20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Graph Models of Network Behavior in Environmental Planning

Gianni Fenu[a,*], Pier Luigi Pau[a]

[a]*Department of Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Italy*

Abstract

Policies to protect the environment in Europe and in the rest of the world have been adjusted to take into account the network behavior of conglomerates of nature protection areas. Network behavior can emerge from the natural configuration of habitat patches, or be induced by the establishment of habitat corridors. Careful planning is required to protect and improve the network behavior in existing sites; this has prompted researchers to build graph models of ecological networks, and apply complex network analysis to improve the understanding of their features. However, the most common approach is to keep the focus on a single species, meant to be representative of most species within the area under analysis, or especially important with respect to conservation issues. In this paper, data pertaining to land use types found within sites making up the "Natura 2000" ecological network is used to provide a high-level view of the network, and propose a framework for study, in which similarity measures are used as a criterion to suggest guidelines for land management.

Keywords: Complex networks; Ecological networks; Knowledge discovery; Similarity

1. Introduction

As society evolves and shows an increasing phenomenon of urbanization, activities related with the protection and safekeeping of endangered habitats and species have undergone changes, as well. The establishment of nature protection areas with limited goals has proven to be inadequate in many situations, and policies have been shifted toward a consideration of large scale, wide area goals, with the introduction of the concept of ecological networks[1]. This has repercussions on land management and planning, which are now as challenging as ever. In order to improve the understanding of network behavior in nature protection areas within a geographical region, complex network analysis techniques are being employed in this field. While this has proven useful, turning analysis results into guidelines for land management is still perceived as a considerable challenge.

In this paper, current practices in the analysis of ecological networks are revised, and an approach to extend current methods to take into account the goals of land management is proposed. In Section 2, further details are given regarding ecological networks as a concept. Section 3 provides an introduction to graphs as a model for ecological networks. In Section 4, basic information on Natura 2000 is provided and methods to build graph models from data on

* Corresponding author. Tel.: +39-070-675-8759; fax: +39-070-675-8779.
  E-mail address: fenu@unica.it

Natura 2000 sites are discussed. A test case where Sardinian sites are modeled as a network is the topic of Section 5, with a revision of the process of graph building and some observations on the properties of constructed graphs. Lastly, in Section 6, conclusions are drawn and possibilities for future work are discussed.

2. Basics of Ecological Networks

The sustainment and expansion of human activities of every kind have increasingly modified land use, resulting in a loss of habitats and the fragmentation of those still in existence. Nature preservation areas have been created with the goal of protecting endangered habitats and species, but on occasion, this approach has met with limited success, since the effectiveness of preservation areas is negatively affected by insufficient size or excessive distance from similar habitats. To address this problem, the current mindset involves the creation of nature protection areas following a reticular model[2], in which each area is supposed to provide a contribution to larger preservation goals with respect to nature and biodiversity, rather than being setup only with a specific goal in mind. Ecological networks are the intended result of this change in policies concerning nature protection areas.

Network behavior can emerge from spontaneous migration or dispersal of species, or be induced by means of land planning. Most commonly, this happens with the establishment of man-made 'habitat corridors' (or 'green corridors'), defined as linear strips of habitat connecting two or more larger patches of habitat, surrounded by a dissimilar matrix[3]. Alternatively, corridors may be implemented as a set of disconnected patches, referred to as 'stepping stones'. While habitat corridors haven't been widely accepted yet as a solution to biodiversity issues, their implementation has resulted in a good effect for biodiversity at least in the short-term for specific cases[4].

From the point of view of land management and planning, a common problem is that the actual implementation of corridors is to be adapted to the needs of the target species[5]. The number of species in existing habitat patches, however, may be very large, even when only protected species are accounted for. Designing a man-made habitat corridor to be suitable for a large number of species is a considerable challenge. Since the same is true for preservation areas, however, it is possible to borrow approaches that have proven to be effective in the management of areas. A recurring concept is that of identifying an 'umbrella species', defined as a 'species with large area requirements, which if given sufficient protected habitat area, will bring many other species under protection'[6]. Essentially, a single species is chosen such that, by implementing measures to be beneficial to that species, the end result is an improvement for a wide number of other species as well. This concept has proven successful in many situations, particularly when dealing with strict time constraints in the activities of analysis and conservation planning[7]; however, its limitations ought to be considered carefully.

3. Graph Models for Ecological Networks

Models for ecological networks can be built using two fundamentally different approaches. One is a structural approach, based on a representation of the elements that make up the network. Essentially, this corresponds to answering the question of what the ecological network is like. The other possibility is a functional approach, which keeps the focus on describing how the ecological network actually works.

Graphs as mathematical objects are fit to represent networks of many different types, being made up of two kinds of basic elements, i.e. nodes and edges, which represent entities and links, respectively. Complex systems are often modeled as weighted graphs, i.e. graphs in which a different weight is assigned to each edge, to represent either a distance or a strength of the link.

To model an ecological network as a graph, one can use a node to represent each site or habitat patch, depending on the desired scale; then, edges will represent existing or potential connections between nodes. Theoretically, both approaches described above are possible, i.e. edges can be used to model structural or functional connectivity. If a structural approach is chosen, the presence of edges and their weights should reflect the presence of physical features, which may influence transit between sites in a positive or negative way. If a functional approach is chosen, edges should represent actual or potential flows of species migration. Since these can vary substantially for different species, the study of functional connectivity is generally performed for a chosen target species. In fact, since it has proven difficult to find meaningful variables to assess edge weights based on structural connectivity, this is considered the

most successful approach[8]. Instead, the description and analysis of ecological networks from a structural point of view is performed with Geographic Information System tools, such as ArcGIS and QGIS.

Once a proper model has been established, the graph representing an ecological network can be analyzed to extrapolate advanced network properties, which may not be simply derived from its geographical map. The first and most important observation is whether the graph is connected (i.e. a path through existing edges allows movement between any two nodes), or made up of several connected components. Then, several more measures of connectivity become possible to extract. For instance, network elements can be ranked by importance taking reference from node degree, i.e. the number of nodes adjacent to a given one, and betweenness centrality, which expresses a proportion of shortest paths including a node or an edge. If nodes with a high betweenness centrality index are removed from a network to represent a loss of their functionality, a larger number of shortest paths is affected, and this can have an impact on network connectivity[9]. This is particularly true when the large scale effect is considered; the end result may be the fragmentation of the network into several isolated components, while the connectivity of local agglomerates of patches may be unaffected, if they are strongly organized in cliques[10].

4. Application to Natura 2000 Sites

In recent times, the European Union has extended its nature and biodiversity policies towards the establishment of a network of nature protection areas, denominated "Natura 2000". The elements of this network are nature protection sites designated as Special Protection Areas (SPA), in accordance with the EU Birds Directive (2009/147/EC), and Special Areas of Conservation (SAC), in accordance with the EU Habitats Directive (92/43/EEC). The latter are designated in a two-step process which involves a preliminary designation as Site of Community Interest (SCI). A site can be designated both as a SPA and a SAC (or SCI) at the same time, and it is possible for the boundaries of SPAs to overlap with those of SACs and SCIs. Sites designated as SPA do not overlap each other, but they may be adjacent; the same is true for SCIs and SACs. Compared to sites, linking elements are not as thoroughly regulated, as network behavior is not necessarily the result of the establishment of man-made corridors.

From the point of view of nature conservation, land planning involves the identification of threats to the continued presence of protected species and habitats, and the proposal of a course of action to address any standing issue. Land planning requires careful consideration of technical, regulatory, and political aspects. It is desirable that software-aided land planning tools be designed to provide assistance in evaluating technical aspects at different scales, and the work toward graph models that properly represent functional connectivity of ecological networks is part of an endeavor to take steps in that direction. In the remainder of the document, alternatives for graph models are discussed and proposed, while keeping in mind the need to fulfill two main goals: that of describing network behavior based on known data, and that of suggesting lines of development for future land planning.

4.1. Reference Model Basics

Graph models of the ecological network made up of Natura 2000 sites in a geographical area can be built for comparison and analysis. Available data includes reports of species presence, geographical data (site perimeters and coordinates of site centroids), and data on land use from the CORINE (Coordination of Information on the Environment) program managed within the European Environment Agency. Data on the presence of species was extracted from reports on Natura 2000 sites filed by on-site researchers, in accordance with the Standard Data Form released with Commission Implementing Decision 2011/484/EU, as part of activities periodically scheduled for Natura 2000 sites. Only the species of interest listed in Annex II to Directive 92/43/EEC are considered, as reports on other species tend to be fragmented and not standardized.

First, a list of nodes is determined. Each site will be represented by a node in the graph model. When any site is designated as a SPA and a SCI at the same time, only one node is created. If the boundaries of a SPA and a SCI overlap, two nodes are still created, but they will be considered to be at zero distance from each other at the linking stage.

Secondly, criteria to draw edges between pairs of nodes are to be considered. These criteria can be based on:

- Geographical distance below a definite threshold, which in turn can be calculated between boundaries or between centroids;
- Presence of a target species in both sites under consideration;
- Similarity between sites above a threshold value, based on vectors representing land use;
- Similarity between sites above a threshold value, based on vectors representing species presence.

Criteria can be combined to obtain a more meaningful model. In literature, a common choice is to link pair of sites where a target species is present and the distance between them is within the dispersal distance of the target species. Distance between sites can be calculated by importing perimeter and centroid data into a spatial database, such as SQLite with the Spatialite extension. Current spatial databases will be able to compute distance between boundaries or between centroids, but in both cases, result values are to be treated as an approximation, since they are calculated on a map projection.

4.2. Estimation of Similarity

A similarity value between sites is calculated by representing site attributes as vectors, and using common techniques for vector comparison, such as the Jaccard Coefficient for binary vectors, and cosine distance for the general case. In this study, vectors for each site are built in different ways:

- using the presence or absence of each Annex II species in the target site as a binary attribute;
- using the number of occurrences of land use types in the target site as an integer attribute.

Once vectors are built, the Jaccard Coefficient of a pair of vectors can be used as a similarity value. It is defined as:

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}},$$ (1)

where f_{11} is the number of attributes that is set to *true* in both vectors, and f_{01} and f_{10} are the counts of attributes that are set to *true* only in the second and first vector being considered, respectively. Essentially, when calculating Jaccard coefficients from data on species presence, f_{11} is the number of species that is present on both sites in a pair, and the sum of three terms in the denominator is the total number of species found in at least one of the sites under consideration. The Jaccard Coefficient takes a value from 0 to 1, where 1 is the maximum similarity.

To estimate similarity between sites with respect to land use, additional data available separately from the CORINE program is used. Different land use types correspond to different CORINE Land Cover codes, categorized in five levels of detail. A CORINE Land Cover code consists of up to five digits, where each digits adds a level of detail; for instance, all codes where the first digit is a 2 correspond to agricultural areas, and all codes beginning with a 3 correspond to forests and semi-natural areas. Adding a level of degree to the latter, codes beginning with 3.1 identify forests, and codes beginning with 3.2 identify areas with herbaceous vegetation; further digits are used to identify different kinds of forest in the 3.1.x branch, and likewise for other code branches.

The open source QGIS[11] software was used to compute the overlapping of Natura 2000 site boundaries with data on land use from the CORINE program. Since data on the fourth and fifth level of details were not always available, only the first three levels were used. The vector for each site was built by counting the number of land patches corresponding to each 3-digit code and intersecting the site, regardless of whether the land patch is completely or only partly within the site boundaries. The similarity between pairs of vectors can be calculated using the formula for cosine similarity:

$$cos(x, y) = \frac{x \cdot y}{\|x\| \, \|y\|},$$ (2)

where x and y are the vector representing a pair of sites, $x \cdot y$ is their scalar product, and $\|v\|$ is the magnitude of a vector v. Cosine distance also takes values between 0 and 1, where a higher value means higher similarity. This measure of similarity is common for the comparison of text documents based on the occurrence count of keywords, which makes part of document categorization. Its success in document comparison is determined by the fact that it is independent
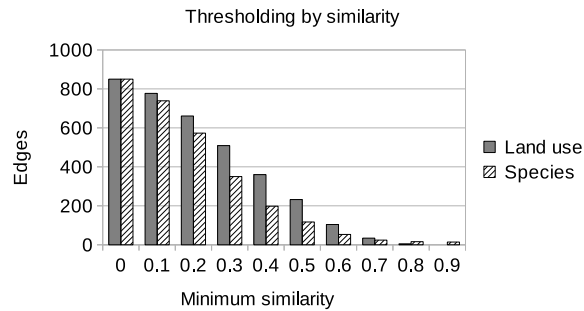
Fig. 1. Comparison of the number of edges in the graph representation of the ecological network, after thresholding by similarity of land use and species presence, calculated using Jaccard coefficients.

of document length. In the context of this study, this is essentially equivalent to independence from the geographical extension of sites.

To allow for more meaningful comparison with the similarity of species presence, Jaccard coefficients can be calculated for pairs of vectors representing land use, by treating each non-zero value as a 1.

5. Analysis Setup and Results

To act as a benchmark case for this study, Natura 2000 sites in Sardinia (one of the two main Italian islands) are used for building reference models as described in the previous section. At the time of this writing, a total of 124 sites are recognized, including SPA and SCI; no site in Sardinia has been designated as a SAC. Since land use data is currently unavailable for 7 sites, only 117 sites shall be considered for building graphs based on similarity scores. Reference graphs are built for a maximum distance between sites of 30 Km, but since all distances are estimated, the actual distance may be larger in some cases. With this setup, the resulting graph comprises 117 nodes and 850 edges (Figure 2a).

5.1. Thresholding by similarity

As the analysis of graph models is supposed to contribute as a factor in land managers' decisions, including that of establishing habitat corridors, it can be sensible to remove edges where a similarity value between the corresponding node pair is below a set threshold.

Table 1 reports the way the number of edges decreases as a threshold is applied, based on three different similarity scores: Jaccard coefficients based on species presence and land use, and cosine similarity of land use. Figure 1 shows results for Jaccard coefficients in a graphical manner.

Table 1. Number of edges after thresholding by minimum similarity.

Threshold	Jaccard Coefficient Species-based	Jaccard Coefficient Land use-based	Cosine similarity Land use-based
0.0	850	850	850
0.1	739	777	766
0.2	573	661	679
0.3	350	509	556
0.4	198	360	437
0.5	117	232	338
0.6	53	104	243
0.7	24	34	153
0.8	16	5	95
0.9	14	0	41

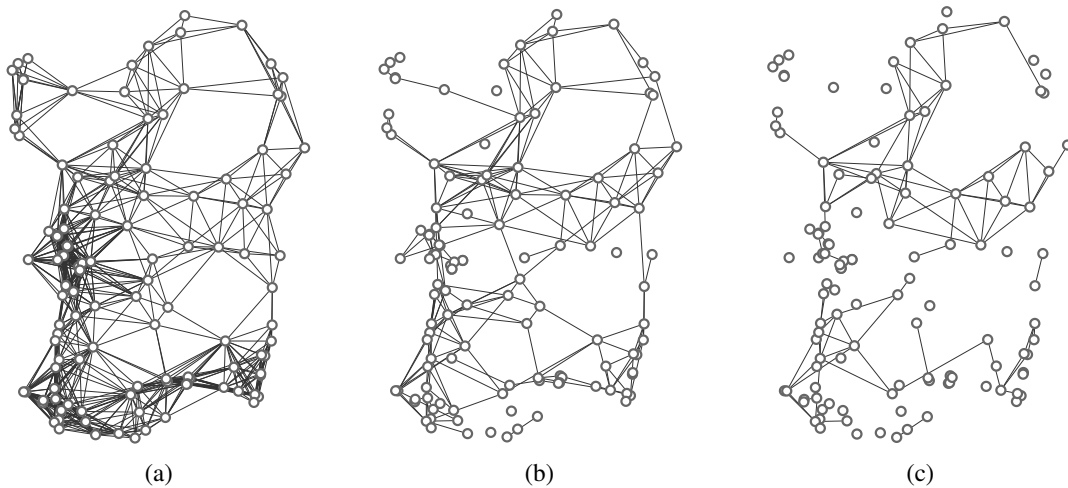(a)　　　　　　　　　　　(b)　　　　　　　　　　　(c)

Fig. 2. Sample networks built from Sardinian Natura 2000 sites. Nodes within a maximum distance of 30 Km between boundaries are linked by an edge. The position of each node corresponds to the coordinates of the site centroid. (a) Ignoring similarity between sites. (b) Following a comparison of pairs of sites by land use, edges are removed for Jaccard coefficients below 0.5. (c) Likewise, with edges removed for Jaccard coefficients below 0.6.

It is notable that, if similarity of species presence is used, when the minimum similarity to keep an edge is increased, the number of edges decreases sharply for low thresholds; at 0.3 threshold, more than half of the edges have been cut. If Jaccard similarity of land use is chosen, the observed decrease in the number of edges is not as steep for low thresholds; a drop to less than half of the original edges is observed only at 0.4. However, the number of edges does drop to zero over the 0.8 mark. Cosine similarity scores are reported mainly for comparison, to support the notion that Jaccard coefficients require a lower threshold to cut similar amounts of edges from the network. The remainder of this study shall focus on Jaccard coefficients.

The comparison of graphs built with different similarity threshold is also useful to infer differences between geographical areas. Figure 2b provides a view of the network built from Sardinian Natura 2000 sites as discussed above, where a 0.5 threshold for Jaccard similarity of land use is applied. The largest connected component comprises a large majority of nodes, with only a few isolated nodes and node pairs making up own components. The graph in Figure 2c was built using the same method, but using a threshold of 0.6 for Jaccard similarity. In this instance, aside of the increased number of isolated nodes, two large connected components can be identified, which roughly correspond to a subdivision into Northern and Southern Sardinia. This observation can be used to support the notion that morphological differences exist between the two large areas.

Moving on, if the similarity threshold is further increased to 0.7, the number of edges is reduced to about one third of those in the network built with a threshold of 0.6. It can be argued that a threshold of 0.7 is a strong requirement for the linking of nodes.

5.2. Graph models for single species

In section 4.1, it was mentioned that the presence of a single species can be used as a criterion to link nodes. This essentially serves the purpose of contributing to describe the current status of the ecological network with respect to the chosen species. Networks constructed with this method are often analyzed using complex network techniques to extrapolate their advanced properties, evaluate their degree of connectivity, and identify their most central nodes, for example by using the betweenness centrality index as reference[12]. This index corresponds to the proportion of
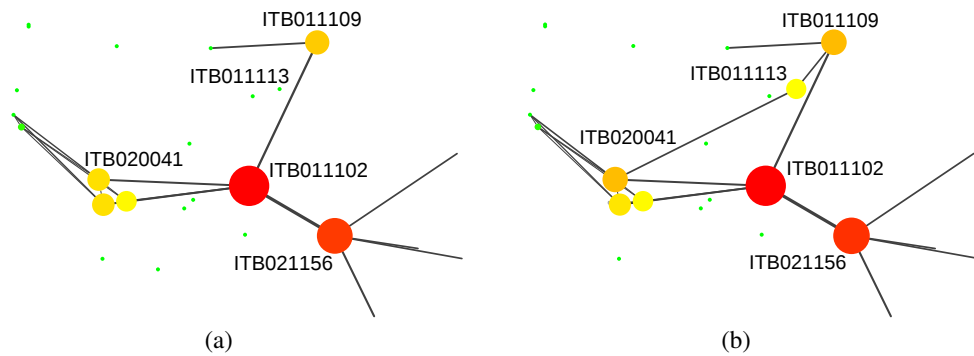
Fig. 3. Example of a single-species network for *Rhinolophus hipposideros*. Administrative codes of some relevant sites are reported. A higher value of the betweenness centrality index of each node is represented by a larger size of the node. (a) Generated network. (b) A proposed modification, which provides alternative paths and slightly lowers the betweenness centrality index of the most central node in the previous network.

shortest paths that include a network element (typically a node, but betweenness centrality can be calculated for edges, as well). More precisely, it is expressed for nodes by

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{3}$$

where V is the set of nodes (when a graph is made up of connected components, each component should be considered separately), σ_{st} is the number of shortest paths from node s to node t, and $\sigma_{st}(v)$ is the number of said shortest paths that include v, the node for which the index is calculated. It is often useful to normalize this index to a scale of 0 to 1, dividing it by the number of node pairs, expressed by $(N-1)(N-2)/2$, where N is the number of nodes.

Central nodes can be associated with vulnerabilities of the network, i.e. those areas where a loss of habitat patches can have the greatest adverse impact on network behavior at a large scale. The activity of identifying these areas is referred to as 'vulnerability assessment' and it is a common goal for which complex network analysis is performed, in those fields of application where different kinds of infrastructure are modeled as graphs (e.g. power grids, computer networks). Ecological networks can be considered part of this category. Moreover, the comparison with similarity-based graphs can provide guidelines for the proposal of ways to reduce the impact of vulnerabilities.

To provide an example, a portion of the network built for *Rhinolophus hipposideros* is shown in Figure 3a. The open source tool Cytoscape [13] version 3.3.0 was used to perform network analysis and visualization. The highest value for the betweenness centrality index in this portion of the network is found on the node labeled ITB011102, which means that the removal of this node (or the edges linked to it) can be expected to have a very adverse effect on the network, in terms of increase of shortest path lengths or even the disconnection of network components. If the goal of limiting this impact were to be considered, the establishment of habitat corridors could be proposed, evaluating candidates by land use similarity scores, and with the help of similarity graphs described in Subsection 5.1. In this example, the use of node ITB011113 as an alternative path between ITB020041 and ITB011109 (see Figure 3b) could be supported by the argument that the involved node pairs have a similarity score above 0.7. Analysis of the modified version of the network reveals a slight decrease in betweenness centrality for node ITB011102 (down to 0.614 from the value of 0.667 in the unmodified network), which is in support of the usefulness of the proposed modification, although it stands to reason that the centrality of node ITB021156 would have to be addressed at the same time for the effect to be relevant. Together with observations from different studies, this consideration can be useful for land managers to evaluate possible approaches to reducing vulnerabilities and enhancing network behavior in specific areas.

6. Conclusions and Future Work

The analysis of ecological networks is a growing topic which holds many challenges and opportunities for research. Complex network analysis techniques have often been used to improve the understanding of network behavior in

nature protection areas within certain regions, but current methods tend to focus on narrow cases, even limited to one species, which can result in a need to represent a network with multiple graphs, making it difficult to understand high-level properties of the network as a whole. In this paper, to work toward that goal, the construction of networks using site similarity as a criterion to link nodes is discussed. This can also serve the purpose of identifying interesting cases among graphs which keep their focus on a single species, or as a complement to the activity of vulnerability assessment.

In future work, more graphs built from data on single species will be analyzed, with a focus on finding methods to uncover cases where a comparison with similarity-based graphs is significant. Additionally, there will be more work on assessing properties of networks using graphs built from cosine similarity of land use.

Acknowledgements

References

1. Vimal, R., Mathevet, R., Thompson, J.D.. The changing landscape of ecological networks. *Journal for Nature Conservation* 2012; **20**(1):49–55. URL: `http://www.sciencedirect.com/science/article/pii/S1617138111000471`. doi:10.1016/j.jnc.2011.08.001.
2. Jongman, R.H.G.. Nature conservation planning in Europe: developing ecological networks. *Landscape and Urban Planning* 1995; **32**(3):169–183. URL: `http://www.sciencedirect.com/science/article/pii/0169204695001970`. doi:10.1016/0169-2046(95)00197-O.
3. Beier, P., Noss, R.F.. Do Habitat Corridors Provide Connectivity? *Conservation Biology* 1998; **12**(6):1241–1252. URL: `http://onlinelibrary.wiley.com/doi/10.1111/j.1523-1739.1998.98036.x/abstract`. doi:10.1111/j.1523-1739.1998.98036.x.
4. Gilbert-Norton, L., Wilson, R., Stevens, J.R., Beard, K.H.. A meta-analytic review of corridor effectiveness. *Conservation Biology: The Journal of the Society for Conservation Biology* 2010; **24**(3):660–668. doi:10.1111/j.1523-1739.2010.01450.x.
5. Fleury, A.M., Brown, R.D.. A framework for the design of wildlife conservation corridors With specific application to southwestern Ontario. *Landscape and Urban Planning* 1997; **37**(3–4):163–186. URL: `http://www.sciencedirect.com/science/article/pii/S0169204697800023`. doi:10.1016/S0169-2046(97)80002-3.
6. Caro, T.M.. Umbrella species: critique and lessons from East Africa. *Animal Conservation* 2003; **6**(2):171–181. URL: `http://onlinelibrary.wiley.com/doi/10.1017/S1367943003003214/abstract`. doi:10.1017/S1367943003003214.
7. Roberge, J.M., Angelstam, P.. Usefulness of the Umbrella Species Concept as a Conservation Tool. *Conservation Biology* 2004; **18**(1):76–85. URL: `http://onlinelibrary.wiley.com/doi/10.1111/j.1523-1739.2004.00450.x/abstract`. doi:10.1111/j.1523-1739.2004.00450.x.
8. Urban, D.L., Minor, E.S., Treml, E.A., Schick, R.S.. Graph models of habitat mosaics. *Ecology Letters* 2009; **12**(3):260–273. URL: `http://onlinelibrary.wiley.com/doi/10.1111/j.1461-0248.2008.01271.x/abstract`. doi:10.1111/j.1461-0248.2008.01271.x.
9. Bodin, Ö., Norberg, J.. A Network Approach for Analyzing Spatially Structured Populations in Fragmented Landscape. *Landscape Ecology* 2006; **22**(1):31–44. URL: `http://link.springer.com/article/10.1007/s10980-006-9015-0`. doi:10.1007/s10980-006-9015-0.
10. Estrada, E., Bodin, Ö.. Using Network Centrality Measures to Manage Landscape Connectivity. *Ecological Applications* 2008; **18**(7):1810–1825. URL: `http://onlinelibrary.wiley.com/doi/10.1890/07-1419.1/abstract`. doi:10.1890/07-1419.1.
11. QGIS Development Team, . *QGIS Geographic Information System*. Open Source Geospatial Foundation; 2009. URL: `http://qgis.osgeo.org`.
12. Mazaris, A.D., Papanikolaou, A.D., Barbet-Massin, M., Kallimanis, A.S., Jiguet, F., Schmeller, D.S., et al. Evaluating the Connectivity of a Protected Areas' Network under the Prism of Global Change: The Efficiency of the European Natura 2000 Network for Four Birds of Prey. *PLoS ONE* 2013; **8**(3):e59640. URL: `http://dx.doi.org/10.1371/journal.pone.0059640`. doi:10.1371/journal.pone.0059640.
13. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 2003; **13**(11):2498–2504. doi:10.1101/gr.1239303.