PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 426

Taksonomia 26
Klasyfikacja i analiza danych – teoria i zastosowania

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2016

Redaktor Wydawnictwa: Agnieszka Flasińska

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 stronach internetowych
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

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

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Wstęp

W dniach 14–16 września 2015 r. w Hotelu Novotel Gdańsk Marina w Gdańsku odbyła się XXIV Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS (XXIX Konferencja Taksonomiczna) „Klasyfikacja i analiza danych – teoria i zastosowania", zorganizowana przez Sekcję Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego oraz Katedrę Statystyki Wydziału Zarządzania Uniwersytetu Gdańskiego. Przewodniczącymi Komitetu Organizacyjnego konferencji byli prof. dr hab. Mirosław Szreder oraz dr hab. Krzysztof Najman, prof. nadzw. UG, sekretarzami naukowymi dr hab. Kamila Migdał-Najman, prof. nadzw. UG oraz dr hab. Anna Zamojska, prof. nadzw. UG, a sekretarzem organizacyjnym Anna Nowicka z Fundacji Rozwoju Uniwersytetu Gdańskiego.

Konferencja Naukowa została dofinansowana ze środków Narodowego Banku Polskiego.

Zakres tematyczny konferencji obejmował takie zagadnienia, jak:

a) teoria (taksonomia, analiza dyskryminacyjna, metody porządkowania liniowego, metody statystycznej analizy wielowymiarowej, metody analizy zmiennych ciągłych, metody analizy zmiennych dyskretnych, metody analizy danych symbolicznych, metody graficzne),

b) zastosowania (analiza danych finansowych, analiza danych marketingowych, analiza danych przestrzennych, inne zastosowania analizy danych – medycyna, psychologia, archeologia, itd., aplikacje komputerowe metod statystycznych).

Zasadniczymi celami konferencji SKAD były prezentacja osiągnięć i wymiana doświadczeń z zakresu teoretycznych i aplikacyjnych zagadnień klasyfikacji i analizy danych. Konferencja stanowi coroczne forum służące podsumowaniu obecnego stanu wiedzy, przedstawieniu i promocji dokonań nowatorskich oraz wskazaniu kierunków dalszych prac i badań.

W konferencji wzięło udział 81 osób. Byli to pracownicy oraz doktoranci następujących uczelni i instytucji: AGH w Krakowie, Politechniki Łódzkiej, Politechniki Gdańskiej, Politechniki Opolskiej, Politechniki Wrocławskiej, Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Szkoły Głównej Handlowej w Warszawie, Uniwersytetu im. Adama Mickiewicza w Poznaniu, Uniwersytetu Ekonomicznego w Katowicach, Uniwersytetu Ekonomicznego w Krakowie, Uniwersytetu Ekonomicznego w Poznaniu, Uniwersytetu Ekonomicznego we Wrocławiu, Uniwersytetu Gdańskiego, Uniwersytetu Jana Kochanowskiego w Kielcach, Uniwersytetu Łódzkiego, Uniwersytetu Mikołaja Kopernika w Toruniu, Uniwersytetu Przyrodniczego w Poznaniu, Uniwersytetu Szczecińskiego, Uniwer-

sytetu w Białymstoku, Wyższej Szkoły Bankowej w Toruniu, a także przedstawiciele NBP i PBS Sp. z o.o.

W trakcie dwóch sesji plenarnych oraz trzynastu sesji równoległych wygłoszono 58 referatów poświęconych aspektom teoretycznym i aplikacyjnym zagadnienia klasyfikacji i analizy danych. Odbyła się również sesja plakatowa, na której zaprezentowano 14 plakatów. Obradom w poszczególnych sesjach konferencji przewodniczyli profesorowie: Józef Pociecha, Eugeniusz Gatnar, Tadeusz Trzaskalik, Krzysztof Jajuga, Marek Walesiak, Barbara Pawełek, Feliks Wysocki, Ewa Roszkowska, Andrzej Sokołowski, Andrzej Bąk, Tadeusz Kufel, Mirosław Krzyśko, Krzysztof Najman, Małgorzata Rószkiewicz, Mirosław Szreder.

Teksty 25 recenzowanych artykułów naukowych stanowią zawartość prezentowanej publikacji z serii „Taksonomia" nr 26. Pozostałe recenzowane artykuły znajdują się w „Taksonomii" nr 27.

W pierwszym dniu konferencji odbyło się posiedzenie członków Sekcji Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego, któremu przewodniczył prof. dr hab. Józef Pociecha. Ustalono plan przebiegu zebrania obejmujący następujące punkty:

A. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS.
B. Informacje dotyczące planowanych konferencji krajowych i zagranicznych.
C. Organizacja konferencji SKAD PTS w latach 2016 i 2017.
D. Wybór przedstawiciela Rady Sekcji SKAD PTS do IFCS.
E. Dyskusja nad kierunkami rozwoju działalności Sekcji.

Prof. dr hab. Józef Pociecha otworzył posiedzenie Sekcji SKAD PTS. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS przedstawiła sekretarz naukowy Sekcji dr hab. Barbara Pawełek, prof. nadzw. UEK. Poinformowała, że obecnie Sekcja liczy 231 członków. Przypomniała, że na stronie internetowej Sekcji znajdują się regulamin, a także deklaracja członkowska. Poinformowała, że zostały opublikowane zeszyty z serii „Taksonomia" nr 24 i 25 (PN UE we Wrocławiu nr 384 i 385). W „Przeglądzie Statystycznym" (zeszyt 4/2014) ukazało się sprawozdanie z ubiegłorocznej konferencji SKAD, która odbyła się w Międzyzdrojach, w dniach 8–10 września 2014 r. Prof. Barbara Pawełek przedstawiła także informacje dotyczące działalności międzynarodowej oraz udziału w ważnych konferencjach członków i sympatyków SKAD.

W konferencji Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS – International Federation of Classification Societies) w dniach 6–8 lipca 2015 r. w Bolonii, zorganizowanej przez Università di Bologna, udział wzięło 19 osób z Polski (w tym 17 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 79,0%). Ponadto prof. Józef Pociecha był członkiem Komitetu Naukowego Konferencji z ramienia SKAD, członkiem Międzynarodowego Komitetu Nagród IFCS oraz organizatorem i przewodniczącym sesji nt. „Classification models for forecasting of economic processes".

W konferencji „European Conference on Data Analysis" (Colchester, 2–4 września 2015 r.) zorganizowanej przez The German Classification Society (GfKl) we współpracy z The British Classification Society (BCS) i Sekcją Klasyfikacji i Analizy Danych PTS (SKAD) udział wzięło 18 osób z Polski (w tym 14 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 66,0%). Ponadto profesorowie Krzysztof Jajuga oraz Józef Pociecha byli członkami Komitetu Naukowego konferencji, prof. Andrzej Dudek został poproszony przez organizatorów o przygotowanie referatu i wygłoszenie na Sesji Plenarnej „Cluster analysis in XXI century, new methods and tendencies", prof. Krzysztof Jajuga był przewodniczącym sesji plenarnej, przewodniczącym sesji nt. „Finance and economics II" oraz organizatorem i przewodniczącym sesji nt. „Data analysis in finance", prof. Józef Pociecha był organizatorem i przewodniczącym sesji nt. „Outliers in classification procedures – theory and practice", prof. Andrzej Dudek był przewodniczącym sesji nt. „Machine learning and knowledge discovery II".

Kolejny punkt posiedzenia Sekcji obejmował zapowiedzi najbliższych konferencji krajowych i zagranicznych, których tematyka jest zgodna z profilem Sekcji. Prof. dr hab. Józef Pociecha poinformował o dwóch wybranych konferencjach krajowych (były to XXXIV Konferencja Naukowa „Multivariate Statistical Analysis MSA 2015", Łódź, 16–18 listopada 2015 r. i X Międzynarodowa Konferencja Naukowa im. Profesora Aleksandra Zeliasia nt. „Modelowanie i prognozowanie zjawisk społeczno-gospodarczych", Zakopane, 10–13 maja 2016 r.) oraz o trzech wybranych konferencjach zagranicznych. Konferencja „European Conference on Data Analysis" odbędzie się na Uniwersytecie Ekonomicznym we Wrocławiu w dniach 26–28 września 2017 r. W przeddzień tej konferencji, tj. 25.09.2017 r., odbędzie się Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2017". Następna konferencja Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS) odbędzie się w 2017 r. w Tokio. W 2019 r. Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2019" organizuje prof. Andreas Geyer-Schultz w Karlsruhe.

W następnym punkcie posiedzenia podjęto kwestię organizacji kolejnych konferencji SKAD. SKAD 2016 zorganizuje Katedra Metod Statystycznych Wydziału Ekonomiczno-Socjologicznego Uniwersytetu Łódzkiego.

W kolejnej części zebrania dokonano wyboru przedstawiciela Rady Sekcji SKAD PTS do IFCS na kadencję 2016–2019. Powołano Komisję Skrutacyjną, której przewodniczącym został prof. Tadeusz Kufel, a członkami dr hab. Iwona Konarzewska i dr Dominik Rozkrut. Profesor Józef Pociecha poprosił zebranych o proponowanie kandydatur zgłaszając jednocześnie prof. Andrzeja Sokołowskiego. Wobec braku następnych kandydatur listę zamknięto. Komisja Skrutacyjna przeprowadziła głosowanie tajne. W głosowaniu uczestniczyło 41 członków Sekcji. Profesor Andrzej Sokołowski został przedstawicielem Rady Sekcji SKAD PTS do

IFCS na kadencję 2016–2019, uzyskując następujący wynik: 39 głosów na „tak", 1 głos na „nie", 1 głos był nieważny.

W ostatnim punkcie zebrania dyskutowano nad kierunkami rozwoju działalności Sekcji obejmującymi następujące problemy: udział w międzynarodowym ruchu naukowym (wspólne granty, publikacje), umiędzynarodowienie konferencji SKAD (uczestnicy zagraniczni, dwujęzyczność konferencji), wydawanie własnego czasopisma.

Profesor Józef Pociecha zamknął posiedzenie Sekcji SKAD.

Krzysztof Jajuga, Marek Walesiak

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 426 • 2016

Taksonomia 26 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

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

GRUPOWANIE DANYCH FUNKCJONALNYCH W ANALIZIE POZIOMU WIEDZY MATURZYSTÓW

FUNCTIONAL DATA CLUSTERING METHODS IN THE ANALYSIS OF HIGH SCHOOL GRADUATES' KNOWLEDGE

DOI: 10.15611/pn.2016.426.21

Streszczenie: W artykule przedstawiono różne podejścia do grupowania danych funkcjonalnych. Do głównych kategorii metod zaliczmy: metody, w których przeprowadza się redukcję wymiaru przed właściwym grupowaniem; nieparametryczne metody wykorzystujące odpowiednią odległość między krzywymi oraz metody bazujące na modelu bazowym. Zaprezentowano przykład grupowania danych funkcjonalnych, który obejmuje analizę poziomu wiedzy uczniów na IV etapie edukacji.

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

Abstract: The article presents different approaches to clustering functional data. The main methods include: dimension reduction before clustering, nonparametric methods using specifying distances between curves and model-based clustering methods. Numerical illustrations includes an analysis of the level of knowledge of students in the fourth stage of education.

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

1. Wstęp

Danymi funkcjonalnymi są krzywe i trajektorie, które są wyrażone za pomocą funkcji. Chociaż dane funkcjonalne często są zależne 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, Costanzo 2005; Daniele 2006].

Niech zmienna $\mathbf{y}_i = (y_i(t_1), y_i(t_2), \ldots, y_i(t_p))$ będzie próbkowym pomiarem zmiennej Y w czasie $t_1, t_2, \ldots, t_p$ dla i-tej jednostki. Dane y_i nazywane są surowymi danymi

funkcjonalnymi (*raw functional data*). Realizacje dyskretne przekształca się w funkcję ciągłą $x_i(t)$ [Hall, Hosseini-Nasab 2006; Ramsay, Silverman 2005].

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, $\phi(t)$ – funkcje tworzące bazę ortonormalną przestrzeni $L_2(I)$, oraz, $x_i(t) \in L_2(I)$, gdzie $L_2(I)$ – przestrzeń Hilberta funkcji całkowalnych z kwadratem na przedziale I wyposażoną 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($k\omega t$), cos($k\omega t$), ...
- 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:

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

Globalna miara aproksymacji dana jest wzorem:

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

Dane funkcjonalne powinny być **wygładzone**, wszelkie chropowatości funkcji traktowane są jako szum, który powinien być całkowicie usunięty. Jako kryterium dopasowania krzywej do danych obserwowalnych przyjmuje się minimalizację kwadratu błędu uzupełnioną o „karę" dla funkcji niewygładzonych:

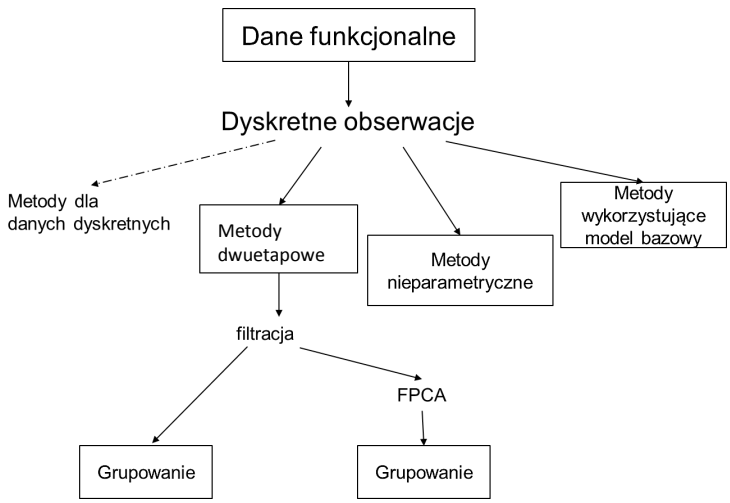
$$\|x_i - \hat{x}_i\|^2 + \alpha\|\hat{x}_i\|^2. \tag{4}$$

Obie normy niekoniecznie są takie same. Druga norma powinna być powiązana z pochodną drugiego rzędu $D^2[\hat{x}_i(t)]$, która jest miarą chropowatości funkcji [Hall, Müller, Wang 2006; Ramsay, Hooker, Graves 2009].

2. Metody wyodrębniania grup dla danych funkcjonalnych

Wyróżnia się cztery podejścia do grupowania danych funkcjonalnych (rys. 1) [Jacques, Preda 2014; Ramsay, Silverman 2005].

W pierwszym dokonuje się dyskretyzacji funkcji w pewnych punktach czasu. Dla dyskretnych obserwacji nie ma potrzeby przekształcania ich w funkcje ciągłe. W tym przypadku wykorzystuje się techniki grupowania dla danych wielowymiarowych.



Rys. 1. Metody wyodrębniania grup

Źródło: opracowanie na podstawie [Jacques, Preda 2014].

Wyróżnia się trzy kategorie metod stosowanych dla danych opisanych za pomocą funkcji. Pierwsza kategoria to **metody dwuetapowe**. W pierwszym etapie przeprowadzana jest filtracja, czyli redukcja wymiaru danych. Dokonuje się jej poprzez aproksymację danych za pomocą krzywych ze skończonego zbioru funkcji. W drugim etapie przeprowadza się właściwe grupowanie danych.

Do drugiej kategorii należą **metody nieparametryczne**. Wśród nich wyróżnia się:

- nieparametryczne techniki (np. *k*-means, hierarchical clustering), w których wykorzystuje się następującą odległość między krzywymi x_i, y_i:

$$d_l(x_i, y_i) = \sqrt{\int \left(x_i^{(l)}(t) - y_i^{(l)}(t)\right)^2 dt}, \tag{5}$$

gdzie $x^{(l)}$ jest l-tą pochodną x;

- techniki bazujące na heurystycznych lub geometrycznych kryteriach wyodrębniające grupy danych funkcjonalnych.

Trzecią kategorię tworzą **metody wykorzystujące model bazowy** (*model-based clustering*). W grupowaniu danych największym problemem jest gęstość prawdopodobieństwa, która nie jest zdefiniowania dla funkcjonalnych zmiennych losowych. Stąd techniki określają gęstości prawdopodobieństwa dla skończonej liczby parametrów opisujących krzywe. W przeciwieństwie do metod dwuetapowych, w których estymacja parametrów następuje przed grupowaniem, w metodach wykorzystujących model bazowy te dwa etapy przeprowadzane są jednocześnie.

W ramach tej kategorii wyróżnia się metody oparte na:

– ładunkach wyznaczonych przez funkcjonalną metodę głównych składowych (główne składowe nie są skorelowane)
– współczynnikach otrzymanych w skończonym zbiorze funkcji bazowych.

W pierwszej metodzie gęstości prawdopodobieństwa są wyznaczane na podstawie ładunków otrzymanych w wyniku przeprowadzenia FPCA. Zakłada się, że główne składowe posiadają rozkład Gaussa i grupowanie następuje w oparciu o średnią następującego modelu [Jacques, Preda 2014]:

$$f_X^{(q)}(x;\theta)=\sum_{k=1}^{K}\pi_k\prod_{j=1}^{q_k}f_{C_j/Z_k=1}\left(c_{jk}(x);\lambda_{jk}\right), \tag{6}$$

gdzie $\theta=\left(\pi_k,\lambda_{1k},\ldots,\lambda_{q_kk}\right),1\le k\le K$ jest wektorem parametrów modelu, K – liczba wyodrębnionych grup.

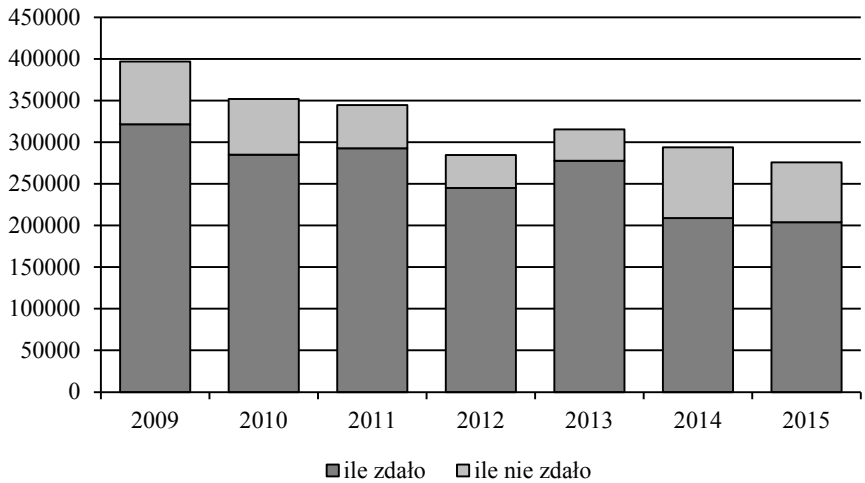
W drugiej metodzie współczynniki funkcji bazowych typu *spline* posiadają rozkład mieszany Gaussa $c_i\sim N(\mu_k,\Sigma)$ dla k-tej grupy. W przeciwieństwie do metod dwuetapowych, w których współczynniki funkcji bazowych są ustalone, tutaj współczynniki są zmiennymi losowymi.

3. Osiągnięcia uczniów na IV etapie edukacji

W Polsce zachodzą bardzo istotne zjawiska demograficzne – maleje liczba osób w wieku szkolnym. W dużej mierze wpływa to na szkolnictwo wyższe, bezpośrednim efektem zmian demograficznych jest spadek liczby studentów. Ważnym aspektem jest także analiza poziomu wiedzy przyszłych studentów. Badaniem objęte będą średnie oceny z poszczególnych przedmiotów otrzymane na egzaminie maturalnym w latach 2009–2015.

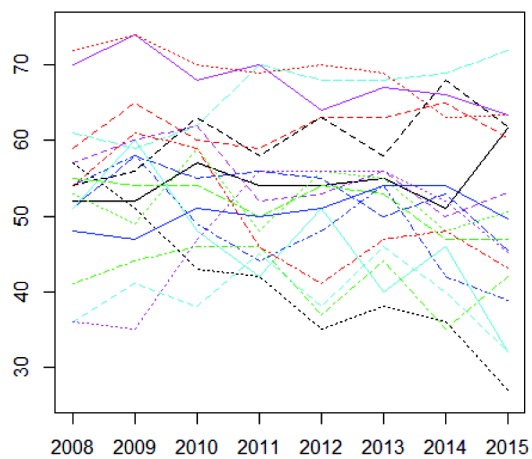
Na rysunku 2 przedstawiono liczbę osób, które zdały maturę lub nie zdały matury w latach 2009–2015. Można zauważyć, że liczba podchodzących do egzaminu maturalnego z roku na rok maleje, a ponadto liczba osób, które nie zdały matury, jest w dwóch ostatnich latach niepokojąco wysoka.

W badaniu przeanalizowano procentowe wyniki maturalne z poszczególnych przedmiotów w latach 2008–2015. Na rysunku 3 przedstawiono obserwacje dyskretne, natomiast na rys. 4 – dane funkcjonalne otrzymane metodą B-spline. Pogrubioną



Rys. 2. Liczba zdających maturę

Źródło: opracowanie własne na podstawie danych z BDL.

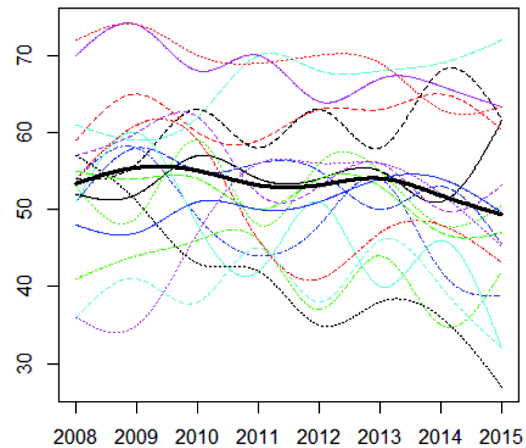


Rys. 3. Średnie wyniki matur (dane oryginalne)

Źródło: opracowanie własne z wykorzystaniem programu R.

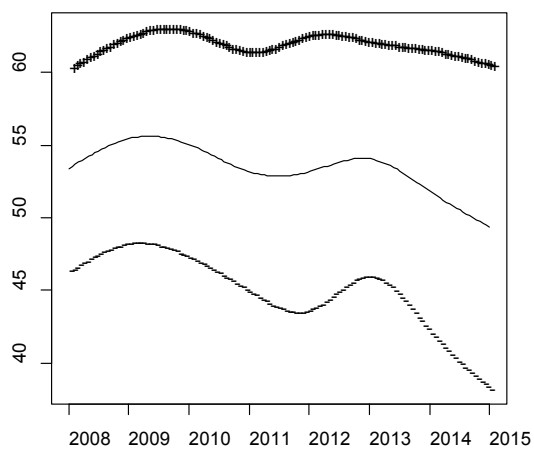
linią zaznaczono średni wynik z wszystkich przedmiotów – można zauważyć tendencję malejącą.

Zastosowano podejście wykorzystujące model bazowy. Za pomocą funkcjonalnej analizy głównych składowych wyodrębniono dwie funkcje składowe. Praktyczne wyjaśnienie funkcjonalnych głównych składowych ułatwiają wykresy odchylenia każdej ze składowych od średniej z wszystkich przedmiotów (rys. 5, 6).



Rys. 4. Średnie wyniki matur (dane funkcjonalne)

Źródło: opracowanie własne z wykorzystaniem programu R.

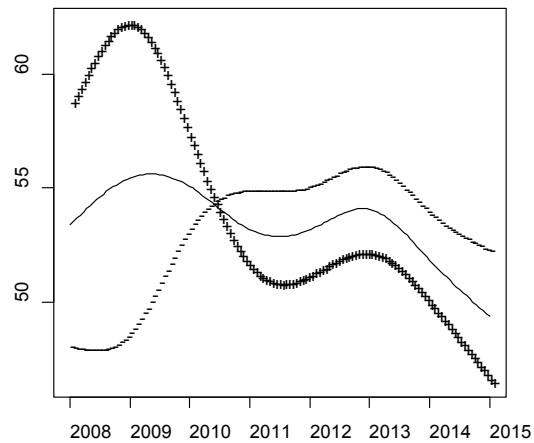


Rys. 5. Składowe FPCA

Źródło: opracowanie własne z wykorzystaniem programu R.

Linia ciągła przedstawia średni wynik maturalny, linia zbudowana ze znaków „+" oznacza średnią powiększoną o tę część składowej, która wyjaśnia zmienność wspólną. Linia zbudowana ze znaków „–" oznacza średnią pomniejszoną o tę wartość.

Pierwsza funkcjonalna główna składowa wyjaśnia 80% zmienności wspólnej, natomiast druga 15%. Pierwsza składowa odpowiada za ogólną tendencję. Dodatni ładunek na tej składowej oznacza, że krzywa opisująca wynik z danego przedmiotu leży powyżej średniej. Druga składowa pokazuje tendencję w pierwszych i ostatnich

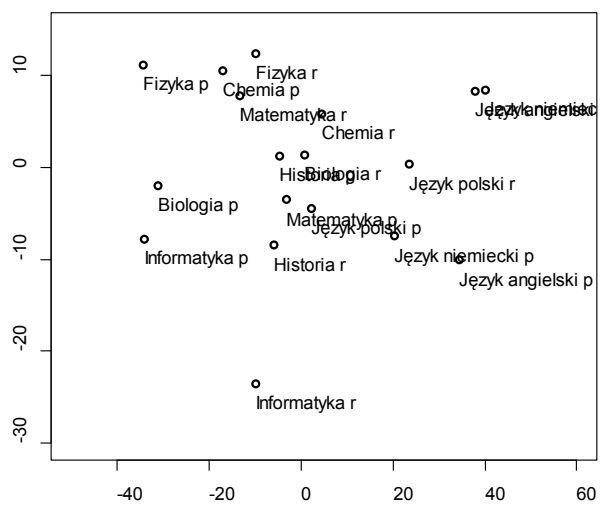


Rys. 6. Składowe FPCA

Źródło: opracowanie własne z wykorzystaniem programu R.

latach w odniesieniu do średniej („początek kontra koniec") porównuje okres do 2010 r. i po 2010 r. ze średnim wynikiem. Dodatni ładunek na drugiej składowej oznacza, że wynik matury z danego przedmiotu na początku badanego okresu był lepszy od średniej, natomiast na końcu wynik był gorszy od średniej.

Na podstawie wyników funkcjonalnej analizy czynnikowej można dokonać wizualizacji danych oraz porównania badanych obiektów. Rysunek 7 zawiera rzut danych na płaszczyznę wyznaczoną przez dwie funkcjonalne główne składowe.



Rys. 7. Obiekty w przestrzeni składowych

Źródło: opracowanie własne z wykorzystaniem programu R.

Najlepsze wyniki maturzyści osiągali z języka niemieckiego i angielskiego na poziomie podstawowym (dodatni ładunek na pierwszej i ujemny na drugiej składowej). Najgorsze wyniki z fizyki podstawowej, chemii podstawowej i matematyki rozszerzonej (ujemny ładunek na pierwszej i dodatni na drugiej składowej).

Bazując na wynikach funkcjonalnej analizy głównych składowych wyodrębniono grupy przedmiotów:

- **grupa 1:** język polski r, język angielski p, język angielski r, język niemiecki p, język niemiecki r, chemia r;
- **grupa 2:** język polski p, matematyka p, matematyka r, biologia p, biologia r, chemia p, fizyka p, fizyka r, historia p, historia r, informatyka p i r.

Najlepsze wyniki maturzyści osiągali z języków oraz z chemii rozszerzonej, ze wszystkich pozostałych przedmiotów wyniki były gorsze.

4. Podsumowanie

Opisane podejścia do grupowania danych funkcjonalnych mają zalety i wady. Metody dla danych dyskretnych w strukturze danych nie uwzględniają zależności od czasu. W metodach dwuetapowych przeprowadza się pierwszy etap, nie uwzględniając głównego celu – grupowania. Metody nieparametryczne są najłatwiejsze w zastosowaniu, jednak nieefektywnie wyodrębniają bardziej złożone grupy. Metody wykorzystujące model bazowy uwzględniają funkcyjną naturę danych, jednocześnie przeprowadzają redukcję i grupowanie.

Wnioski empiryczne wynikające z przeprowadzonego badania: maturzyści mają największe problemy z przedmiotami ścisłymi, zwłaszcza z fizyką na poziomie rozszerzonym i podstawowym, chemią p. podstawowy oraz matematyką p. rozszerzony. Najlepsze wyniki osiągnęli z języka niemieckiego i angielskiego na poziomie podstawowym. Wyniki matur z roku na rok są gorsze zwłaszcza z przedmiotów ścisłych.

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 (7.05.2016).

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:] M. Vichi, P. Monari, S. Mignani, A. Montanari (red.), *New Developments in Classification and Data Analysis*, Springer-Verlag, Berlin, s. 351–358.

Jacques J., Preda C., 2014, *Functional data clustering: A survey*, Advances in Data Analysis and Classification, Springer Verlag (Germany), 8 (3).

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

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