Method of analysis of floral lists of higher plants

S I Ryabukhin

Institute of mining of the far Eastern branch of the Russian Academy of Sciences, 51, Turgenev str., Khabarovsk, 680000, Russia

E-mail: poleem@mail.ru

Abstract. Quantitative methods for analyzing data on living objects are an essential tool for research and modeling the state, stability, and productivity of ecological systems. In this paper, we consider one of the approaches that is useful for quantitative analysis, namely, when the initial information about living objects (in particular, about vascular plants) is presented in the form of descriptions of floral spectra that differ in structure and content (namely, lists of plant species). The proposed approach is based on the application of first-order predicate logic and is intended for the purpose of studying the species richness of plants in ecosystems.

1. Introduction

One of the indicators used to describe the state and sustainability of various ecological systems is the so-called biodiversity assessment. The diversity of plant species in the ecosystem under study is an important component of assessing the biodiversity of the ecosystem as a whole. One way to assess the diversity of plant species is to compare the species richness of different plant communities within an ecosystem. For the purpose of identifying species communities and analyzing their composition, so-called descriptions of floral spectra or, in other words, lists of plant species confined to a particular territory, limited, for example, by regional or administrative contours, are widely used.

A species community is considered to be a set of plant species associated with a habitat (ecotope) and isolated from a certain (for example, regional) set of species. As you know, this separation is due to the existence of different boundaries: biocenotic; soil; geological; climatic, etc. As a rule, the identification of species communities is carried out in the course of in-house processing of the results of field biocenotic studies [1]. The results of such processing are published in the form of various annotated [2] and non-annotated lists of plant species. In the future, we will refer to unannotated lists as "floral lists", meaning by this concept a complete description of the floral spectrum of a certain ecosystem [3]. As is known, various methods of quantitative analysis are used to study the floral spectrum [4]. One of the well-known methods for analyzing species populations based on their comparison is the "extensive classification" method [5]. In this case, views are grouped into classes that depend on how they are grouped. Then the classes are grouped into larger agglomerations, and the procedure for combining classes is performed by re-matching all objects (plant species) included in the classes. The results of extensive classification are usually visualized using dendrograms [5]. If the source of the initial data is floral lists that include a sufficiently large number of elements (for example, as in the case of floral spectra of regions [2, 6]), the use of extensive classification methods and dendrograms becomes extremely difficult and, sometimes, impractical [4].
In our opinion, it is more correct to call the extensive classification method the method of hierarchical intersecting clustering. This clarification, on the one hand, allows us to explain the low practical usefulness of the results of "extensive classification", in particular, the tree hierarchy of classes. On the other hand, it is cluster analysis that can make it possible to isolate clusters - communities of species from the original set of species, compare such clusters, and thus obtain an estimate of the diversity of species in communities. Obviously, if the original set of plant species is represented by a large floral list (for example, with the number of elements more than 10^n|n>3) or a set of lists, then the appropriate list processing algorithms must be used to successfully apply cluster analysis.

On the one hand, the availability of floral lists, including those published on the Internet [6], opens up wide opportunities for researchers. On the other hand, floral lists may be compiled by different researchers and have different structures. For example, the composition of these lists may include the following elements: 1) species names, in particular, the unified Latin and alternative names; 2) indicator of the relative area of the projection of the species on the soil surface as a percentage [7]; 3) various phytological properties of the environment, such as humidity, soil acidity, etc. [8]; 4) characteristics of the species: morphological, biochemical, etc.; 5) geographical data: longitude and latitude of the location of the detection point of the species; names of the range or landscape, etc. It should also be noted that the elements of the floral lists specified in paragraph 5 can be represented in different ways: a) using standard unified names (for example, [9]); b) using phrases from significant and non-significant parts of speech.

As a rule, floral lists are published as text data separated by commas (i.e. the "csv" separator is used). In other words, floral lists are structured data that are not linked by a conceptual schema. Since in most cases floral lists are sets of names, the analysis of elements of such lists requires the use of special methods designed for processing categorical data.

As is known, differentiating quantitative estimation methods based on the presence or absence of plant species in the compared sets are used as criteria for hierarchical intersecting clustering [10]. Therefore, the clustering algorithm for plant species applied to the original set of species represented by the list should be based on the principles of associative analysis of categorical data.

This paper describes a method for analyzing structured descriptions of floral spectra that differ in structure and content. The proposed method is based on the application of first-order predicate logic and is designed to study the diversity of plant species in the floral spectrum of the studied ecosystems.

2. Results
The essence of the proposed method of analysis of floral lists is to apply propositional logic based on the relationships of elements of the floral list. In this case, the list is considered as a set of sets of elements. First-order predicates are used to describe the relationships of elements in sets. In the process of unifying predicates, predicate variables take the symbolic values of list items. In this case, the unified predicates are associated with a group of sets. Then the validity of certain conditions is checked, for example: conditions for maximum similarity or dissimilarity of the received sets; conditions for equivalence (or difference) of elements of sets that perform the same role, and so on. As a result, a set of target predicates is formed that describe the sets of the original floral list. In General, this procedure implements the process of cluster analysis of species present in the list. The peculiarity of this method is the structure of the predicate model, as well as the organization of the predicate unification procedure, which allows you to effectively form and agglomerate sets of species.

2.1. Formal model of a floral list
Based on the general description of the structure and content of floral lists (see the first part of this article), we present such a list in the form of a formal Backus-Nauer scheme (see expression 1).

\[ <L> ::= <S>[,K][,<I>][,<G>] \]
\[ <S> ::= S^1[^{S^2}] \]
\[ <I> ::= I^1[^{I^2}] \]
\[ <G> ::= G^1[^{G^2}] \]
Expression 1 uses the following values: \( L \) – name of the list (corresponds, for example, to a certain floral area); \( S \) – names of the species; \( S^1, S^2 \) – unified and alternative names of the species; \( K \) – indicator of "coverage"; \( I \) – phytointicators; \( I^1, ..., I^m \) - various types of phytointicators; \( G \) – geographical identifiers; \( G^1, ..., G^n \) - various types of geographical identifiers. Note that the order of elements following the type name (s) is not important.

2.2. Predicate interpretation of the floristic elements of the list

First, we transform the list schema elements (see expression 1) into first-order predicates (see expressions 2-5).

\[
(S^1, S^2) \leftrightarrow S(X^{S1}, X^{S2})\{[X^{S1}], ..., [X^{S2}], [X^{S2}], ..., [X^{S2}]\}
\]

\[
K \leftrightarrow K(X^I)[/ X^I,..., X^n_I]
\]

\[
(I^1, ..., I^m) \leftrightarrow (I^1(X^I))[X^I, ..., X^I], ..., I^m(X^I)][X^I, ..., X^I]
\]

\[
(G^1, ..., G^n) \leftrightarrow (G^1(X^G))[X^G, ..., X^G], ..., G^n(X^G)][X^G, ..., X^G]
\]

In expressions 2-5, the variables \( X^{S1}, X^{S2}, X^K, X^I, X^G, X^G \) correspond to the particular values of the list items, and we will assume that the original list is not empty. This assumption will allow us to avoid using quantifiers to determine predicate variables and simplify the presentation significantly.

2.3. Predicate logical model of a floral list

Let's make a General predicate model of the list in the form of a propositional expression (see expression 6).

\[
L(X^{S1}, X^{S2}, X^K, X^I, X^G) \rightarrow S(X^{S1}, X^{S2}) \land [K(X^I) \lor \neg K(X^I)] = [I^m(X^I) \lor \neg I^m(X^I)] = [G^n(X^G) \lor \neg G^n(X^G)]
\]

In expression 6, the predicate \( L(X^{S1}, X^{S2}, X^K, X^I, X^G) \) describes the contents of the list, and \( L \) means the name of the list.

2.4. The scheme of the unification of predicates, grouping sets

We will show a scheme for performing predicate unification using the example of the rule for grouping species by the common value of the geographical identifier (see expression 7).

\[
G^n(x^G) \land S(X^{S1}, X^{S2}) \land L(X^{S1}, X^{S2}, ..., x^G) \rightarrow R(X^{S1}, X^{S2}, x^G, J^k_i)
\]

The truth of the rule in the left part of expression 7 is checked by alternately substituting all values of variables \( X^{S1}, X^{S2} \), that are unified with antecedent predicates \( G^n(x^G) \land S(X^{S1}, X^{S2}) \land L(X^{S1}, X^{S2}, ..., x^G) \), in which the variable \( X^G \) takes a constant value \( x^G \). The consequent \( R(X^{S1}, X^{S2}, x^G, J^k_i) \) describes a group of list items, and the \( J^{k_i} \) is the ID of this group. The group number can be used as this ID (for example, a counter with increment). Similarly, the unification procedure is performed for grouping rules based on the values of \( X^k \) variables (see expression 8) and \( X^m \) (see expression 9).

\[
K(x^K) \land S(X^{S1}, X^{S2}) \land L(X^{S1}, X^{S2}, x^K, ...) \rightarrow R(X^{S1}, X^{S2}, x^K, J^{k_i})
\]

\[
P^m(x^m) \land S(X^{S1}, X^{S2}) \land L(X^{S1}, X^{S2}, x^m, ...) \rightarrow R(X^{S1}, X^{S2}, x^m, J^{k_i})
\]

Pairs of expressions 6-7, 6-8, and 6-9 make up three logical systems that unify and resolve to form a set of three types of groups: \( R(X^{S1}, X^{S2}, x^G, J^{k_i}); R(X^{S1}, X^{S2}, x^K, J^{k_i}); R(X^{S1}, X^{S2}, x^m, J^{k_i}) \).

2.5. Rules for checking agglomeration criteria for set groups

So, clusters are formed based on previously received groups of suites due to the implementation of verification rules, i.e. criteria for combining groups. We will show a scheme for executing such a rule using an example of a group of sets of the type \( R(X^{S1}, X^{S2}, x^G, J^{k_i}) \) (see expression 10).
max\(F(R(X, Y, x^{Gn}, J^R), R(Y, Y, x^{Gn}, J^{R+1})))\rightarrow C(X, Y, x^{Gn}, J^C) \land C(Y, Y, x^{Gn}, J^C)
\)

The predicates \(C(X, Y, x^{Gn}, J^C)\) and \(C(Y, Y, x^{Gn}, J^C)\) in expression 10 describe a single cluster with the \(J^C\) number. This assembly is a consequence of maximizing the similarity function for a pair of sets \(R(X, X, x^{Gn}, J^R)\) and \(R(Y, Y, x^{Gn}, J^{R+1})\) from a group of sets of type \(R(X, X, x^{Gn}, J^R)\). The similarity function \(F\) of sets provides a quantitative estimate of similarity obtained by comparing the names of species, i.e. the values of variables \(X\) and \(Y\). This assessment is called the similarity index of sets of plant species [4,5,10].

We will do the same with sets of groups like \(R(X^{G1}, X^{G2}, x^K, J^R)\) and \(R(X^{G1}, X^{G2}, x^{Im}, J^R)\) and get clusters \(C(X, X, x^K, J^C)\) and \(C(Y, Y, x^{Im}, J^C)\). At the end of this procedure, a set of clusters will be generated, consisting of sets, each of which includes the name of the plant species and a characteristic of the habitat.

2.6 Agglomeration of groups of sets

The cluster aggregation scheme is based on pairwise comparison of previously obtained clusters, checking conditions, and generating new cluster agglomerations (see expression 11).

\[P(VZ) \rightarrow \max(F(C(X, X, V, I^F), C(Y, Y, Z, J^F))) \rightarrow C(X, X, V, Q) \land C(Y, Y, Z, Q)\]

In expression 11, the matching function \(F\) and the maximization function for matching sets \(\max\) are similar to the corresponding functions in expression 10. The predicates \(C(X, X, V, I^F)\) and \(C(Y, Y, Z, J^F)\) are unified by the \(VZ:\
(V, Z)\{\{X^k, ..., X^l, J^F\}, \{X^{in}, ..., X^{im}, J^F\}\}, V \neq Z\)

This variables denote different characteristics of species habitats. Variables \((I^F, J^F)\{I...N\}\) mean cluster IDs. As follows from expression 11, the unification of these predicates will be true if the variables \(V\) and \(Z\) have some relation described by the predicate \(P(VZ)\). For example, this relationship may mean that the habitats of plant species may be comparable in characteristics indicated by the variables \(V\) and \(Z\). The result of agglomeration is a new cluster with the number \(Q\), to which sets \(C(X, X, V, I^F)\) and \(C(Y, Y, Z, J^F)\) are linked.

3. Discussion

The method of analysis of floral lists described in the second part of this article, in fact, implements overlapping hierarchical clustering of a certain set of plant species, information about which is represented by species names.

Let's list the limitations of the proposed method, related to the composition and content of the floral list. First, the floral list should not contain elements equivalent in content (with the exception of alternative names of plant species, if any are present in the list).

Secondly, if there are no clear boundaries [10] of habitats (for example, if you specify the topographic contiguity of floral areas), you need to create additional predicate forms (see p 2.6).

Third, floral lists must be suitable for machine processing, that is, such lists must have markup using special formatting characters (namely “csv”).

Fourth, the initial floristic lists must contain the verified name of the species. In particular, the unified and alternative name of the plant species must constitute a unique pair of values. On the other hand, we should note the sensitivity of this analysis method to the content of the list items, since any character combinations are allowed.

Let's list some aspects of implementing the method. The proposed method of list analysis is effective when using logical programming methods, and its implementation requires the availability of appropriate development environments. The absence of a set of logical conclusions in the proposed predicate propositional model precludes the need to check it for consistency. At the same time, the predicate model ensures that the results of list analysis are independent of the order of its elements, while few currently known methods for analyzing categorical data sets are unstable (for example, the CLOPE method [11]).

Regarding the effectiveness of the method, it should be said that the algorithmic estimation of the
complexity of its implementation is quite acceptable and is described by the function $O=f(m^n)^{k}$, where: $O$ – the complexity estimate; $m$ – the number of species in the list; $n$ – the number of lists; $k$ – the number of conditions and criteria.

It should also be noted that the results of using the associative method of list analysis can be considered adequate if the degree of territorial study of the studied ecosystem (for example, regional) is uniform, the floral spectrum of which is represented by lists of plant species. The proposed method of analysis is tested on the example of floral lists in the abstract of the flora of the Khabarovsk territory [2].

4. Conclusion

The assessment of the state of ecological systems, in particular, the assessment of the state of the diversity of plant species in the study region, can be used for various purposes. For example, for the purpose of conducting model experiments when studying the response of the ecosystem to anthropogenic impacts. To study and clarify the existing boundaries of various ecotopes, as well as for other purposes. Like any other methods of formal quantitative analysis of data on plants and their communities, the proposed method can serve as a means of creating various hypotheses, the need to test which will be one of the reasons for conducting additional field Botanical research.

If the accounting of biological objects (in particular, the accounting of higher plant species) is conducted by recording the presence of such objects in a certain study area, then the use of stochastic (for example, regression) models [12] for the analysis of communities of plant species is impractical.

This paper describes a method for analyzing the results of accounting for higher plant species presented in the form of so-called floral lists of species. On the one hand, floral lists are widely distributed on the Internet, accessible to users and do not require special software tools and skills to read them. On the other hand, as a rule, such lists of species differ in structure and content and can be quite large, which makes it extremely difficult to process and analyze them.

The proposed method can be useful for analysts and developers of software tools designed to analyze categorical data, i.e. data represented by sets of names of various objects, including floral lists of plant species.

5. Acknowledgments

This work was partially supported by the Far East fundamental research program of the Feb RAS (project no. 18-5-104).

References

[1] Ipatov V S 1998 Opisanie fitocenoza: metodicheskie rekomendacii (SPb.)
[2] Shlotgauer S D, Kryukova M V and Antonova L A 2001 Vascular plants of the Khabarovsk territory and their protection (Vladivostok–Khabarovsk: Feb RAS)
[3] Yurcev B A, Kamelin R V 1991 Osnovnye ponyatiya i terminy floristiki (Perm’)
[4] Shitikov V K, Rozenberg G S 2005 Ocenkabioraznoobraziya: Popytka formal’nogo obobshcheniya Kolichestvennye metody ekologii i gidrobiologii (sbornik nauchnyh trudov, posvyashchennyj pamяти A.I. Bakanova) Otv. red. chl.–kor. RAN Rozenberg G S (To’yatti: SamNC RAN)
[5] Pecenko A 1982 Principles and methods of quantitative analysis in faunal research (Moscow: Nauka)
[6] Global Biodiversity Information Facility GBIF (https://www.gbif.org – 01.06.2020)
[7] Ipatov V S 2015 Using projective coverage to assess the role of species in the spatial structure of communities Bot.Shurn. 5 508–12.
[8] Bulohov A D 2004 Fitoindikaciya i ee prakticheskoie primenenie (Bryansk: BGU)
[9] Natural England. List of habitats (http://www.naturalengland.org.uk– 01.06.2020)
[10] Chernov Yu I 2008 Ecology and Biogeography.Selected works (Moscow: KMK Scientific Press Ltd.)
[11] Yang Y, Guan X and You J 2002 CLOPE: a fast & effective clustering algorithm for transactional data Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery & data mining pp 682–87

[12] Alimov A F 2006 The Role of Biological Diversity in Ecosystems Bulletin of the Russian Academy of Sciences 76 (11) 989–94