

## Research topics and trends over the past decade (2001-2013) of Baltic Coleopterology using text mining methods

Yuno Do, Jarosław Skłodowski

Do Y., Skłodowski J. 2014. Research topics and trends over the past decade (2001-2013) of Baltic Coleopterology using text mining methods. *Baltic J. Coleopterol.*, 14(1): 1 – 6.

The Baltic Journal of Coleopterology (BJC) is one of the most important and respected literature sources in the field of coleopterology. Over the past decade (2001-2013), 270 articles in 13 volumes of BJC were published on all aspects of beetles (Coleoptera) including faunistics, taxonomy, biogeography, and ecology. For improving our understanding of coleopterology and in guiding the future of BJC, we reviewed the articles in BJC using bibliometric analysis methods of text mining to summarise the research reported in the scientific literature by a set of indicators. Text mining analysis has shown a clear reduction of journal terms and concepts to two knowledge domains dealing with taxonomy and ecology. In addition, carabid community and distribution subjects were continuously contributed to the journal while relatively new topics such as landscape preference and modeling arose significantly since 2009.

Keywords: bibliometrics, text mining, topic modeling, VOSviewer

*Yuno Do. Department of Biological Sciences, Pusan National University, Busan 609-735, Republic of Korea; e-mail: doyunho@gmail.com*

*Yuno Do, Jarosław Skłodowski. Department of Forest Protection and Ecology, Warsaw University of Life Sciences, Nowoursynowska 159, 02-776 Warszawa, Poland; e-mail: sklodowski@wl.sggw.pl*

### INTRODUCTION

The Baltic Journal of Coleopterology (BJC) is one of the most important and respected literature sources in the field of coleopterology. In the preface of the first volume of BJC published in 2001 (Barševskis, 2001), Arvīds Barševskis as the editor in chief introduced a particular aim of the Journal: “*Papers included in the first volume deal with the latest developments in research into the classification and ecology of beetles in the Baltic region. Special attention is paid to the taxonomy of Palaearctic beetles. Taxonomy will*

*be one of the Journal’s priorities in future. Descriptions of new species from Palaearctic regions, articles dealing with ecological issues, particularly those associated with the conservation habitats in the Baltic states and in the EU member-states, as well as reports about distribution of the species which are new to the fauna of the Baltic region.*”

Over the past decade (2001-2013), 270 articles in 13 volumes of BJC were published on all aspects of beetles (Coleoptera) including faunistics, taxonomy, biogeography, and ecology. Although

BJC contributes to the distribution of knowledge of Coleoptera taxonomy and ecology in the Baltic region, research topics and publication contents of BJC and its change over time have not yet been summarised. A review of this Journal will be helpful in improving our understanding of coleopterology and in guiding the future of BJC. In this paper, we had a look back at the last 10 years of BJC.

We utilised text mining, statistics, and historical and current literature to track the history of BJC, to determine where the Journal has been and where it is currently. For this review of BJC, we used bibliometric analysis methods of text mining to summarise the research reported in the scientific literature by a set of indicators. This process collects and collates publication contents using quantitative statistics on the growth of research, stratified by the year and topic (Maimon and Rokach, 2005; Thelwall, 2008).

## DATA COLLECTION AND ANALYSIS

Core data for bibliometric analysis including text mining were collected from the BJC website (<http://www.bjc.sggw.pl>). Titles, abstracts, and authors information of 270 articles in 13 volumes of BJC were analysed. For identifying topics within a text corpus, we used Paper Machine for Zotero, a program that allows for creating bibliographies and building large coherent texts. The topic model in Paper Machine for Zotero uses the MALLET package to perform latent dirichlet allocation (Blei et al. 2003), and by default displays the 10 most coherent topics, based on the metric system according to Mimno et al. (2007). A word cloud was created using Tagxedo ([www.tagxedo.com](http://www.tagxedo.com)). Word clouds give more prominence to words that appear more frequently in the source text. Displayed words are scaled by their frequency of occurrence. A density map as well as a cluster-density map were created using the VOS mapping and clustering techniques in VOSviewer ([www.papermachines.org](http://www.papermachines.org), [www.vosviewer.com](http://www.vosviewer.com)). Colours indicate the density of terms in the density map, and clusters of related terms in the cluster-density map. Statistical differences of the

number of articles and authors were tested using T-paired tests. In addition, linear regression was used to identify changes in the number of articles and authors. The latter was performed using PASW Statistics 18.

## RESULTS

Of the 270 articles (2001-2013) analyzed, 172 (63.70%) described new taxa and taxonomic data (Fig. 1a) while 92 articles (34.07%) were associated with ecological studies. There was a mean of 20.77 articles published per year (range = 11-33 articles) with peaks in 2003 (30 articles) and 2007 (33 articles). A total of 361 authors had contributed to BJC over the last 13 years (mean = 25.85 authors per year, Fig. 2b). The number of authors per paper was 0.75. The highest number of authors was 44 in 2003. The number of articles published and authors contributing showed no significant change (articles,  $R^2 = -0.005$ ,  $p = 0.811$ ; authors,  $R^2 = 0.017$ ,  $p = 0.668$ ). The number of authors of the first issue (published in April 2001) was significantly higher than the corresponding number for the second issue (October 2001), although the number of articles between the two issues was similar (articles,  $t = 0.31$ ,  $p = 0.38$ ; authors,  $t = 2.01$ ,  $p = 0.03$ ).

In the word cloud, Coleoptera, of course, was the most frequently used word of all articles. Terms associated with taxonomic groups such as Carabidae, Dermestidae, Histeridae, and Staphylinidae were frequently found in articles. Specifically in the work describing new taxa and faunistics, the terms new, species, genus, described, fauna, and data, to name a few, appeared more frequently in the word cloud (Fig. 2). Although the frequency of terms associated with taxonomy, such as assemblage, community, pine, ecological, time, etc., was lower, they were frequently used in the description of studies. Most studies were conducted in the Baltic region, especially in Latvia and in Poland.

Density maps showed a clear concentration of the Journal's terms and concepts in two knowledge domains dealing with Coleoptera taxonomy

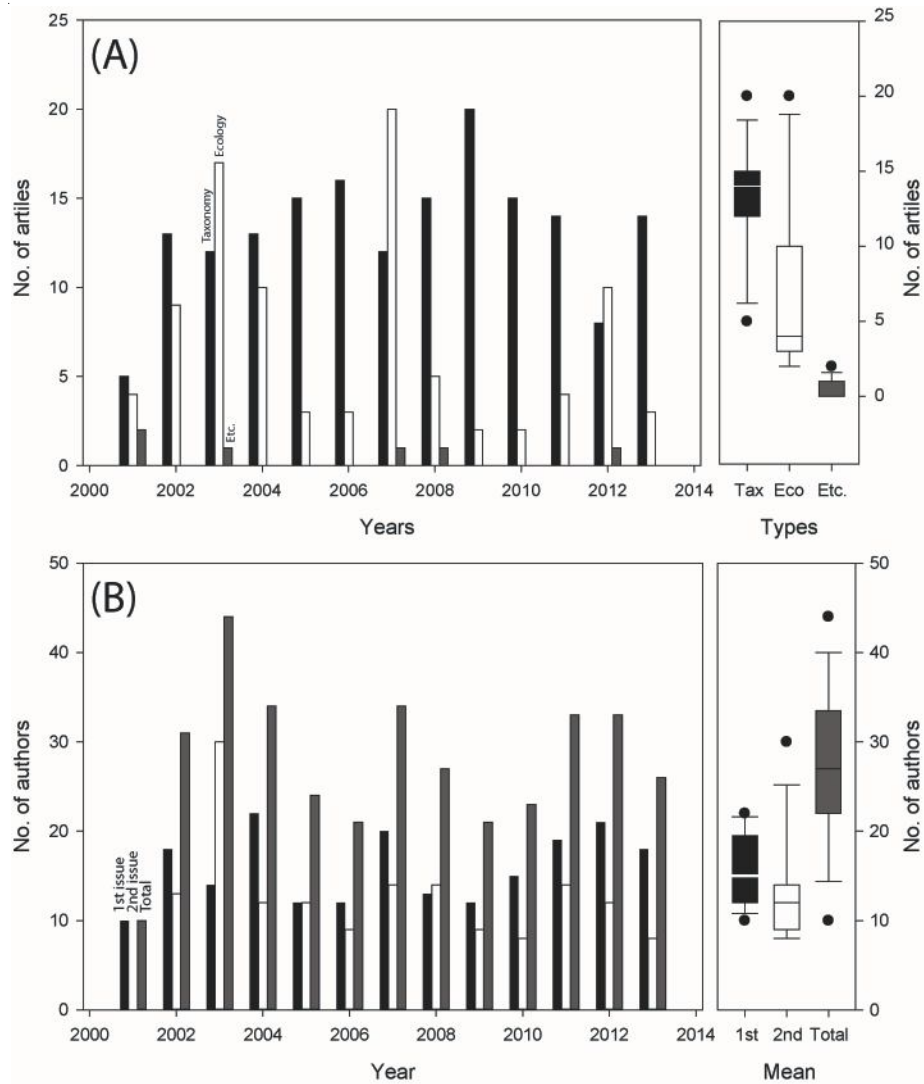


Fig. 1. Change of number of articles and authors in Baltic Journal of Coleopterology during 2001-2013

and ecology (Fig. 3). Taxonomy: In the cluster-density map (Fig. 3a), the green cluster of taxonomic groups comprised terms associated with Carabidae and Dermestidae while the blue cluster represented taxonomic studies of Chrysomelidae. Additionally, the yellow cluster representing Staphylinidae was clearly separate from the other clusters. Ecology: All terms associated with ecological research of Coleoptera coalesced. In the density map, new species and

associated taxa had the highest density (red, Fig. 3b). In addition, carabid beetles, abundance, composition, diversity, and forest had the highest density relating to ecological research. The cluster-density map showed different clusters of the article terms.

Topic modeling showed the changes of publication subjects over the last decade, depicting a clear image of ecological research in BJC (Fig 4).



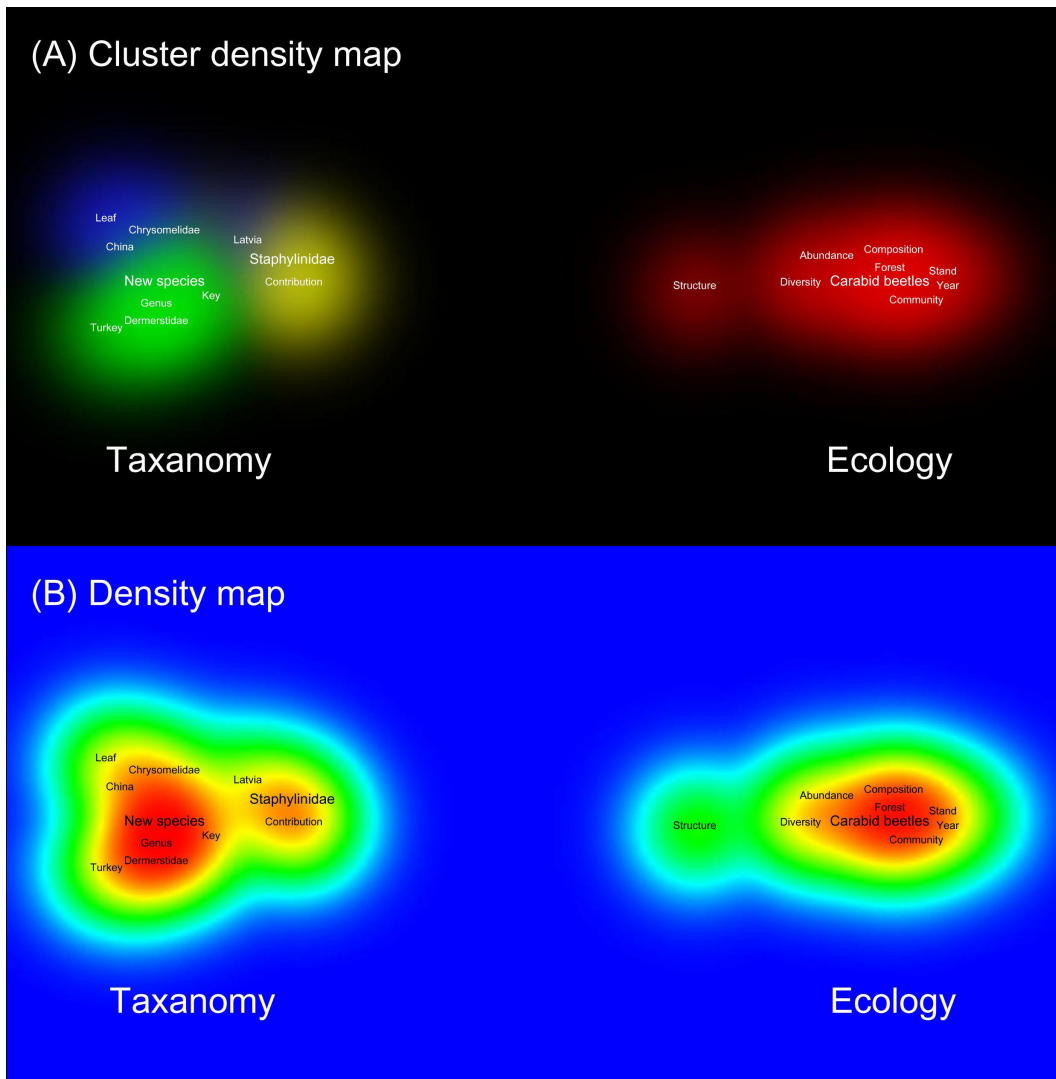


Fig. 3. Cluster density and density map for Baltic Journal of Coleopterology

certain areas and habitats. Most ecological studies were conducted in forests of Poland. Although carabid beetles as indicators responding to change of forest environments (e.g. fragmentation, restoration, management, etc.) are important topics in the Baltic region as well as in carabidology, there may be a paucity of studies of Coleoptera in agricultural land, urbanised areas, and grassland as relatively dominant habitat types in the Baltic region. After 2009, fortu-

nately, research topics have begun to diversify. As BJC enters maturity, the time is right to increase efforts to shed light on its history, efforts that will require cooperation among the publishers and editors as well as academic institutions and professional associations.

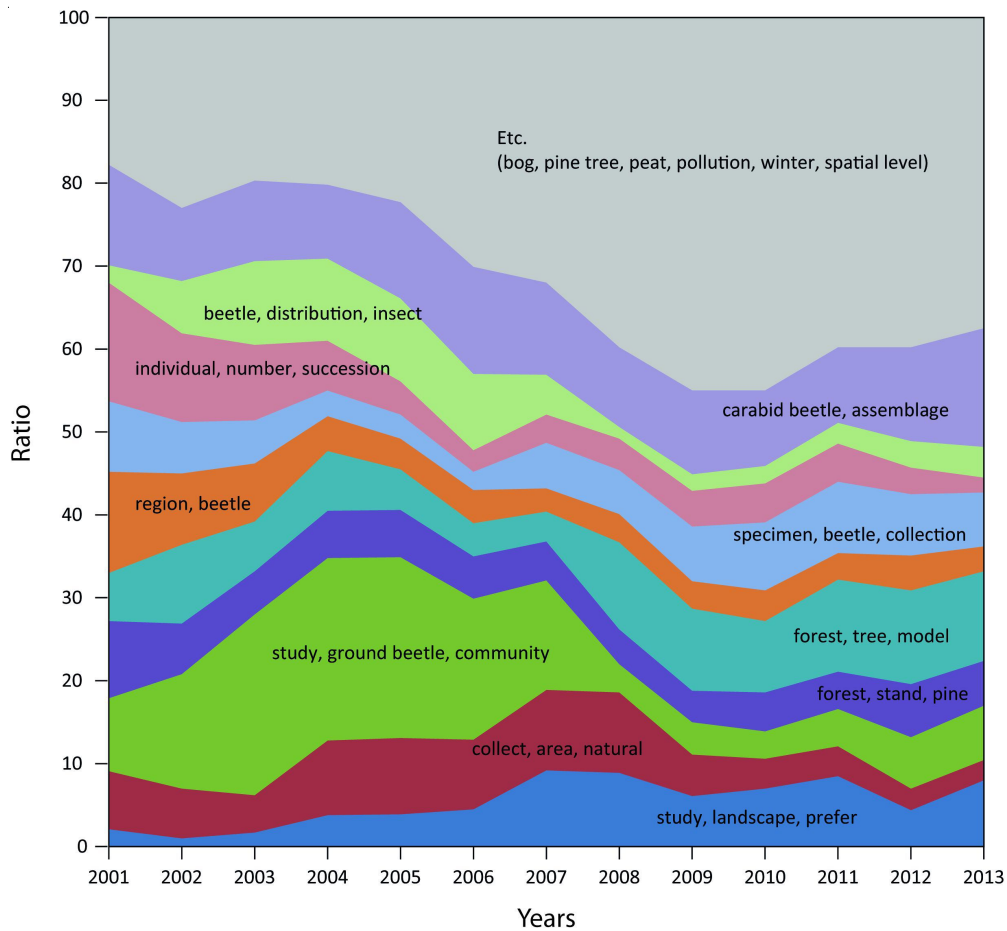


Fig. 4. Topic change of Baltic Journal of Coleopterology

**REFERENCES**

Barševskis A. 2001. Preface. Baltic Journal of Coleopterology 1 (1): 2.

Blei D. M., Ng A. Y., Jordan M. I. 2003. Latent dirichlet allocation. The Journal of Machine Learning research 3: 993-1022.

Maimon O. Z., Rokach L. 2005. Data mining and knowledge discovery handbook, Springer.

Mimno D., McCallum A. 2007. Expertise modeling for matching papers with reviewers.

Thelwall M., 2008. Bibliometrics to webometrics. Journal of Information science 34 (4): 605-621.

*Received: 14.04.2014*  
*Accepted: 10.06.2014*