Colloquium Biometricum 48
2018, 18–26

APPLICATION OF MULTIVARIATE STATISTICAL ANALYZES FOR COMPARISON OF *BETULA* POLLEN SEASONS 2001-2016 IN LUBLIN

Agnieszka Kubik-Komar

Department of Applied Mathematics and Computer Science
University of Life Sciences in Lublin
e-mail: agnieszka.kubik@up.lublin.pl

Summary

The aim of the study was finding similar groups of birch pollen season in Lublin in 2001-2016 based on the pollen seasons parameters such as start, end, duration, peak date, peak value and annual pollen sum. Cluster analysis was used along with an optimal number of clusters algorithms as well as principal components analysis. Seasons were grouped in three clusters: 1) characterized by the late start and short duration, 2) characterized by the late end and low concentration of pollen and 3) characterized by the early start, long duration and high concentration of pollen.

Keywords and phrases: cluster analysis, optimal number of clusters, principal component analysis, pollen seasons

Classification AMS 2010: 62H25, 62H30

1. Introduction

Birch is a common tree in the north-western and central Europe, and its pollen is highly allergenic. In the majority area of Poland, birch pollen seasons start in the first decade of April, however, they differ considerably from each other, mainly due to changing weather conditions, especially because of the difference

in temperature (Stach et al., 2008). In Poland, birch pollen reaches very high concentrations in the atmosphere, and the number of pollen grains in the pollen season is also determined by the conditions in the previous year during the formation of flower buds (Latałowa et al., 2002).

The aim of the study was to analyze the diversity of birch pollen season in Lublin in 2001-2016 based on the characteristics of pollen seasons and to find groups of similar seasons. For this purpose agglomerative cluster analysis was used along with algorithms for determining the optimal number of clusters and principal components analysis. There are many indexes in the literature describing the optimal number of clusters (Charrad et al., 2014). In this paper, two of them were used. 1) CH index, 2) Silhouette index.

All calculations and graphs were made using the R open source software (R Development Core Team, 2013).

2. Materials and methods

Aerobiological data

Pollen monitoring was performed in the period 2001–2016. A Hirst-type sampler (Lanzoni VPPS 2000) was used for pollen trapping. It was placed on the flat roof of a building of the Lublin University of Sciences (in the center of the city) at a height of 18 m (Piotrowska & Kubik-Komar, 2012b).

The following characteristics of pollen season have been set:

- Start – a day of the year, counted from January the 1st, from which the pollen season begins
- End - a day of the year, counted from January the 1st, in which the pollen season ends
- Duration - the difference between End and Start
- Peak.date - maximum daily concentration date
- Peak.value – maximum daily pollen concentration
- Annual.total - total annual pollen sum of average daily pollen concentration.

In order to define the pollen seasons the 95% method was applied, in which the start is defined as the day when 2.5% of the season's catch had been recorded and the end occurs when 97.5% of the total catch had been reached (Andersen, 1991; Stach et al. 2008).

Statistical analysis

Before performing multivariable analyzes, the correlation of features was checked and the parameters strongly correlated with others were removed from the dataset. Moreover, the data were standardized before analysis in order to avoid the effect of the differences in measurement units between the parameters on the values of Euclidean distances.

Hierarchical cluster analysis was applied for finding similarity between seasons and to group them according to studied characteristics. These results were presented graphically in the form of dendrogram where the distance of linkage is a measure of pollen seasons similarity. The joining algorithm was based on the Euclidean distance and the Ward method of linkage. Optimal number of groups was achieved on the base of 1) CH index and 2) Silhouette index.

Since using different algorithms sometimes results in indicating the different optimal number of classes both indices were used to compare the results and make the obtained number more reliable.

1) CH index (Calinski & Harabasz, 1974; Mufti et al., 2000) is defined as pseudo-statistic F in the form of:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}$$

where k denotes the number of clusters, n denotes the number of objects and B(k) and W(k) denote the between and within cluster sums of squares of the partition, respectively. An optimal number of clusters is then defined as a value of k that maximizes CH(k).

2) Silhouette index (Rousseeuw, 1987) measures how close each object in a cluster A is to the object in its neighbouring clusters B.

Thus, for i-th object in A class $s(i)$ is defined in a way:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}},$$

where $a(i)$ is average dissimilarity of i-th object to all other objects of A,

$$b(i) = \min_{C \neq A} d(i, C)$$

for $d(i, C)$ defined as average dissimilarity of i-th object to all objects of $C \neq A$.

The cluster B for which $d(i, B) = b(i)$ is called the neighbour of the i-th object. The value $s(i) \in [-1,1]$ and $s(i)>0$ means that the *i-th* object is well-clustered, s(i)=0 means that *i-th* object could be as well in cluster A as in B, while s(i)<0 means this object is misclassified.

According to the definition presented above one can count Silhouette score for each data item. To obtain one value for the whole dataset average value $\bar{s}$ of $s(i)$ is counted. Each value of k yields a different average silhouette width $\bar{s}(k)$. The optimal number of classes is determined by

$$SC = \max_k(\bar{s}(k))$$

Principal Component Analysis was applied to compare pollen season in the scatterplot and to find the seasonal features that distinguish founded groups.

3. Results

At the beginning, before multivariate analyzes were performed, the correlation between season pollen parameters was checked. The results are presented in the graphical form of correlogram (Fig. 1).

The highest value of correlation was noticed between the start of the pollen season and the peak data (0.899) and between the maximal value of pollen and total annual pollen sum (0.916). This fact resulted in removing Peak.date as well as Peak.value variables from the further analysis.

The remaining variables constituted the dataset for agglomerative cluster analysis, which results presents in the form of dendrogram (Fig. 2)

The structure of the tree in Fig. 2 suggests that the most similar seasons according to pollen characteristics were 2001 and 2006, moreover 2002, 2009 and 2015 as well as 2007 and 2008.

After achieving results of cluster analysis the optimal number of classes was obtained on the basis of the above-described indices. In both cases, k = 3 was considered the optimal value (CH = 7.367, SC = 0.282) and therefore the following 3 groups of pollen seasons were proposed:

- Group 1 covering seasons 2001, 2003, 2004, 2006 and 2013,
- Group 2 constituted by seasons 2005 and 2011,

- Group 3 to which the seasons 2002, 2007, 2008, 2009, 2010, 2012, 2014, 2015 and 2016 were included.

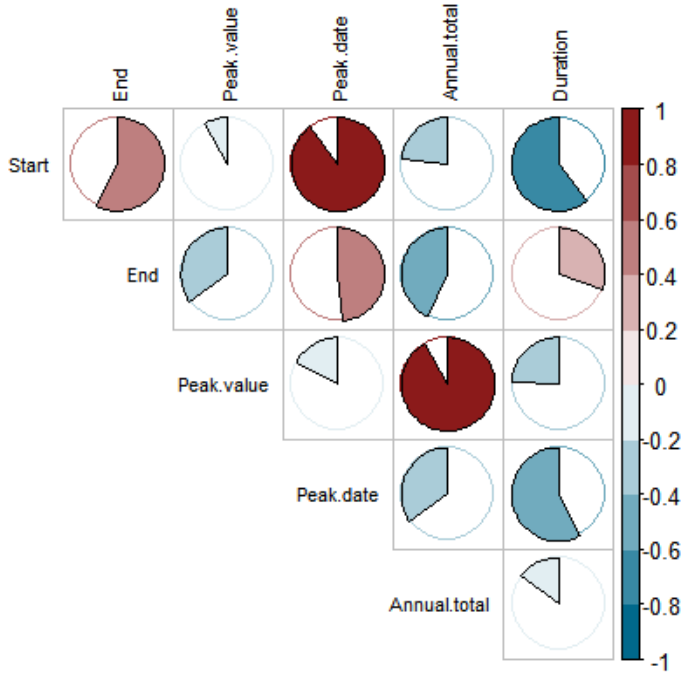

Fig. 1. The correlogram of studied features

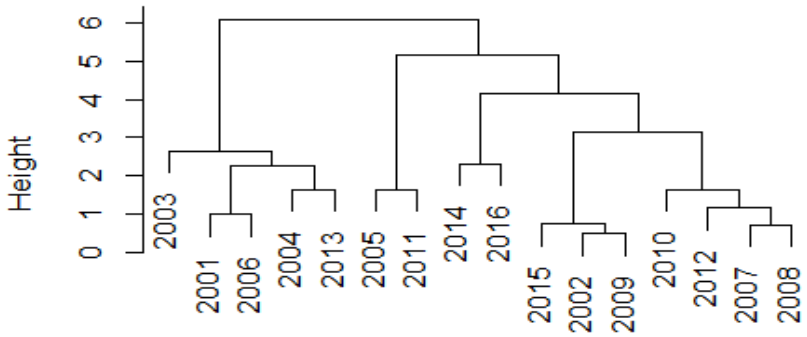

Fig. 2. The dendrogram of pollen seasons

The next step of the analysis was the application of PCA to reduce the size of the data space and to present pollen seasons on a single scatterplot. The aim was to compare them and define those characteristics of the season, which had the greatest impact on the determination of the above groups.

The results of the analysis indicated that only two eigenvalues exceeded 1, thus two new factors were determined, which explained almost 84% of the variability of the whole system. Factor loadings were presented in Table 1.

Table 1. Factor loadings after VARIMAX rotation

	PC1	PC2
START	0.50	0.85
END	0.92	0.06
ANNUAL.TOTAL	-0.75	0.05
DURATION	0.31	-0.93

Taking into consideration the coefficients in Table 1, an attempt was made to interpret the designated new variables. High PC1 values had seasons ending late with a low concentration of pollen in the air, while short seasons with a late beginning were characterized by high PC2 values.

Factor scores of studied seasons are presented in Fig. 3, in which additionally selected groups of clusters were marked.

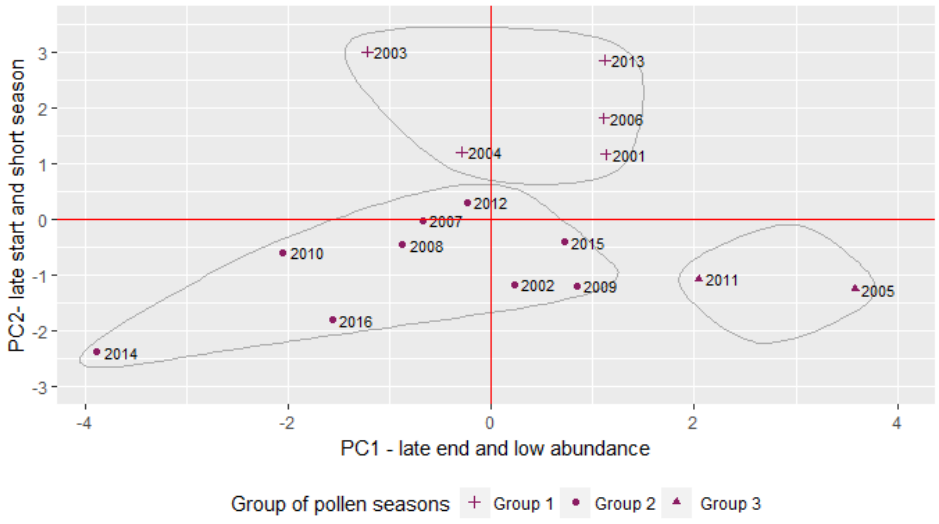

Fig. 3. The scatterplot of pollen seasons factor scores in PC1-PC2 coordinate system

Based on this chart, it can be concluded that Group 1 includes seasons characterized by a late start and short duration, while in Group 2 are seasons beginning late and not very abundant. Group 3 constituted the remaining pollen seasons and the variation in the terms of pollen season characteristics was the largest in this group. In most cases, these are seasons beginning early and lasting longer. The variability mainly concerned other features of the season, as one can notices seasons ending later with a smaller abundance like 2002, 2009 or 2015, as well as the 2014 season, characterized by very early start and end and extremely high concentration of pollen in the air.

4. Discussion

Grouping of similar pollen seasons is a proceeding found in literature in the field of aerobiology. In the papers of Piotrowska & Kubik-Komar, (2012a), and Malkiewicz & Klaczak, (2011), the seasons of the Poaceae family were compared in terms of pollen seasons dynamics by the hierarchical clustering as well as the k-means method. In addition, similar this paper, in the first article mentioned above, the pollen season parameters as grouping features were also used.

A completely variant example of the hierarchical cluster analysis application in aerobiological research is the grouping of particular days of the holm oak seasons in order to model diurnal pollen cycles for each sampling site based on the bi-hourly counts (Hernández-Ceballos et al., 2015). Another example is a comparison of 13 sites located in different biogeographical areas of Central and Eastern Europe based on two features of the *Artemisia* pollen season (Grewling et al., 2012).

However, according to the knowledge of the author, any algorithm of the optimal classes number determination after cluster analysis was not applied, in papers covering this area.

PCA appearing in literature in this field was most often used to compare pollen seasons of different pollen species in terms of weather (Piotrowska & Kubik-Komar, 2012a; Piotrowska & Kubik-Komar, 2012b) In the paper by González Parrado et al. (2009) weather-related parameters and pollen counts in one dataset were used, which changes the way the results of this study could be interpreted.

Comparing the result of PCA of Piotrowska & Kubik-Komar (2012b), with groups constituted in this paper one can conclude that there is no relationship between seasonal weather conditions and characteristics of season. Although it can be noticed that the 2005 season, which ended late and was characterized by

low abundance, was a cool season with inclement weather, and the 2003 and 2006 seasons belonging to the Group 1 were extremely warm, but that's not enough to draw general conclusions. All the more so because it is known that characteristics of *Betula* pollen season are affected by the weather, especially the temperature, before pollen season beginning - of February and March as well as of the year preceding pollen (Piotrowska & Kubik-Komar, 2012b).

Other results of this research are very difficult to refer to literature due to the facts that pollen seasons of different taxa do not overlap, different periods of pollen seasons are tested by authors, as well as pollen season is largely determined by climatic and geographic conditions.

5. Conclusions

On the basis of multivariate analyzes, three groups of similar seasons have been distinguished in the studied period of time. The seasons in Group 1 were characterized by the late start and short duration and were as follow: 2001, 2003, 2004, 2006, 2013. Seasons 2005 and 2011 in Group 2 were characterized by the late end and low concentration of pollen. The rest of seasons constituted Group 3 and varied the most according to studied factors, however, most of them started rather early and lasted quite long and the concentration of pollen during these seasons was mostly high.

Acknowledgments

The author thanks dr hab. K. Piotrowska-Weryszko and prof. E. Weryszko-Chmielewska, Department of Botany, University of Life Sciences in Lublin, for providing pollen data for this study.

References

Andersen T. B. (1991). A model to predict the beginning of the pollen season. *Grana*, 30(1), 269–275. https://doi.org/10.1080/00173139109427810.

Calinski T., Harabasz J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*. https://doi.org/10.1080/03610927408827101.

Charrad M., Ghazzali N., Boiteau V., Niknafs A. (2014). **NbClust** : An *R* Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61(6). https://doi.org/10.18637/jss.v061.i06.

González Parrado Z., Valencia Barrera R. M., Fuertes Rodríguez C. R., Vega Maray A. M., Pérez Romero R., Fraile R., Fernández González D. (2009). Alternative statistical methods for

interpreting airborne Alder (Alnus glutimosa (L.) Gaertner) pollen concentrations. *International Journal of Biometeorology*, 53(1), 1–9. https://doi.org/10.1007/s00484-008-0184-1.

Grewling Ł., Šikoparija B., Skjøth C. A., Radišić P., Apatini D., Magyar D., Smith M. (2012). Variation in Artemisia pollen seasons in Central and Eastern Europe. *Agricultural and Forest Meteorology*. https://doi.org/10.1016/j.agrformet.2012.02.013.

Hernández-Ceballos M. A., García-Mozo H., Galán C. (2015). Cluster analysis of intradiurnal holm oak pollen cycles at peri-urban and rural sampling sites in southwestern Spain. *International Journal of Biometeorology*. https://doi.org/10.1007/s00484-014-0910-9.

Latałowa M., Miętus M., Uruska A. (2002). Seasonal variations in the atmospheric Betula pollen count in Gdańsk (southern Baltic coast) in relation to meteorological parameters. *Aerobiologia*. https://doi.org/10.1023/A:1014905611834.

Malkiewicz M., Klaczak K. (2011). Analysis of the grass (Poaceae L) pollen seasons in Wrocław , 2003-2010. *Acta Agrobotanica*, 64(4), 2003–2010.

Mufti G. B., Bertrand P., Moubarki L. El. (2000). Determining the number of groups from measures of cluster stability. *Proceedings of International Symposium on Applied Stochastic Models and Data Analysis*, (January 2005), 404–413.

Piotrowska K., Kubik-Komar A. (2012a). A comparative analysis of Poaceae pollen seasons in Lublin (Poland). *Acta Agrobotanica*, 65(4), 39–48. https://doi.org/10.5586/aa.2012.020.

Piotrowska K., Kubik-Komar A. (2012b). The effect of meteorological factors on airborne Betula pollen concentrations in Lublin (Poland). *Aerobiologia*, 28(4), 467–479. https://doi.org/10.1007/s10453-012-9249-z.

R Development Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. *R Foundation for Statistical Computing, Vienna, Austria.*

Rousseeuw P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Stach A., Emberlin J., Smith M., Adams-Groom B., Myszkowska D. (2008). Factors that determine the severity of *Betula* spp. pollen seasons in Poland (Poznań and Kraków) and the United Kingdom (Worcester and London). *International Journal of Biometeorology*, 52, 311-321.