz mniejsze. Test osypiska polega na znalezieniu miejsca, na prawo od którego regularny spadek wartości własnych staje się wolniejszy. Nie należy wyodrębniać więcej czynników niż znajduje się po lewej stronie tego punktu.

[12] Całość obliczeń została wykonana w programie Matlab R2011a (Ramsay, Matlab, 2005).

[13] W przypadku klasycznej analizy składowych głównych otrzymujemy, że pierwsza składowa główna wyjaśnia 77,34% zmienności, natomiast druga — 8,21%. W sumie składowe te wyjaśniają 85,55% zmienności. Co jest wynikiem znacznie gorszym od metody FPCA.

Wykr. 1. PKB *PER CAPITA* W 1999 R.

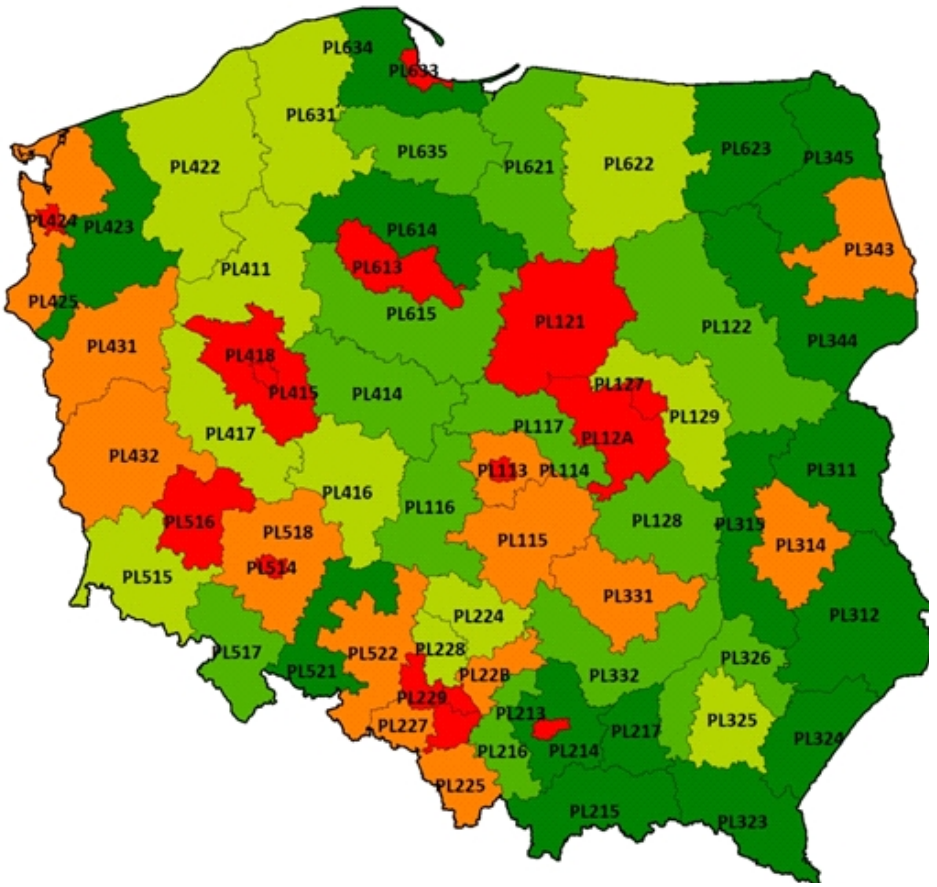


PKB *per capita*

od 9000 do 25400 (15)
od 7700 do 9000 (9)
od 6800 do 7700 (15)
od 6100 do 6800 (13)
od 5000 do 6100 (14)

Ź r ó d ł o: opracowanie własne.

Wykr. 2. PKB *PER CAPITA* W 2008 R.



PKB *per capita*

od 14900 do 41600 (15)

od 12000 do 14900 (13)

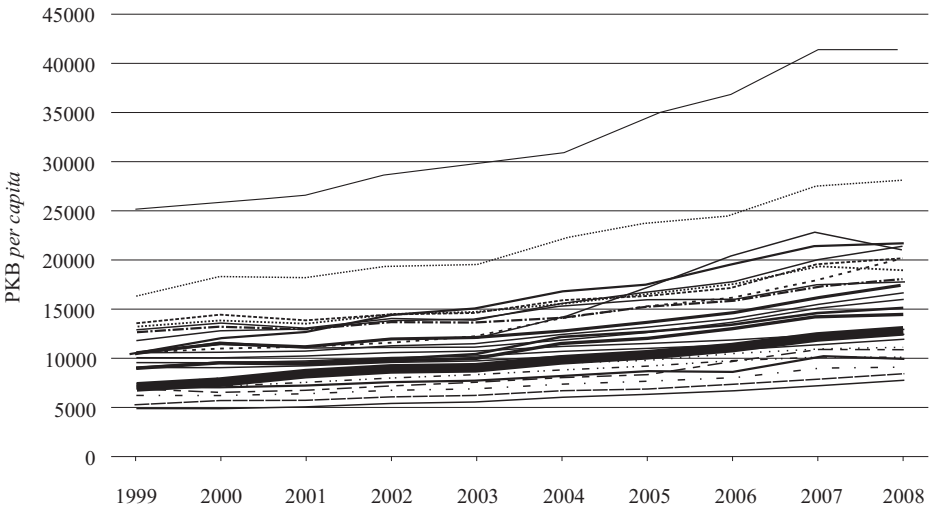
od 11000 do 12000 (11)

od 9900 do 11000 (12)

od 8100 do 9900 (15)

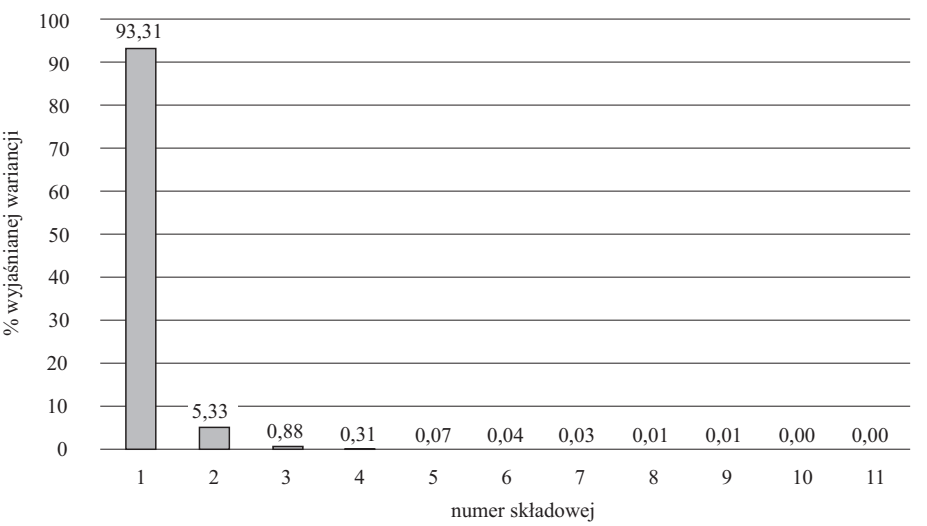
Źródło: jak przy wykr. 1.

Wykr. 3. TRAJEKTORIE PKB *PER CAPITA*



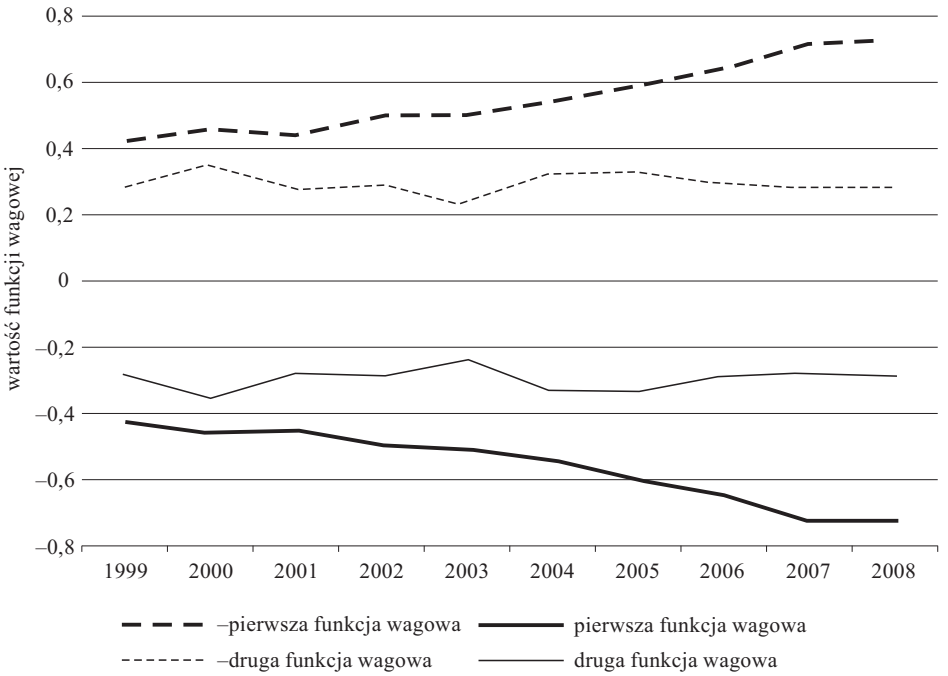
Ź r ó d ł o: jak przy wykr. 1.

Wykr. 4. WYKRES PIARGOWY (osypiska)



Ź r ó d ł o: jak przy wykr. 1.

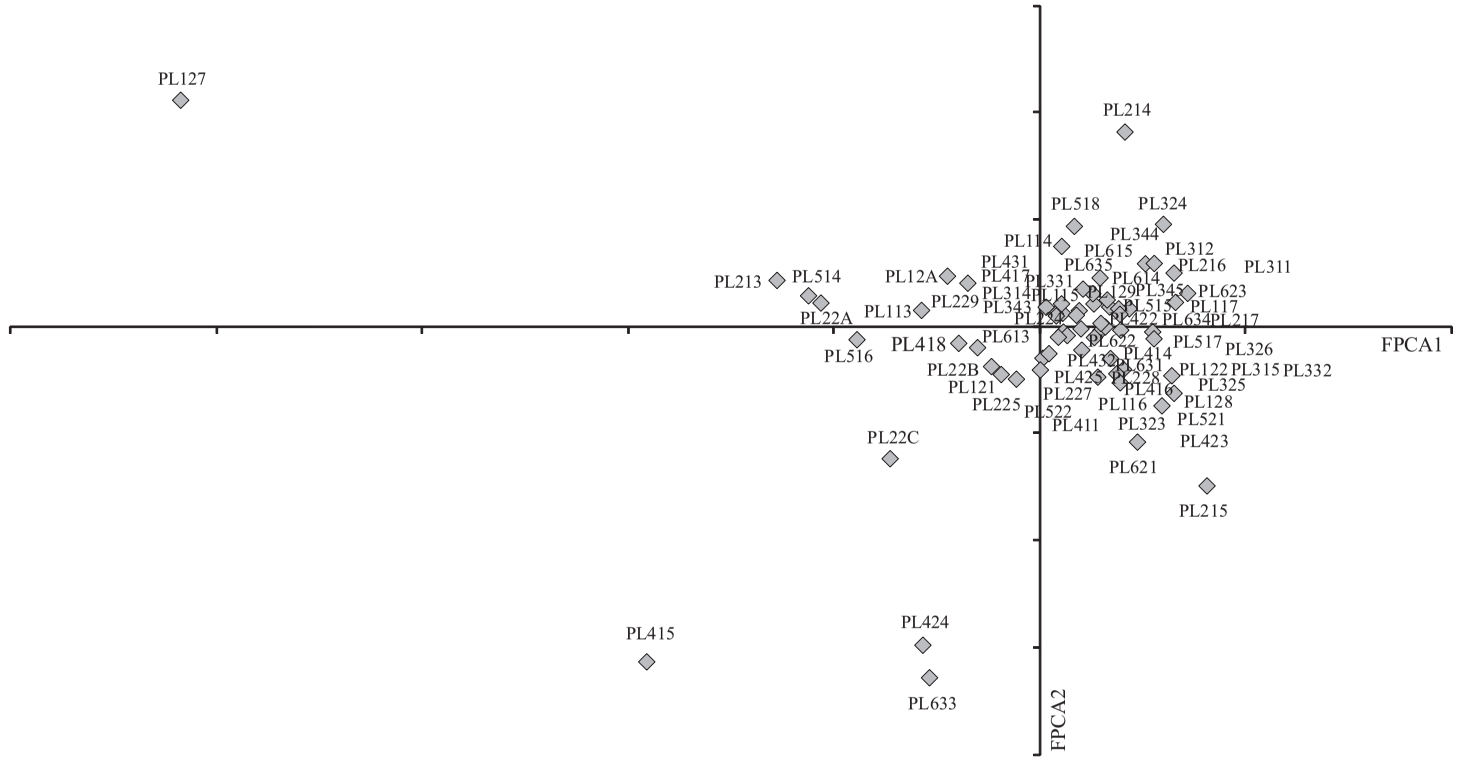
Wykr. 5. DWIE PIERWSZE FUNKCJE WAGOWE



Ź r ó d ł o: jak przy wykr. 1.

Na koniec pozostaje najważniejsza wizualizacja. Mianowicie, rzut danych na przestrzeń dwóch pierwszych funkcjonalnych składowych głównych, który ma nam pokazać, które regiony zachowywały się podobnie w rozpatrywanym okresie pod względem zmian oraz poziomu PKB *per capita*. Jeśli punkty leżą blisko siebie oznacza to, że regiony zachowywały się pod tym względem podobnie. Na wykr. 6 widzimy, że np. Szczecin (PL424) oraz podregion trójmiejski (PL633) zachowywały się w badanym okresie pod względem poziomu oraz zmian PKB *per capita* podobnie, natomiast stolica Polski (PL127) wyróżniała się zdecydowanie.

Wykr. 6. RZUT DANYCH NA UKŁAD DWÓCH PIERWSZYCH FUNKCJONALNYCH SKŁADOWYCH GŁÓWNYCH



Ź r ó d ł o: jak przy wykr. 1.

Wnioski końcowe

Przydatność tej metody do analiz ekonomicznych wydaje się być niezmiernie ciekawa i ważna ze statystycznego punktu widzenia. Trudno wskazać wszystkie możliwe zastosowania, ale zauważmy, że jednym z nich może być analiza stóp zwrotu walut czy spółek giełdowych. Można stwierdzić, że użyteczność tej metody w analizach regionalnych jest istotna, ponieważ na jej podstawie dokonujemy porównań w zachowaniu poszczególnych obiektów analizy z perspektywy analizowanego okresu. W naszym przypadku była to analiza podregionów w okresie 1999—2008. Z obserwacji tej wynika, że regiony o podobnym poziomie zjawiska wykazywały podobieństwo w utrzymaniu lub zmianie poziomu PKB *per capita*, a regiony o wyższym PKB *per capita*, np. podregion stołeczny, zachowywały się zupełnie inaczej. Takie wnioski mogą okazać się przydatne do formułowania zaleceń w polityce regionalnej.

Przydatność tej analizy znajduje także odzwierciedlenie w badaniu konkurencyjności regionalnej. Naszym zdaniem, uzyskane wyniki pokazują użyteczność opisanej metodologii w wizualizacji oraz wnioskowaniu na temat ekonomicznych szeregów czasowych z nieco innej perspektywy niż było to czynione do tej pory przy użyciu powszechnie wykorzystywanych metod, jak chociażby klasyczna analiza składowych głównych czy analiza czynnikowa. Klasyczne metody, pomimo że prostsze, nie pokazują jednak pełnego obrazu sytuacji. Z jednej strony mogą one pomijać fakt, że dane mają postać szeregu czasowego i z tego powodu istotna jest kolejność zmiennych (np. klasyczna analiza składowych głównych). Z drugiej strony mogą nie mieć statystycznych własności (np. szereg ciekawych wizualizacji), które daje nam zaproponowana metoda.

---

dr Tomasz Górecki — *Uniwersytet im. Adama Mickiewicza w Poznaniu*, **dr hab. Ewa Łaźniewska** — profesor *Uniwersytetu Ekonomicznego w Poznaniu*

LITERATURA

Górecki T., Krzyśko M. (2012), *Functional Principal Components Analysis*, [w:] *Data analysis methods and its applications*, red. J. Pociecha, R. Decker, Wydawnictwo C. H. Beck

Hastie T., Tibshirani R., Friedman J. (2009), *The Elements of Statistical Learning*, Springer

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

McQuarrie A. D. R., Tsai C. L. (1998), *Regression and Time Series Model Selection*, World Scientific

Ramsay J., Matlab R. (2005), *S-Plus Functions for Functional Data Analysis*, McGill University, ftp://ego.psych.mcgill.ca/pub/FDAfuns.pdf

Ramsay J. O., Dalzell C. J. (1991), *Some tools for functional data analysis,* „Journal of the Royal Statistical Society", Series B (Statistical Methodology), No. 53

Ramsay J. O., Silverman B. W. (2002), *Applied Functional Data Analysis: methods and case studies*, Springer

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

Wahba G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia

SUMMARY

The article describes the use of functional GDP principal components analysis from the exploratory point of view. This analysis is designed to show the variation in the entire sample, not only discrete observations. This is a technique that is often used as an introduction (e.g., dimensionality reduction and data visualization) for further analysis. This approach emphasizes the significant characteristics of associated statistically data sets and provides useful tools for statistical analysis of economic data sets. The study concerns the GDP per capita in Poland. The results showed that the two main functional components explain 98,64% of the variation. This means that the presented visualization and interpretation of the results is reliable.

РЕЗЮМЕ

В статье характеризуется использование функционального анализа основных компонентов ВВП с исследовательской точки зрения. Этот анализ должен показать изменения во всей выборке, а не только дискретные наблюдения. Эту технику часто используют в качестве введения (напр. сокращение размерности или визуализация данных) для дальнейшего анализа. Такой подход подчеркивает важные статистические характеристики связанных наборов данных, а также предоставляет полезные инструменты для статистического анализа экономических наборов данных. Исследование касается ВВП на душу населения в Польше. Полученные результаты показали, что два основных функциональных компонента объясняют 98,64% вариаций, то есть представленная визуализация и интерпретация результатов является надежной.