sciendo

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2019, Vol. 23, No. 4
ISSN 1507-3866; e-ISSN 2449-9994

FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS ON THE EXEMPLE OF THE ACHIEVEMENTS OF STUDENTS IN THE YEARS 2009-2017

Mirosława Sztemberg-Lewandowska
Wroclaw University of Economics and Business, Wroclaw, Poland
e-mail: miroslawa.sztemberg-lewandowska@ue.wroc.pl

ORCID: 0000-0001-5915-8388

DOI: 10.15611/eada.2019.4.02
JEL Classification: C38

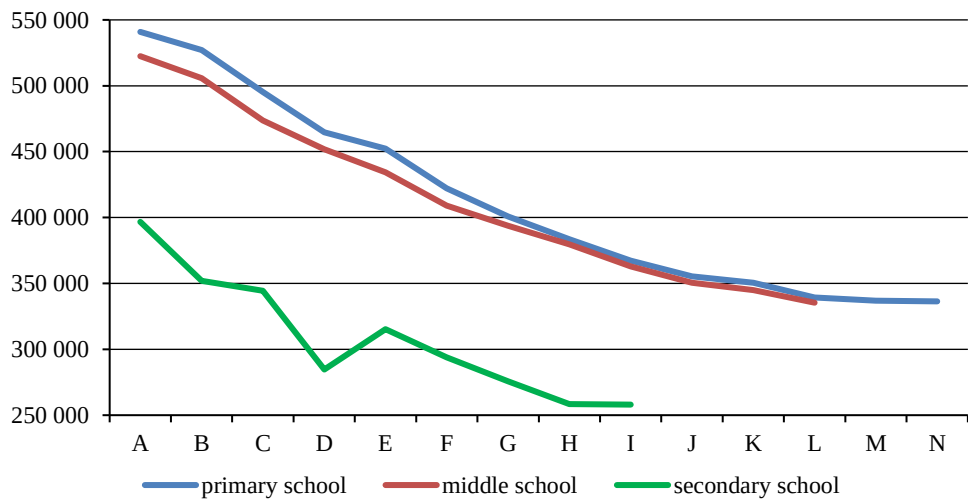
Abstract: The functional principal components analysis joins the advantages of the principal components analysis and provide analysis of dynamic data. The main difference in both methods is the type of data the PCA is based on multivariate data, whereas the FPCA on the functional data including curves and trajectories, i.e. a series of individual observations, not a single observation, as usual. The functional principal components analysis with functional data, will be used in the analysis. This method allows the analysis of dynamic data. The purpose of the article is to apply of functional principal components analysis to the problem of student's achievements. The article was compared the level of students' knowledge during different stages of education in 2009-2017. The analysis covers the average exam results after the II, III and IV stage of education.

Keywords: level of students' knowledge, functional data, longitudinal data, functional principal components analysis.

1. Introduction

The demographic situation has a direct impact on higher education. Since 1990, Polish higher education has been undergoing the period of continuous and dynamic growth, the basis of which was the demographic boom lasting almost fifteen years. The first symptoms of this demographic trend's downturn were observed in the declining number of students since 2006. In each subsequent year a drop in the number of 19-year-olds was recorded, and thus a reversing tendency in the demographic trend, so far beneficial for Polish higher education. This process has been going on till now, causing a gradual decline in the number of people in the traditional higher education student age.

Figure 1 shows the number of students graduating from the subsequent stages of education. On *x*-axis the letter A refers to the students who in 2004 completed primary school (SP), in 2007 – middle school (G) and in 2010 – secondary school (M), thus theoretically people born in the same year. It is noticeable that small local extremes in the number of SP students correspond to greater fluctuations in the number of secondary school students. The number of students graduating from SP in 2006 and secondary school in 2012 reached the minimum, whereas a year later – the local maximum. The number of SP graduates has been stabilizing since 2014, i.e. these students who completed middle school in 2017 and will graduate from secondary school in 2020. The number of secondary school graduates has been stabilizing since 2016.



A	SP 2003	G 2006	M 2009
B	SP 2004	G 2007	M 2010
C	SP 2005	G 2008	M 2011
D	SP 2006	G 2009	M 2012
E	SP 2007	G 2010	M 2013
F	SP 2008	G 2011	M 2014
G	SP 2009	G 2012	M 2015
H	SP 2010	G 2013	M 2016
I	SP 2011	G 2014	M 2017
J	SP 2012	G 2015	
K	SP 2013	G 2016	
L	SP 2014		
M	SP 2015		

Fig. 1. Number of students graduating from subsequent stages of education

Source: the author's compilation based on the Central Examining Board data.

Another important aspect affecting higher education is the level of knowledge presented by the future students. For this purpose, the level of knowledge acquired by students at subsequent stages of education in the period 2009-2017 was compared. The conducted analysis covered the average grades obtained at the exams after completing the II, III and IV stage of education. The method of functional principal component analysis was used for the purposes of the analysis.

The analysis is an extension of the research from 2017 [Sztemberg--Lewandowska 2017].

2. Method

The functional principal components analysis consists in the transformation of primary variables into a set of new mutually orthogonal variables referred to as the principal components [Harman 1975]. The functional principal components analysis has the advantages of a classical analysis of principal components and additionally allows analysing dynamic data. The main difference between these two methods lies in the type of data used: PCA (principal components analysis) is based on multidimensional data, whereas FPCA (functional principal components analysis) on functional data. Functional data take the form of curves and trajectories, i.e. a series of individual observations, rather than a single observation as usual [Hall, Hosseini-Nasab 2006; Górecki, Krzyśko 2012].

Before initiating the procedure of principal components analysis, empirical data should be converted into functional data.

Let variable $\boldsymbol{y_i} = \left(y_i(t_1), y_i(t_2), \cdots, y_i(t_p)\right)$ represent a sample measurement of Y variable over time $t_1, t_2, \cdots, t_p$ for *i*-th unit $(i = 1,2, \ldots, n)$. y*i* data are referred to as raw functional data. Discrete realisations are transformed into a continuous function $x_i(t)$. $\boldsymbol{X_t} = \left(x_1(t), x_2(t), \cdots, x_n(t)\right)$ set is called a functional dataset [Daniele 2006; Hall, Müller, Wang 2006].

The functional principal components analysis consists in finding the principal components explaining the most common variability of all variables. The problem is to determine the $x_i(t)$ function for which determining the principal components is possible. The $x_i(t)$ function is presented as the linear combination of the basic functions:

$$x_i(t) = \sum_j c_{ij}\varphi_j(t), \tag{1}$$

where: c_{ij} – coefficients of linear combination,

$\varphi_j(t)$ – functions forming the orthonormal basis for $L_2(I)$ space, and $x_i(t) \in L_2(I)$, where $L_2(I)$ – Hilbert space of integrable functions with a square on interval I equipped with a scalar product

$$\langle u, v\rangle = \int_I u(t)v(t)dt.$$

The following basic functions are most often used:

- monomials 1, t, t_2, t_3, . . . , t_k, ...
- Fourier functions (for cyclic data) 1, sin(ωt), cos(ωt), sin(2ωt), cos(2ωt), sin(3ωt), cos(3ωt), . . . , sin(kωt), cos(kωt), . . .
- B-spline functions, which have the following properties:
  - each basic function is a "spline" of m rank functions in the points called nodes
  - the sum, difference and linear combination of these basis functions is still a B-spline function.

B-spline functions were used as basic functions in the article.

The integral of the squared error is adopted as the matching criterion for each curve:

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

where $x_i \; and \; \hat{x}_i$ represent the observed and matched curves. The global measure of approximation is given by the formula:

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

Functional data should be smoothed as any function roughness is treated as noise, which should be removed as far as possible. The measure of the function roughness is its second rank derivative and smoothed functions should take small values of this derivative [Silverman 1996].

The functional data prepared in this way are used in the analysis of the principal components. In the functional PCA, each principal component is expressed by the principal component weight function, also referred to as eigenfunction $\xi_j(t)$ depending on time [Daniele 2006; Hall, Hosseini-Nasab 2006]. The eigenfunction maximizes the variance of principal component functions:

$$v(t,s) \stackrel{\text{def}}{=} \frac{1}{n-1} \sum_{i=1}^{n} \{x_i(t) - \bar{x}(t)\} \{x_i(s) - \bar{x}(s)\}. \tag{4}$$

Similarly to the classic PCA, the problem of the functional one is the distribution of the function variance:

$$v(t,s) = \sum_j \lambda_j \xi_j(t) \, \xi_j(s), \tag{5}$$

where $\lambda_j, \xi_j(t)$ meet the eigenequation:

$$\langle v(u), \xi_j \rangle = \lambda_j \xi_j(s) \tag{6}$$

and eigenvalues are positive and non-decreasing:

$$\lambda_j \stackrel{\text{def}}{=} \int_T \; \xi_j(t) v(t,s) \xi_j(s) dt ds. \tag{7}$$

Eigenfunctions satisfy the following condition:

$$\int_T \; \xi_j^2(t) dt = 1 \;\; \text{and} \;\; \int_T \; \xi_j \; (t) \xi_i(t) dt = 0 \; (i < j). \tag{8}$$

Eigenfunctions define the principal components of variations between x_i sample functions [Ingrassia, Costanzo 2005; Hall et al. 2006; Górecki, Krzyśko 2012a].

The purpose of the article is to apply functional principal components analysis to the problem of student's achievements.

3. II stage of education

Since April 2002 a general exam has been conducted for students who finish the sixth grade of primary school. The standards of examination requirements constitute the basis for conducting such an exam. Up to 2014 the standards were grouped into five cross-curricular categories:

- reading,
- writing,
- reasoning,
- using information,
- applying knowledge in practice.

In 2015 the exam consisted of two parts:

- first part – checking knowledge in Polish and maths,
- second part – checking knowledge in a modern foreign language.

A sixth grader takes the exam in one of the following foreign languages: English, Spanish, French, Russian, German or Italian. The student can choose only the language he/she learns at school as a compulsory subject.

Figure 2 shows the average percentage results of the exam in individual subjects at the end of sixth grade in the years 2002-2014.

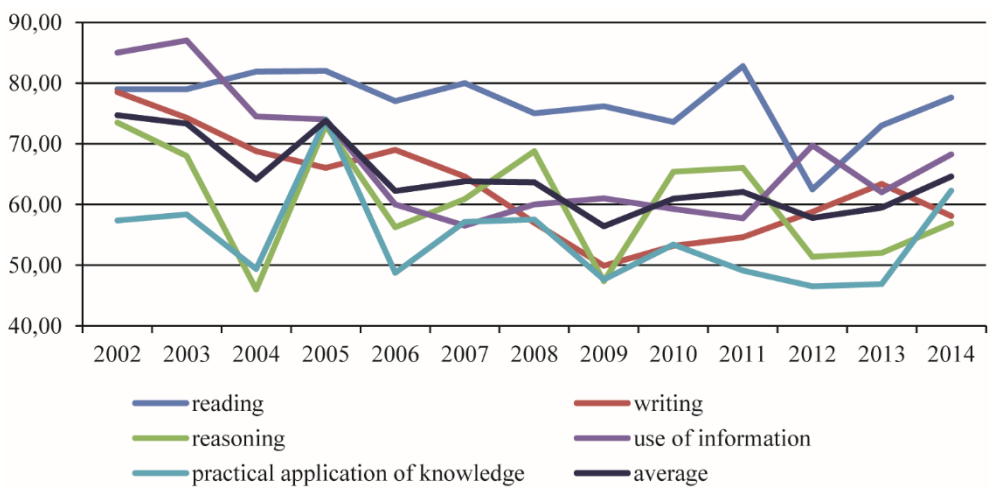


Fig. 2. Average exam results at the end of sixth grade of primary school in individual subjects [%]

Source: author's compilation based on the Central Examining Board data.

The average result in all subjects is marked in black, and shows a decreasing tendency in the period 2002-2014 [Sztemberg-Lewandowska 2017].

Two weight functions were distinguished using the functional principal component analysis. A practical explanation of the functional principal components is facilitated by the graphs of each weight function deviation from the mean value in all subjects (Figure 3).

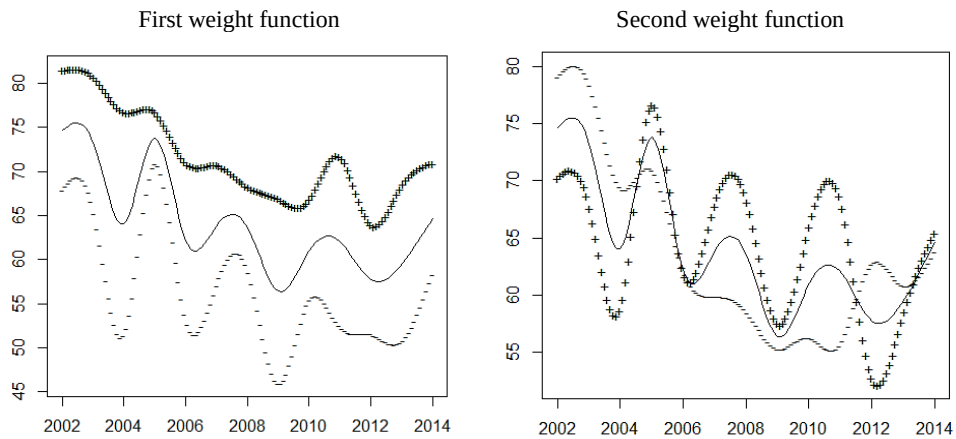


Fig. 3. Weight function

Source: the author's compilation using R software (fda package).

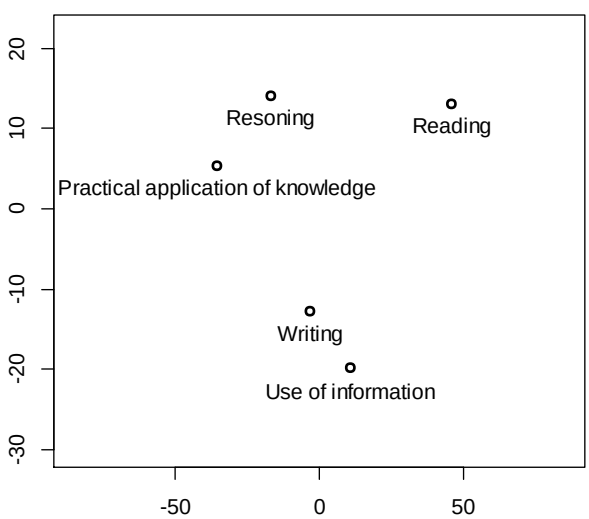


Fig. 4. Objects in the space of functional principal components

Source: the author's compilation using R software (fda package).

The first functional principal component explains 72% of the common variability, whereas the second one – 18%. The first component refers to the general trend. A positive charge on this component means that the curve describing the result in a given subject is above the average. The second component shows the tendency in borderline and middle years in relation to the average. A positive charge of the second component means that the exam result in a given subject before 2005 and after 2011 was worse than the average, while in 2005-2011 the result was better than the average [Sztemberg-Lewandowska 2017].

Based on the results of the functional principal components analysis it is possible to visualize the data and compare the analysed results from individual subjects. Figure 4 presents the projection of data on a plane defined by two functional principal components.

Students received the best results in reading – above average. Writing and using information was at average level, the situation was better before 2005 and after 2011. Applying knowledge in practice and reasoning were below average.

4. III stage of education

The exam in the III grade of middle school checks the knowledge and skills specified in the core curriculum of general education in relation to selected subjects taught at the III and previous stages of education.

Before 2009 middle school students took only the maths part, and humanities and natural science part. In the period 2009-2011 the middle school graduation exam consisted of three following parts:

- humanities
- maths and natural science
- modern foreign language.

Since 2012 separate scopes were distinguished within each part:

- humanities in the scope of history and social studies and in the scope of Polish language
- maths and natural science in the scope of life sciences and in the scope of maths
- modern foreign language at basic and extended level.

A middle school student takes an exam in one of the following foreign languages: English, German, Spanish, French, Russian, Italian or Ukrainian. A student can choose only the language he/she learns at school as a compulsory subject.

Every middle school learner is committed to take a modern foreign language exam at basic level. An extended level exam is compulsory only for students who, at the exam, chose the language they were learning in the primary school as well. Other middle school students can also take it if they want to check their language skills.

The exam has a written form. Taking it is the condition for the middle school graduation, but the minimum score which the exam taker has to achieve is not defined, therefore no-one fails the exam.

Figure 5 presents the average percentage results of the exam in individual subjects after the third class of middle school in 2006-2017. The average score in the period 2006-2017, in all subjects is marked in black.

Two weight functions were distinguished using the functional principal components analysis, the graphs of these functions' deviations from the average score in all subjects are presented in Figure 6.

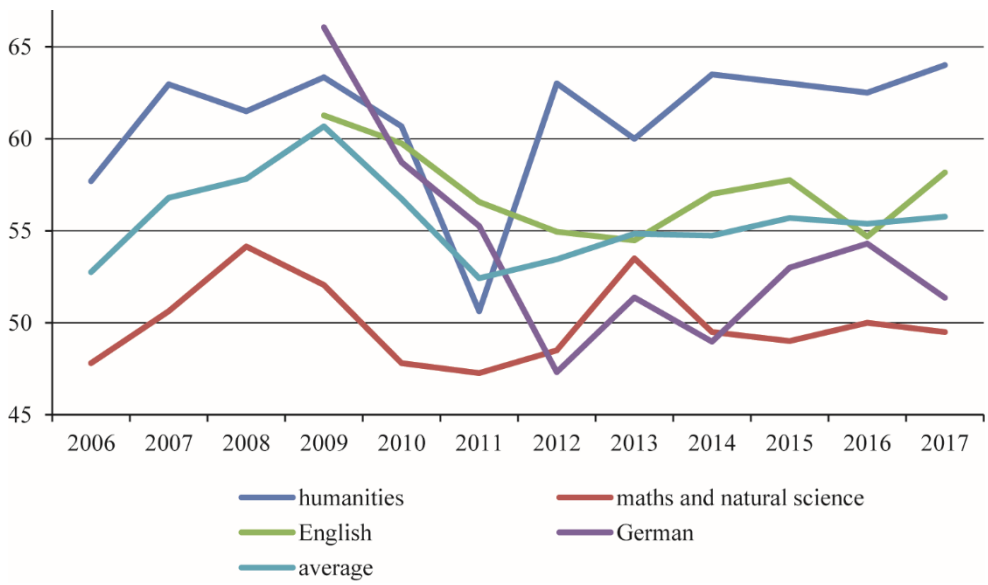


Fig. 5. Average exam results at the end of third grade of middle school in individual subjects [%]

Source: the author's compilation based on the Central Examining Board data.

The first functional principal component explains 74.5% of the common variability and the second one – 18.0%. The first component refers to the general trend. A negative charge on this component means that the curve describing the result in a given subject is below the average. The second component shows the tendency in the first and middle years in relation to the average: it compares the period till 2012 and 2012-2015 against the average score. A negative charge of the second component means that the exam result in a given subject at the beginning of the analysed period was lower than the average. in 2012-2015 the result was higher than the average.

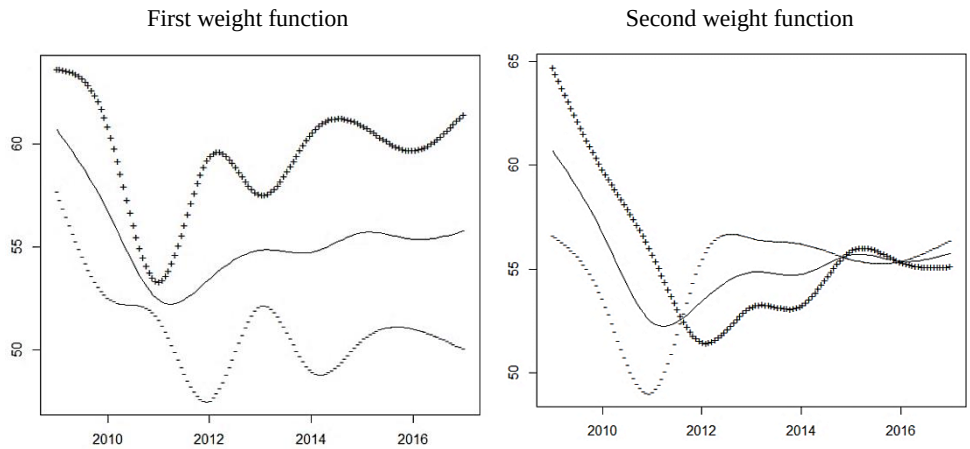


Fig. 6. FPCA weight functions

Source: the author's compilation using R software (fda package).

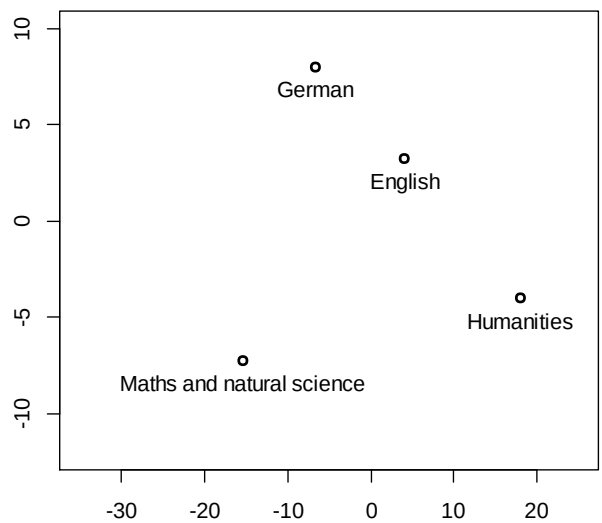


Fig. 7. Objects in the space of components

Source: the author's compilation using R software (fda package).

The visualization of objects in the space of component functions (Figure 7) allows drawing the following conclusions:

- humanities: the achieved exam result was higher than the average, in 2006--2012 the situation was worse than in 2012-2015,

- English and German: the achieved exam result was at average level, at the beginning of the analysed period it was above average and after 2012 was lower than average,
- maths and natural science part: the achieved test result was below average, in 2006-2012 the situation was worse than in 2012-2017.

5. IV stage of education

A graduate called the ‘old matura’ (the old type of the secondary school graduation exam taken before 2015) is obliged to take two exams in the oral part and three exams in the written part. The compulsory oral part comprises:

- exam in Polish language (without specifying the level),
- exam in a modern foreign language (without specifying the level).

The compulsory written part exams comprises:

- exam in Polish language (at basic level),
- exam in maths (at extended level),
- exam in a modern foreign language (at basic level).

A graduate called the ‘new matura’ (since 2015) in its written part also takes the exam in the chosen additional subject (at extended level).

In order to receive the graduation diploma one must obtain at least 30% points from an exam in each compulsory subject in the oral part and at least 30% points from an exam in each compulsory subject in the written part.

.

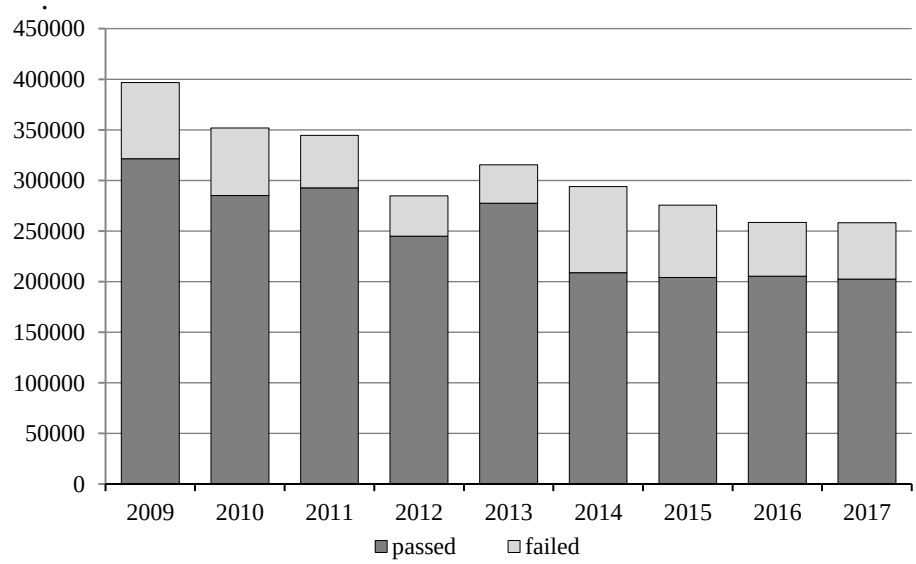


Fig. 8. Number of people taking the ‘matura’ exam

Source: the author’s compilation based on the Central Examining Board data.

Figure 8 shows the number of people who passed and failed ‘matura’ in the period 2009-2017. It can be noted that the number of students taking the secondary school graduation exam shows a declining tendency, and moreover the number of people who did not pass ‘matura’ in recent years is alarmingly high

Figure 9 presents the average percentage results obtained in ‘matura’ exam in individual subjects in the period 2008-2017. The average score in all subjects is marked in bold – the declining trend is visible. The situation has been stabilizing since 2016

Two weight functions were distinguished using the functional principal components analysis. The graphs showing these functions deviations from the average score in all subjects are presented in Figure 10.

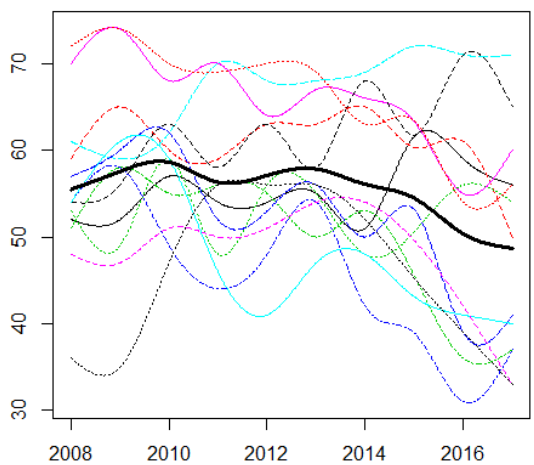


Fig. 9. Average results in the ‘matura’ exams (functional data)

Source: the author’s compilation using R software (fda package).

The first functional principal component explains 72.3% of the common variability and the second one – 17.1%. The first component refers to the general trend. A negative charge on this component means that the curve describing the result in a given subject is below the average. The second component shows the tendency in the first and last years in relation to the average (beginning versus end) and compares the period 2008-2012 and 2013-2017 against the average score. A negative charge of the second component means that ‘matura’ result in a given subject at the beginning of the analysed period was lower than the average, whereas in the end the result was better than the average.

Next the projection of data was made on the plane determined by two functional principal components (Figure 11).

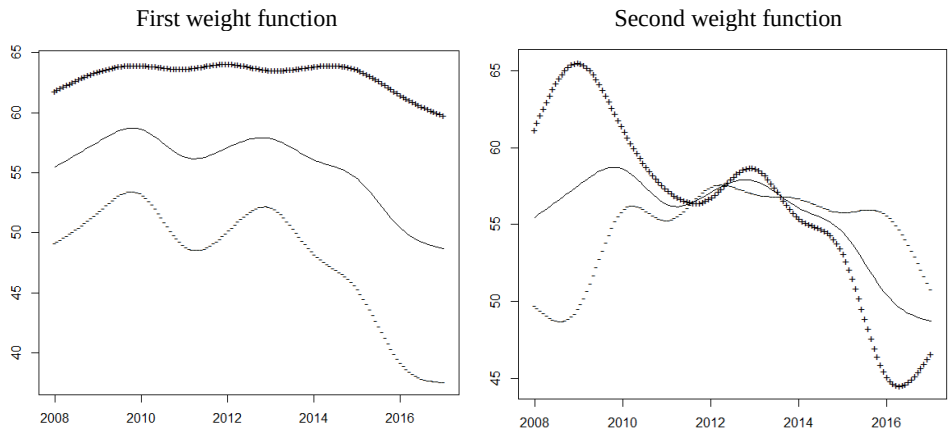


Fig. 10. FPCA weight functions

Source: the author's compilation using R software (fda package).

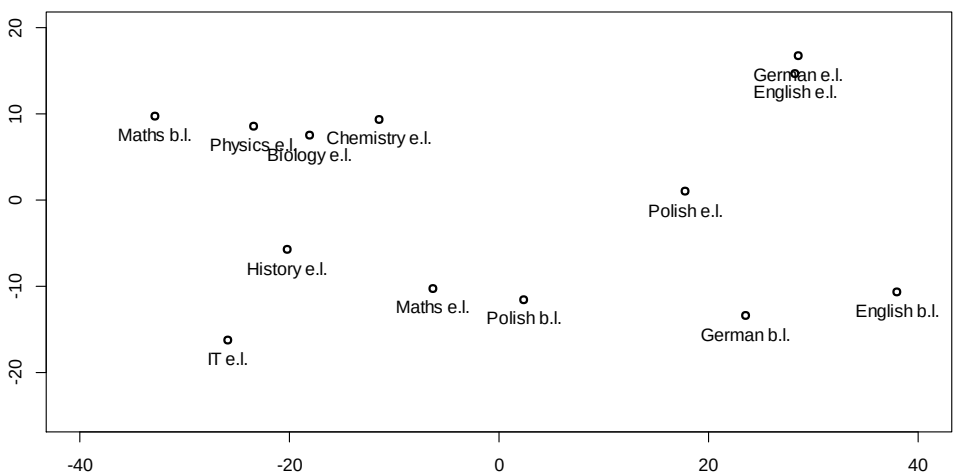


Fig. 11. Objects in the space of components

Source: the author's compilation using R software (fda package).

The best results were obtained by the secondary school graduates in German and English at basic level (a positive charge on the first and a negative one on the second component). The results in German and English at extended level were higher than the average, however worse in recent years than at the beginning of the analysed period. The worst results were recorded in maths, physics, biology and chemistry at extended level a (negative charge on the first and a positive one on the second component).

6. Conclusion

The number of sixth graders in primary schools has been stable since 2014, whereas the number of students in the last year of secondary school has presented a stable level since 2016.

At all levels of education a declining trend is noticeable in terms of the average exam results. Starting from primary school, students were obtaining better results in humanities, while maths and natural science were causing more problems. Primary school leavers obtained an above average score in reading, however the application of knowledge in practice turned out to be below the average level in all subjects. In the case of the middle school exam, the humanities part was passed with an above average result, while the results of maths and natural science part were below the average in all the subjects. 'Matura' graduates have the biggest problems in science, predominantly in maths, physics, biology, chemistry (extended level). They achieved their best scores in German and English at basic level.

The current study confirmed the conclusions of the 2017 study on science subjects. Students have more problems with science than with humanities. In recent years there has been a stabilization in the number of primary school students and especially in high school. The situation in the level of knowledge of students in stages third and fourth has improved in the last two years.

The presented analysis shows the usefulness of the functional principal components analysis in the visualization and inference of economic time series from a slightly different perspective than has been done so far. The classical methods, although simpler, do not show the full picture of the analysed phenomenon. They disregard the fact that data take the form of a time series and for this reason the sequence of variables is important (e.g. classical principal components analysis). FPCA, additionally, facilitates interesting visualizations, which classical methods do not allow.

Bibliography

Daniele M., 2006, *Functional Principal Components Analysis to Study Environmental Data*, http://old.sis-statistica.org/files/pdf/atti/Spontanee%202006_677-680.pdf.

Górecki T., Krzyśko M., 2012, *A kernel version of functional principal components analysis*, Statistics in Transition, 13(3), pp. 559-568.

Górecki T., Krzyśko M., 2012a, *Functional Principal Components Analysis*, [in:] *Data Analysis Methods and its Applications* (eds. J. Pociecha, R. Decker), C.H. Beck, pp.71-87.

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, pp. 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, pp. 1493-1517.

Harman H., 1975, *Modern Factor Analysis*, The University of Chicago Press.
http://www.sis-statistica.it/files/pdf/atti/Spontanee%202006_677-680.pdf.
Ingrassia S., Costanzo G.D., 2005, *Functional Principal Component Analysis of Financial Time Series*, [in:] M. Vichi, P. Monari, S. Mignani, A. Montanari (eds.), *New Developments in Classification and Data Analysis*, Springer-Verlag, Berlin, pp. 351-358.
Ramsay J.O., Hooker G., Graves S., 2009, *Functional Data Analysis with R and MATLAB*, Springer.
Ramsay J.O., Silverman B.W., 2005, *Functional Data Analysis*, Springer.
Silverman B.W., 1996, *Smoothed functional principal components analysis by choice of norm*, Ann. Statist., 24, pp. 1-24.
Sztemberg-Lewandowska M., 2017, *The achievements of students at the II-IV stages of education using functional principal component analysis*, Statistics in Transition New Series, March 2017, vol. 18, no. 1, pp. 1-12.

FUNKCJONALNA ANALIZA GŁÓWNYCH SKŁADOWYCH W BADANIU OSIĄGNIĘĆ UCZNIÓW W LATACH 2009-2017

Streszczenie: Analiza funkcjonalna wykorzystuje dane funkcjonalne, tzn. krzywe i trajektorie, czyli ciągi indywidualnych obserwacji, nie na pojedynczej obserwacji. Funkcjonalna analiza głównych składowych polega na przekształceniu funkcjonalnych zmiennych pierwotnych w zbiór nowych wzajemnie ortogonalnych zmiennych, nazywanych głównymi składowymi. Zastosowanie metody dla danych funkcjonalnych umożliwia analizę danych o charakterze dynamicznym. Celem artykułu jest wykorzystanie funkcjonalnej analizy głównych składowych do porównania poziomu wiedzy uczniów na kolejnych etapach edukacji w latach 2009-2017. Badaniem objęto średnie oceny otrzymane na egzaminach po zakończeniu II, III i IV etapu edukacji. W analizie wykorzystano funkcjonalną analizę głównych składowych, bazującą na danych funkcjonalnych. Metoda ta umożliwia analizę danych o charakterze dynamicznym.

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