PRINCIPAL COMPONENT ANALYSIS IN THE STUDY OF STRUCTURE OF THE BEST POLISH DECATHLON COMPETITORS FROM THE PERIOD BETWEEN 1985–2015

Bartosz Dziadek,[1, A, B, C, D] Janusz Iskra,[2, A, B] Krzysztof Przednowek[1, A, C, D]

[1] University of Rzeszow, Faculty of Physical Education, Poland
[2] Opole University of Technology, Faculty of Physical Education and Physiotherapy, Poland
[A] Study Design; [B] Data Collection; [C] Statistical Analysis; [D] Manuscript Preparation

Address for correspondence:
Bartosz Dziadek
Faculty of Physical Education, University of Rzeszow
Towarnickiego 3 street, 35-959, Rzeszow, Poland
E-mail: bdziadek@ur.edu.pl

Abstract The modern decathlon is a sport consisting of ten different events held over two days, played by men. Depending on the complexity of combined events, variety of events (runs, throws, jumps), the multi-stage, time-consuming and difficult training process the sport is considered as one of the most difficult. The analysis of careers of the best decathlon participants and applying advanced data-mining methods can help define the patterns occurring between each decathlon event and the final result.

The research material encompasses career data of the 25 top competitors from Poland in years 1985–2015. Principal component analysis (PCA) was used in the research in order to designate new uncorrelated variables (components), representing input data across a new plane.

Data analysis involved appointment of correlations between the events, determining the number of main components taken into account in further studies, analysis of the weight of each variable in formation of main components as well as visualisation and interpretation of results in the new plane described by the determined main components.

Through the implementation of PCA method in the process of analysis it was possible to designate over 69% of compound data volatility with the use of the first three components. The first component, comprised of seven variables, displays the largest share in the total variability. The study of the relationship between variables in the new plane displayed strong correlations between sprint events (100 m, 110 m hurdles) and long jump and pole vault. No correlations between the 1,500 m run and other events were found.

Key words decathlon, sport career, principal component analysis

Introduction

The modern decathlon is a combined event, played by men, comprising ten events held over two days. This complex form of track and field competition is characterised by varied composition of events (runs, throws, jumps), differing in terms of effort and technique and requires a competitor to possess a high degree of motor skills (speed, strength, endurance) and perfectly mastered technique (IAAF, 2018; Quercetani, 2000).

Decathlon is one of few athletic competitions, in which the results from each component are recalculated into points and summed, with the final result being the measure of the competitor's proficiency. Because of the number of component events and complexity of that combined event the training process of a decathlon competitor is a lengthy, multi-stage and difficult process (Socha, 1977).

Because of that, scientific research of decathlon's structure turned out to be crucial, as it could assist the competitors and trainers in optimising the training process. Over the course of the years there were numerous works, which employed advanced mathematical and statistical methods in analysis. Furdal (1986) highlighted in his paper the importance of studies carried out as early as in the 1950s and 60s (Karvonen, Niemi, 1953; Zaciorski, Godik, 1963) which utilised factor analysis for checking the correlations between the total result and component events. In the same work the author carried out an analysis on the basis of 158 world's best decathlon competitors using oblique component method and principal profile method, which resulted in determining tree correlated event groups in decathlon and separation of six main competitor profiles in relation to acquired results (Furdal, 1986).

Currently, there are also works, in which advanced calculation methods (e.g. factor analysis, configural frequency analysis) were used to define the type of competitors which are successful in decathlon because of their aptitude and motor skills (Bilić, 2015; Stemmler, Bäumler, 2005), as well as to study the hidden structure of decathlon (Ertel, 2011).

Park, Zatsiorsky (2011) in their work used the principal component analysis (PCA) to designate new variables describing the structure of decathlon on the basis of the results of competitors competing during Olympic Games in years 1988–2008. Another work the principal component analysis was used to reduce the number of variables describing the career progress of decathlon competitors competing in years 2004–2013, the results of which were used to predict the results of future events with the use of artificial neural networks (Chen, Zhang, 2016).

The goal of the work was the implementation of advanced data-mining methods, in this case – the PCA method, to reduce the number of variables describing the structure of decathlon on the basis of career progress of the 25 best Polish competitors, competing in years 1985–2015.

Methods

The study material involved data of career progress of the 25 best Polish competitors from Poland, competing in years 1985–2015. The competitors were selected on the basis of their personal records and length of their sports careers, which could be no shorter than 5 years. The accumulated data (194 participations) involved the best entries of competitors in each year of their careers, containing final results and partial results from each of the ten component events expressed in metric units and in points. The database was created on the basis of the provided PZLA result database (PZLA, 2016) and published PZLA statistical yearbooks for the period between 1985–2015 (PZLA, 1985–2006).

In studies carried out to reduce the number of variables characterising the career progress of the best Polish decathlon participants the principal component analysis (PCA) was used. PCA method involving matrix operations is used for multidimensional data exploration, projection and visualisation. PCA analysis results in new principal components which are linear combinations of vectors subjected to analysis regarding maximisation of variance description. The designated components, which represent multidimensional input date in a new plane contain the most important data regarding volatility of the acquired and analysed study material. Reduction of data dimensionality is carried out by studying the acquired eigenvalues of principal components describing the percentage of described

data variance. The choice of the appropriate number of components included in the following analyses is based on the criterion adopted by the researcher, e.g. Kaiser's criterion, scree test or others. This thesis utilises Kaiser's criterion, which eliminates from further analysis principal components of eigenvalues of less than 1 (Bishop, 2006; Daszykowski, Walczak, 2008; Hardle, Simar, 2007; Kassambara, 2017; STHDA, 2018).

All analyses were carried out with the use of R programming language with additional packages (R Core Team, 2018).

Results

Utilising the accumulated research material involving the career progress of 25 Polish decathlon competitors competing in years 1985–2015 basic statistics were designated, presented in Table 1. The highest result in the group of Polish combined event competitors was achieved by Sebastian Chmara, who scored 8566 pts in Alhama de Murcia in 1998, taking first place and beating Polish record (Matthews, 2013).

Table 1. Basic statistics

Variable	Symbol	Unit	Best results		Worse results		$\bar{x}$		Me	
Personal best	PB	[pt]	8566		7253		7652		7551	
100 M run	X100m	[s] [pt]	10.76	915	12.44	567	11.36	783	11.33	789
Long jump	HJ	[m] [pt]	7.75	997	5.49	479	6.91	794	6.96	804
Shot put	SP	[m] [pt]	16.03	853	9.56	459	13.10	674	13.31	687
High jump	HJ	[m] [pt]	2.15	944	1.56	434	1.95	759	1.95	758
400 M run	X400m	[s] [pt]	48.27	896	57.08	518	51.12	765	50.89	774
110 M hurdles	X110m	[s] [pt]	14.32	934	18.70	459	15.41	803	15.31	813
Discus throw	DT	[m] [pt]	49.86	867	22.11	312	38.91	643	38.98	644
Pole vault	PV	[m] [pt]	5.20	972	2.40	220	4.23	687	4.28	695
Javelin throw	JT	[m] [pt]	67.48	851	28.25	275	51.59	613	51.96	618
1,500 M run	X1500m	[s] [pt]	260.12	811	346.01	324	287.92	634	286.68	639

In accordance to Daszykowski, Walczak (2008) the element being decisive in decreasing the data dimensionality through substitution of some variables by new variables being linear combinations of original parameters are correlations between individual variables. In order to study the correlation between component events of decathlon and the impact of these events on the final result of the competition the values of Pearson's linear correlation were calculated r_{xy}, which are presented in Figure 1.

By analysing the values of correlation coefficient, a crucial and significant impact of nine component events on the final result of decathlon was observed ($r_{xy} > 0.59$). The smallest correlations were found between total result and the 1,500 m run where the value of correlation coefficient was at the level of -0.29. By studying the correlations between the component events of decathlon, the highest values were determined for discus throw and shot put ($r_{xy} = 0.72$) as well as 100 m run and long jump ($r_{xy} = –0.64$) and 400 m run where the r_{xy} was at 0.64. The smallest correlations were observed for the variable of 1,500 m run. The exception was the correlation of that event with 400 m run, which amounted to $r_{xy} = 0.53$.

The result of the PCA analysis was creation of 10 new principal components describing the total variance of original data. On the basis of the acquired eigenvalues of components (Table 2), which describe value of variance of contained data, and the assumed Kaiser's criterion the first three principal components were chosen, which describe 69.77% volatility of acquired data. The graphic interpretation of eigenvalues of principal components is shown in Figure 2.

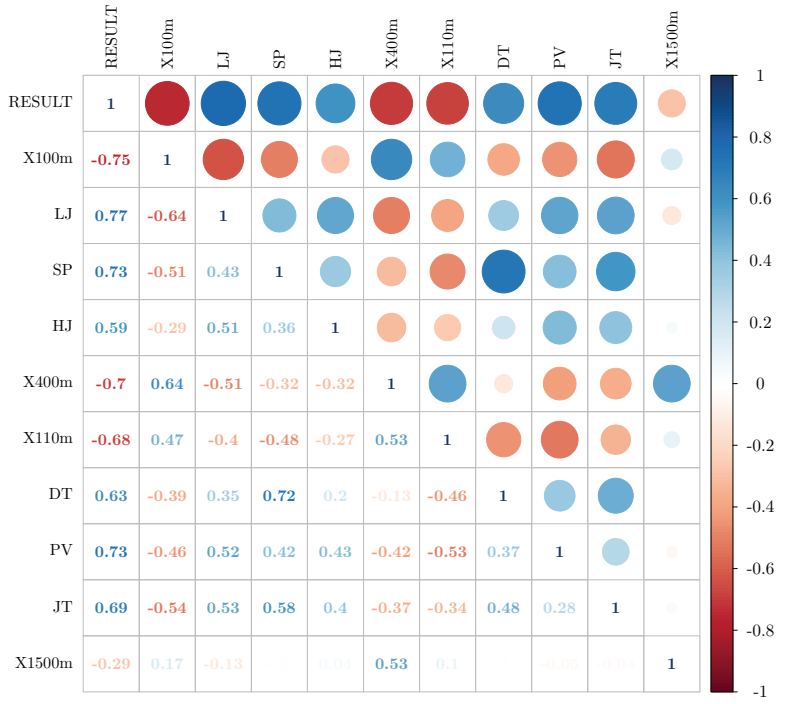

Figure 1. Values of Pearson's linear correlation coefficients for 10 component events

Table 2. Eigenvalues of principal components

	Eigenvalues	Percentage of variances	The cumulative percentage of variances
Component 1	4.53	45.27	45.27
Component 2	1.45	14.49	59.76
Component 3	1.00	10.01	69.77
Component 4	0.83	8.32	78.09
Component 5	0.62	6.17	84.27
Component 6	0.50	4.97	89.23
Component 7	0.34	3.44	92.67
Component 8	0.34	3.41	96.09
Component 9	0.23	2.33	98.42
Component 10	0.16	1.58	100.00

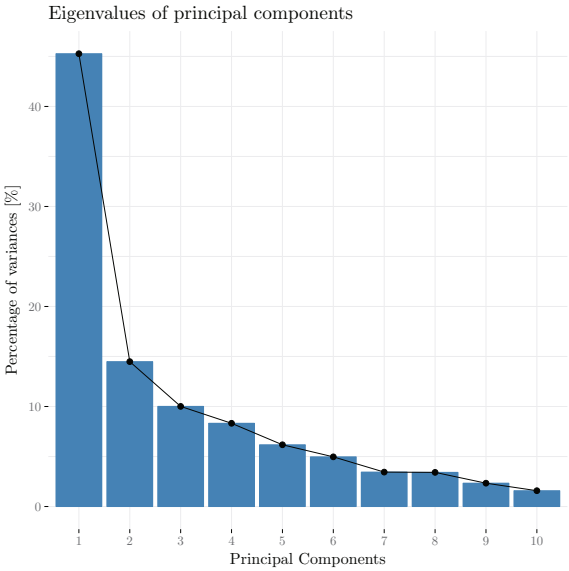

Figure 2. Eigenvalues of principal components – graphic interpretation

Each variable of original data included in PCA analysis has a specific contribution to building new components. Percentages of variables describing the first three principal components were placed in Table 3, and their graphic form was presented in Figures 3–5. By studying the structure of the first component seven explanatory variables were found, among which the highest percentage was reached by variables corresponding to events: 100 m run, long jump and shot put. The remaining variable with values above the red dashed line (Figure 3) have a less severe but important impact on the quality of information stored in the first principal component. The second principal component (Figure 4) is composed of variables storing the result data of 1,500 m run (43.81%), 400 m run (26.45%) and discus throw (14.89%). The structure of the last principal component (Figure 5) contains the variable of high jump, which explains 39.95% of variability and variables of discus throw and 1,500 m run.

Table 3. Share of each variable in each principal component (%)

Variable	Component 1	Component 2	Component 3
X100m	13.85	1.67	0.09
HJ	13.32	0.37	9.81
SP	12.62	8.99	8.88
HJ	7.21	1.05	39.95
X400m	10.64	26.45	0.00
X110m	10.91	0.07	3.21
DT	8.96	14.89	21.74
PV	10.43	0.05	5.32
JT	11.12	2.68	0.13
X1500m	0.95	43.81	10.87

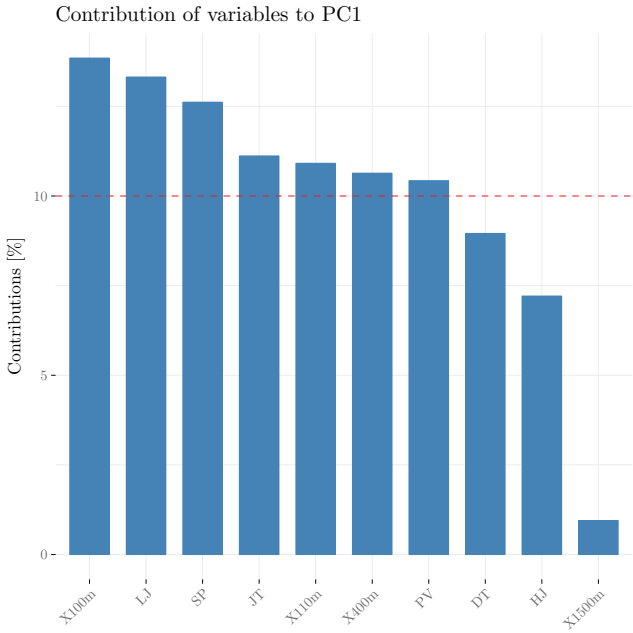

Figure 3. Share of variables in the first principal component

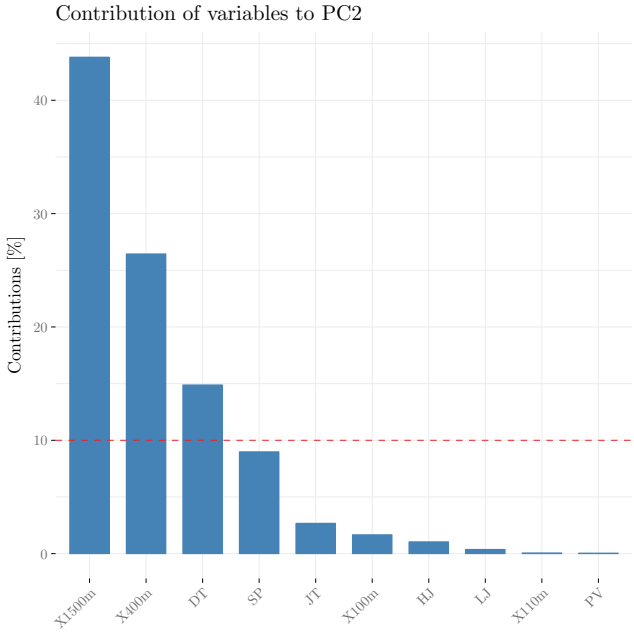

Figure 4. Share of variables in the second principal component

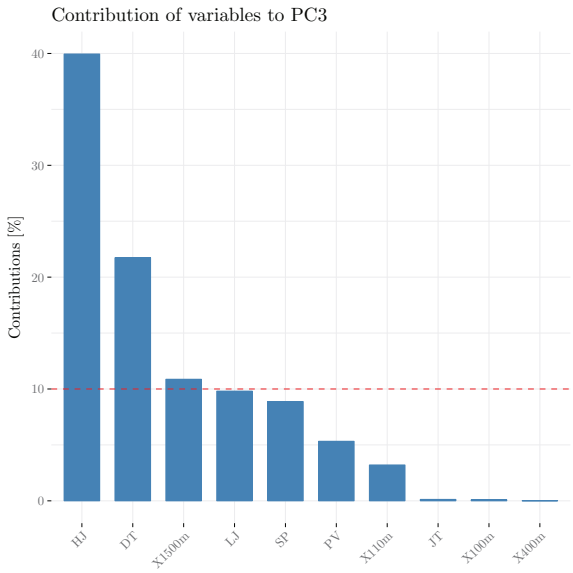

Figure 5. Share of variables in the third principal component

The structural correlation between variables and principal components can also be illustrated on a factor map (Figures 6–8). Factor map is used for graphical presentation of result on a plane described by the created principal components, and shows the correlation strength and coefficient of correlation between variables and the quality of their representation, as well as helps define (description, naming) the principal components. On the basis of the chart creating the plane described by the first two principal components (59.76% of the total variance) and drawn variables (Figure 6) a significant correlation between 100 m and 110 m sprints was observed, which negatively correlate with long jump and pole vault events. Also observed was a positive correlation between throw events. The variable storing data on the results achieved by decathlon competitors in 1,500 m run does not indicate correlation with the rest of the variables, except for an insignificant correlation with 400 m run. In the chart described by the second and third component a large share of 1,500 m run in creation of the second component was observed, which, similarly as before, does not indicate correlations with other events except for 400 m run. By projecting the variables on the ordinate axis formed by the third component a large share of high jump variable and discus throw can be seen in the structure of the third variable which confirms the previous observations (Figure 5). In addition, in the spaces described by the included principal components, significant positive relationships between the variable discus throw and shot put (DT and SP) were observed.

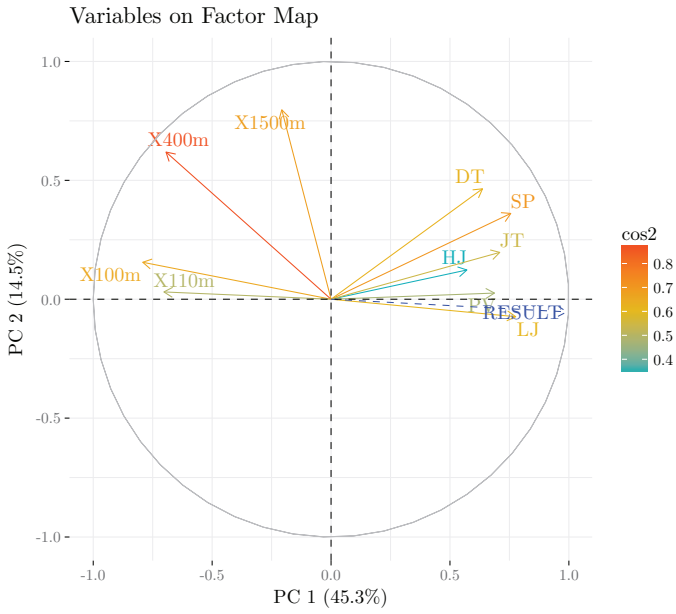

Figure 6. Projection of variables on the factor map described by the first and second principal component

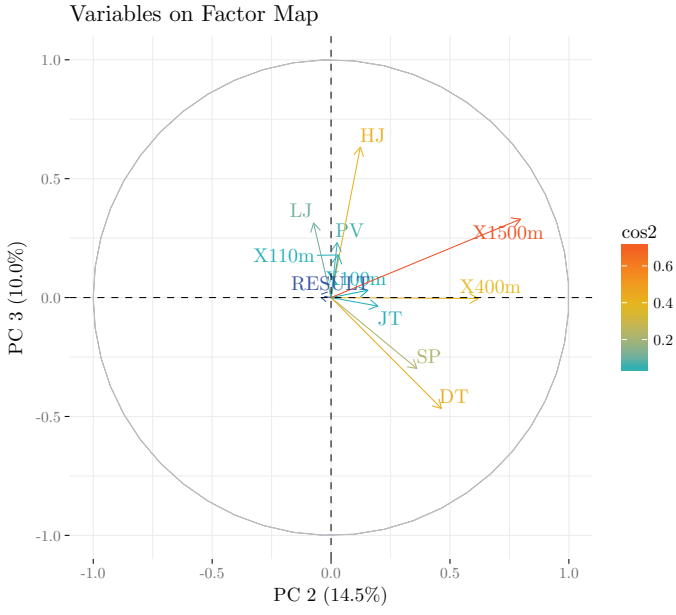

Figure 7. Projection of variables on the factor map described by the second and third principal component

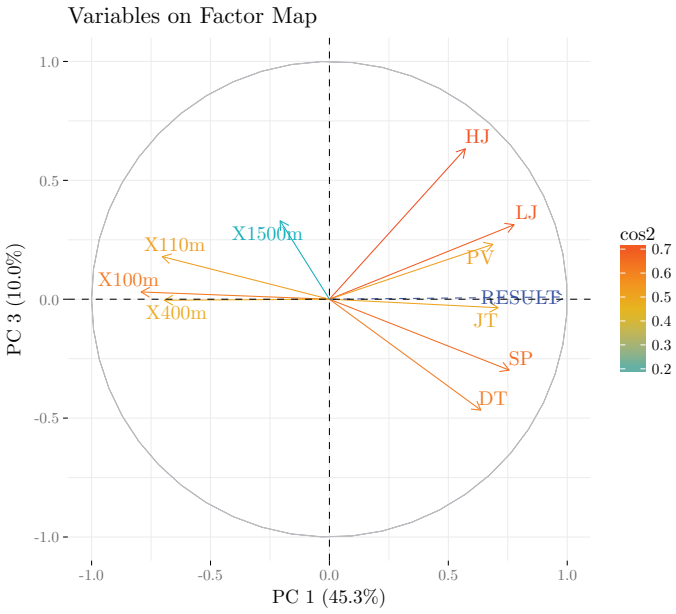

Figure 8. Projection of variables on the factor map described by the first and third principal component

Discussion

Study of the structure of decathlon, searching for correlations between component events and the final result can be practically implemented and significantly improve the training process of a competitor. Mathematical and statistic methods were often used in research involving the complexity of this form of combined event competition. The acquired research material encompassing the career progress of 25 Polish decathlon competitors, competing in years 1985–2015, was used in principal component analysis (PCA) in order to decrease dimensionality and define the structure of new components representing the collected data.

The results achieved by the athletes in all ten events have significance and impact on the final, total result of decathlon (Furdal, 1986; Nowak, 1989). Pearson's linear correlation analysis carried out on the studied group has shown that almost all events have significant impact on the final results (for nine events $r_{xy} > 0.59$). The event with the smallest influence on the final result of a decathlon competitor is 1,500 m run, with the value of correlation coefficient at –0.29. Because 1,500 m run has a completely different character from the remaining events (endurance performance versus speed/strength preparation), while the motor requirements of decathlon involve mostly speed/strength effort (Vana, 2003), the speed/strength preparation is an important, while difficult to achieve, element of physical preparation (Dziadek, Iskra, Przednowek, 2016). On the basis of the analysis of the value of correlation coefficients between the decathlon competition, it was found that the most dependent on each other were the discus throw and shot put ($r_{xy} = 0.72$). Similar observations were observed by Furdal (1986), who analysed the results obtained by the 158 athletes starting in 1980-1983, who determined the largest dependencies for these two throwing competitions ($r_{xy} = 0.61$). Walaszczyk (1998), presented analysis relationships between the decathlon

components in three successive Olympic cycles (1985–1996) among of the 50 world's best decathletes, the largest correlation values in the analysed periods also concerned the discus throw and shot puts (r_{xy} = 0.59; 0.56; 0.76). The importance of the relationship between the discus throw and the shot put was also presented by Socha (1977) and Iskra (1990), where the calculated correlation coefficient were, respectively, r_{xy} = 0.56 and r_{xy} = 0.54.

Principal component analysis led to definition of 10 new components describing the total variance of the collected research material. Thanks to the Kaiser's criterion employed by us, the further part of the studies involved only the first three principal components which described over 69% of variables. The same criterion was used by Park, Zatsiorsky in their work (2011), who also employed three principal components, which described 70% of data variance regarding the competitors' performance competing in Olympic Games in years 1988–2008.

By using the factor map and internal structure of components it was observed that the first component (describing 45.3% of the total variance) is composed mostly of variables involving 100 m run and long jump, which are typical speed events and shot put which involves mostly strength and technique. The composition of the second component (14.5% of the total variance) involves considerable information content regarding 1,500 m and 400 m runs, which require endurance and run tactics, while the third component (variance at 10.0%) is composed mostly of high jump, the results of which are determined by speed and explosive strength of lower limbs of the competitor.

Conclusion

The employed method of principal component analysis and the obtained results allowed us to draw the following conclusions:

- the first three principal components describe 69% variables of acquired data,
- the largest share in the total variance is displayed by the first principal component, which is described by 100 m run, long jump and shot put,
- the study of the relationship between variables in the new plane displayed strong correlations between sprint events (100 m, 110 m hurdles) and long jump and pole vault which suggest that speed and power are most important abilities in selection for the decathlon,
- no significant correlations between the 1,500 m run and other events were found.

References

Bilić, M. (2015). Determination of taxonomic type structures of top decathlon athlets. *Acta Kinesiologica*, *9* (1), 20–23.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. NY: Springer Science+Business Media, LLC.

Chen, C., Zhang, B. (2016). Development trend of world men decathlon scores BP neural network analysis. *National Convention on Sports Science of China*.

Daszykowski, M., Walczak, B. (2008) Analiza czynników głównych i inne metody eksploracji danych. Retrieved from: http://www.chemometria.us.edu.pl (27.04.2018).

Dziadek, B., Iskra, J., Przednowek, K. (2016). Running preparation and the final decathlon score in terms of sports career development. *Physical Activity Review*, *4*, 115–123.

Ertel, S. (2011). Exploratory factor analysis revealing complex structure. *Elsevier*, *50* (2),196–200

Furdal, S. (1986). *Analizy modelowe wyników w dziesięcioboju*. Warszawa: Raport Instytutu Sportu.

Hardle, W., Simar, L. (2007). *Applied Multivariate Statistical Analysis*. L.A.: Springer Science & Business Media.

IAAF – *International Association of Athletics Federations* (2018). Retrieved from: http://www.iaaf.org/home.

Iskra, J. (1990). *Struktura czynnikowa dziesięcioboju*. Wrocław: WSWF Wrocław.

Karvonen, M.J., Niemi, M. (1953). Factor analysis of performance in track and field events. *European Journal of Applied Physiology*, *2* (15), 127–133.

Kassambara, A. (2017). *Practical Guide to Principal Component Methods in R*. STHDA. Retrieved from: http://www.sthda.com.

Matthews, P. (2013). *Athletics 2013. The International Track and Field Annual.* Cheltenham: Sports Book Ltd.

Nowak, L. (1989). *Optymalizacja osiągnięć w dziesięcioboju w kolejnych latach rozwoju zawodniczego*. Kraków: AWF Kraków.

Park, J., Zatsiorsky, V.M. (2011). Multivariate Statistical Analysis of Decathlon Performance Results in Olympic Athletes (1988–2008). *International Journal of Sport and Health Sciences*, *5* (5), 779–782.

PZLA – Polish Athletics Association (2018). Retrieved from: http://www.pzla.pl (1.03.2018).

PZLA – *Zestawienie tabel (1985–2006)*. Warszawa: Dział Sportowo-Techniczny.

Quercetani, R.L. (2000). *Athletics. A history of modern track and field athletics – Men and Woman (1860–2000)*. Mediolan: SEP Editrice.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from: https://www.R-project.org.

Socha, S. (1977). *Sterowanie treningiem i prognozowanie wyników w wielobojach*. Katowice: WSWF Katowice.

Stemmler, M., Bäumler, G. (2005). The Detection of Types among Decathletes using Configural Frequency Analysis (CFA). *Psychology Science*, *47* (3/4), 447–466.

STHDA (2018). *Principal component analysis*. Retrieved from: http://www.sthda.com.

Vana, Z. (2003). The training of the best decathletes. *New Studies in Athletics*, *4*, 15–30.

Walaszczyk, A. (1998). *Wybrane uwarunkowania osobnicze osiągnięć sportowych kobiet i mężczyzn w wielobojach lekkoatletycznych*. Katowice: AWF Katowice.

Zaciorski, W, Godik, M. (1963). Osnovnye faktory trenirovannosti v legkoatleticeskom desjatiborie (opyt faktornogo analiza). *Teorija i Praktika Fiziczeskoj Kultury*, *8*, 27–30.

Cite this article as: Dziadek, B., Iskra, J., Przednowek, K. (2018). Principal Component Analysis in the Study of Structure of the Best Polish Decathlon Competitors from the Period between 1985–2015. *Central European Journal of Sport Sciences and Medicine*, *3* (23), 77–87. DOI: 10.18276/cej.2018.3-08.