Species Pattern Analysis in Long-Term Ecological Data Using Statistical and Biclustering Approach

Hyeonjeong Lee
Bio-Intelligence & Data Mining Laboratory, Graduate School of Electronics,
Kyungpook National University,
Republic of Korea
e-mail: dic1224@naver.com

Miyoung Shin
School of Electronics Engineering,
Kyungpook National University,
Republic of Korea
e-mail: shinmy@knu.ac.kr

***Abstract*—Analyzing long-term ecological data and appropriate visualization techniques are important for understanding biodiversity mechanisms and predicting effects of environmental changes. In this study, we applied an unconventional approach of finding species pattern, the tendency of species abundance monthly and annually in long-term ecological data, by using statistical and biclustering methods. We tended to find out the similarity between each species after summarizing long-term dataset, and then visualized a correlation matrix and network, which exhibit significant statistical association with each other. For detecting species sets frequently appearing together or showing similar variation in abundance, we also employed a clustering based association mining. For experiments, we used weekly abundance butterfly data from the Environmental Change Network (ECN) in the UK. We could find out how often sets of species show the repeated pattern in long-term species abundance data. The approaches we have described can enable researchers to gain insight of many other relationships like between various species and environmental factors. In addition, combining our methods with detailed analyses or assumptions, such as genetic associations between species and functional subsystems may especially be effective in further analysis.**

Keywords- long-term ecological data; association mining; visualization; species set; species abundance pattern.

I. INTRODUCTION

Analyzing long-term ecological data is important for understanding biodiversity mechanisms and predicting effects of environmental changes. Several long-term environmental monitoring projects, such as Terrestrial Ecosystem Research Network (TERN), National Ecological Observatory Network (NEON), and Long-Term Ecological Research (LTER) have attempted to manage and share records of climate and species in international networks [1]-[3]. Accordingly, appropriate data analyses and visualization methods play a significant role in providing more insights into underlying trends in long-term ecological data. Many studies have attempted to search various patterns or trends of species, mostly plotting abundance of individual species across time, without regarding for associations between species [4][5]. In this study, we aim to find sets of species showing similar abundance pattern in long-term ecological data. The rest of this paper is organized as follows. In Section Methods, our proposed methodology is represented. Section Results and Discussion draws the experimental results and discussion.

II. METHODS

In this study, we applied an unconventional approach of finding species pattern, the tendency of species abundance monthly and annually in long-term ecological data, by using statistical and biclustering methods. We first summarized the long-term dataset, and then tended to detect the presence of interesting trend by calculating the similarity between each species statistically. After that we visualized a correlation matrix and network, which exhibit significant statistical association with each other.

For detecting species sets frequently appearing together or showing similar variation in abundance, we also employed a clustering based association mining. Association rule mining finds interesting itemsets (in this case, sets of species) that occur frequently in a dataset [6][7], and biclustering clusters rows and columns of a data matrix simultaneously [8]. We applied the BiMax clustering algorithm which is relatively faster than traditional approaches like Apriori algorithm, since Apriori often create too many rules and is time consuming. For performing the BiMax clustering to find associated species under certain condition, we first constructed experimental data in a such way that rows and columns represent species and samples (all months in 18 years), respectively. After that we utilized discretization to assign either 0 (less than average abundance) or 1 (more than average abundance) to each value of monthly species abundance. That is, we only focused on species sets that appearing more than average abundance in every months. By doing so, we could find out interesting species-sets in the rule form of {species set} $\Rightarrow$ {major month and its abundance percentage}. We also visualized the species sets into species abundance heatmap illustrating monthly and annually repeated abundance pattern of species sets.

III. RESULTS AND DISCUSSION

For experiments, we used butterfly data from the Environmental Change Network (ECN) in the UK [9]. Weekly abundance records of 29 kinds of butterflies from

 ISBN: 978-1-61208-518-0

1994 to 2012 are analyzed, showing several possible species sets. We visualized the overall trend of dataset as shown in Fig.1. Correlation between each species in long-term data are shown in Fig.2. The color and width of boxes in Fig.2 (a) and edges in Fig. 2 (b) represent how much two species show the similar abundance in long-term data. The size of node in Fig. 2 (b) is proportional to its degree or the number of edges. We could also find out how often sets of species show the repeated pattern in long-term species abundance data by applying biclustering based association mining on month-species summarized dataset (Fig.3). We represented species sets as association rule forms, for example, {"Red admiral", "Meadow brown"} ⇒ {JUN 5.6%, JUL 61.1%}, which indicates that a species set including two species "Red admiral" and "Meadow brown" is appearing in June and July as a percentage of 5.6 and 61.1, respectively. Fig. 4 illustrates top 100 interesting species sets at a glance. Relationships between species and climate, i.e., temperature and wind speed are however not found in this experiment. Nevertheless, the results show that our approach have general use in finding the species sets for addressing species abundance patterns of interest. It might be showing better performance if more ecological data are accumulated in recent years or near future.

The approaches we have described can enable researchers to gain insight of many other relationships, for example, between various species and environmental factors. In addition, combining our methods with detailed analyses or assumptions, such as genetic associations between species and functional subsystems may especially be effective in further analysis.

ACKNOWLEDGMENTS

This subject is supported by Korea Ministry of Environment (MOE) as "Public Technology Program based on Environmental Policy (2014000210003)."

REFERENCES

[1] Terrestrial Ecosystem Research Network: TERN. [Online]. Available from: http://www.tern.org.au/

[2] M. Keller, D. S. Schimel, W. W. Hargrove, and F. M. Hoffman, "A continental strategy for the National Ecological Observatory Network," The Ecological Society of America, pp. 282-284, 2008.

[3] J. T. Callahan, "Long-term ecological research," BioScience, vol. 34, pp. 363-367, 1984.

[4] S. Benham, "The Environmental Change Network at Alice Holt Research Forest," Forestry Commission, pp. 1-12, 2008, ISSN: 1756-5758, ISBN: 973-0-85538-762-4.

[5] B. J. McGill, et al., "Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework," Ecology letters, vol. 10, pp. 995-1015, 2007, doi: 10.1111/j.1461-0248.2007.01094.x.

[6] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "A novel biclustering approach to association rule mining for predicting HIV-1-human protein interactions," PLoS One, vol. 7, e32289, 2012.

[7] R. Giugno, A. Pulvirenti, L. Cascione, G. Pigola, and A. Ferro, "MIDClass: Microarray data classification by association rules and gene expression interfals," PLoS One, vol. 8, e69873, 2013.

[8] A. Prelić, et al. "A systematic comparison and evaluation of biclustering methods for gene expression data," Bioinformatics, vol. 22, pp. 1122-1129, 2006.

[9] M. D. Morecroft, et al., "The UK Environmental Change Network: emerging trends in the composition of plant and animal communities and the physical environment," Biological Conservation, vol. 142, pp. 2814-2832, 2009.

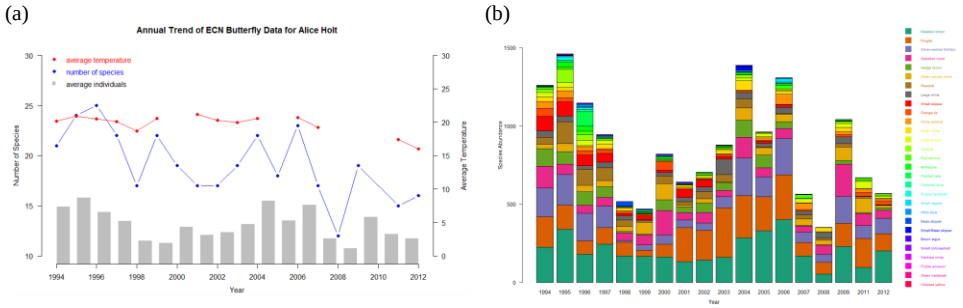

Figure 1. Overall trend of long-term ECN butterfly data for Alice Holt from 1994 to 2012: (a) annual trend of average temperature, number of species, and average individuals of butterflies (b) bar graph of butterfly species abundance and ratio at each year.

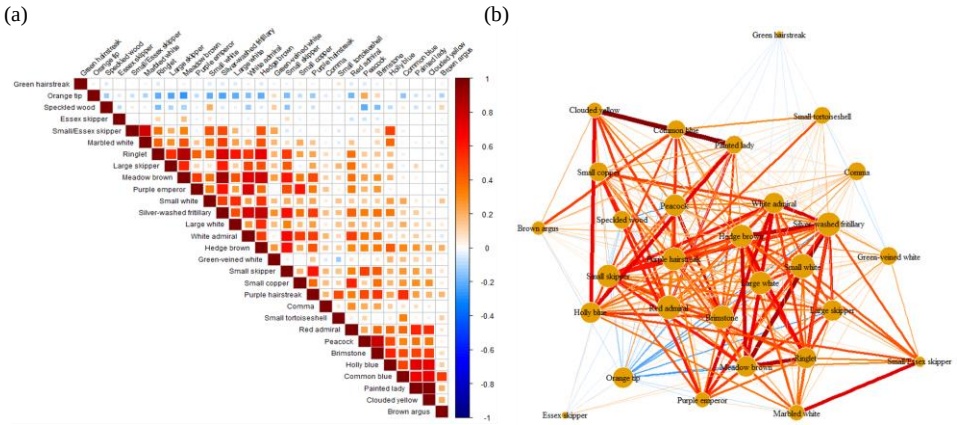

Figure 2. Correlation between each species are represented as: (a) species correlation matrix, and (b) species correlation network.

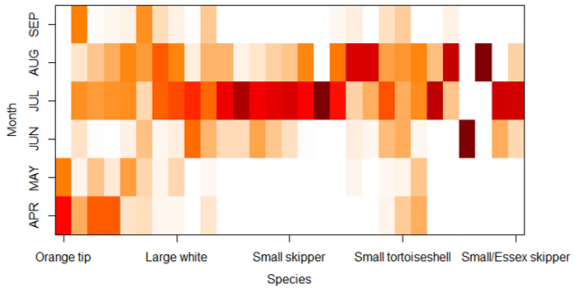

Figure 3. Species-month Summarized data with rows indicating months and columns indicating species of butterflies. The color represents how much individuals of each species are shown in each month on average from 1994 to 2012. This summarized data are used for further analysis.

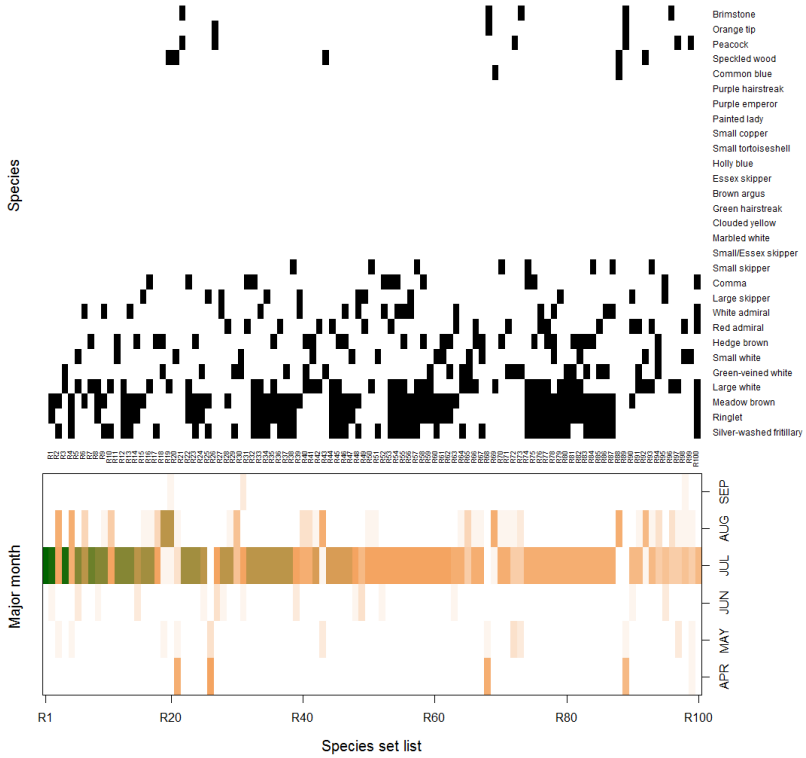

Figure 4. Top 100 interesting species sets frequently shown in each months among the datasets