

mangal – making complex ecological network analysis simpler

T. Poisot, B. Baiser, J. Dunne, S. Kéfi, F. Massol, N. Mouquet, T. Romanuk, D. Stouffer, S. Wood, D. Gravel

Jan. 2014

1 The study of ecological networks is severely limited by (i) the difficulty to access data, (ii) the lack of a
2 standardized way to link meta-data with interactions, and (iii) the disparity of formats in which ecologi-
3 cal networks themselves are represented. To overcome these limitations, we conceived a data specification
4 for ecological networks. We implemented a database respecting this standard, and released a R package (
5 `rmangal`) allowing users to programmatically access, curate, and deposit data on ecological interactions. In
6 this article, we show how these tools, in conjunctions with other frameworks for the programmatic manipu-
7 lation of open ecological data, streamlines the analysis process, and improves replicability and reproducibility
8 of ecological networks studies.

9 Introduction

10 Ecological networks enable ecologists to accommodate the complexity of natural communities, and to discover mech-
11 anisms contributing to their persistence, stability, resilience, and functioning. Most of the “early” studies of ecological
12 networks were focused on understanding how the structure of interactions within one location affected the ecological
13 properties of this local community. Such analyses revealed the contribution of ‘average’ network properties, such as the
14 buffering impact of modularity on species loss (Pimm *et al.* 1991,), the increase in robustness to extinctions along with
15 increases in connectance (Dunne *et al.* 2002), and the fact that organization of interactions maximizes biodiversity (Bas-
16 tolla *et al.* 2009). More recently, new studies introduced the idea that networks can vary from one realization to another.
17 They can be meaningfully compared, either to understand the importance of environmental gradients on the realization of
18 ecological interactions [atylanakis_habitat_2007], or to understand the mechanisms behind variation in the structure of
19 ecological networks (Poisot *et al.* 2012). Yet, meta-analyses of a large number of ecological networks are still extremely
20 rare, and most of the studies comparing several networks do so within the limit of particular systems (Schleuning *et al.*
21 2011; Dalsgaard *et al.* 2013). The severe shortage of data in the field also restricts the scope of large-scale analyses.

22 An increasing number of approaches are being put forth to *predict* the structure of ecological networks, either relying on
23 latent variables (Rohr *et al.* 2010) or actual traits (Gravel *et al.* 2013). Such approaches, so as to be adequately calibrated,

1 require easily accessible data. Comparing the efficiency of different methods is also facilitated if there is an homogeneous
2 way of representing ecological interactions, and the associated metadata. In this paper, we (i) establish the need of a data
3 specification serving as a *lingua franca* among network ecologists, (ii) describe this data specification, and (iii) describe
4 `rmangal`, a R package and companion database relying on this data specification. The `rmangal` package allows to easily
5 retrieve, but also deposit, ecological interaction networks data from a database. We provide some use cases showing how
6 this new approach makes complex analyzes simpler, and allows for the integration of new tools to manipulate biodiversity
7 resources.

8 **Networks need a data specification**

9 Ecological networks are (often) stored as an *adjacency matrix* (or as the quantitative link matrix), that is a series of 0 and 1
10 indicating, respectively, the absence and presence of an interaction. This format is extremely convenient for *use* (as most
11 network analysis packages, *e.g.* `bipartite`, `betalink`, `foodweb`, require data to be presented this way), but is extremely
12 inefficient at *storing* meta-data. In most cases, an adjacency matrix informs on the identity of species (in cases where
13 rows and columns headers are present), and the presence or absence of interactions. If other data about the environment
14 (*e.g.* where the network was sampled) or the species (*e.g.* the population size, trait distribution, or other observations) are
15 available, they are most either given in other files, or as accompanying text. In both cases, making a programmatic link
16 between interaction data and relevant meta-data is difficult and error-prone.

17 By contrast, a data specification (*i.e.* a set of precise instructions detailing how each object should be represented) provides
18 a common language for network ecologists to interact, and ensure that, regardless of their source, data can be used in a
19 shared workflow. Most importantly, a data specification describes how data are *exchanged*. Each group retains the ability
20 to store the data in the format that is most convenient for in-house use, and only needs to provide export options (*e.g.*
21 through an API, *i.e.* a programmatic interface running on a webserver, returning data in response to queries in a pre-
22 determined language) respecting the data specification. This approach ensures that *all* data can be used in meta-analyses,
23 and increases the impact of data (Piwowar *et al.* 2007; Piwowar & Vision 2013).

24 **Elements of the data specification**

25 The data specification (Fig. 1) is built around the idea that (ecological) networks are collections of relationships between
26 ecological objects, each element having particular meta-data associated. In this section, we detail the way networks
27 are represented in the `mangal` specification. An interactive webpage with the elements of the data specification can be
28 found online at <http://mangal.uqar.ca/doc/spec/>. The data specification is available either at the API root (*e.g.*

1 <http://mangal.uqar.ca/api/v1/?format=json>), or can be viewed using the `whatIs` function from the R package
2 (see *Supp. Mat. 1*). Rather than giving an exhaustive list of the data specification (which is available online at the
3 aforementioned URL), this section serves as an overview of each element, and how they interact.

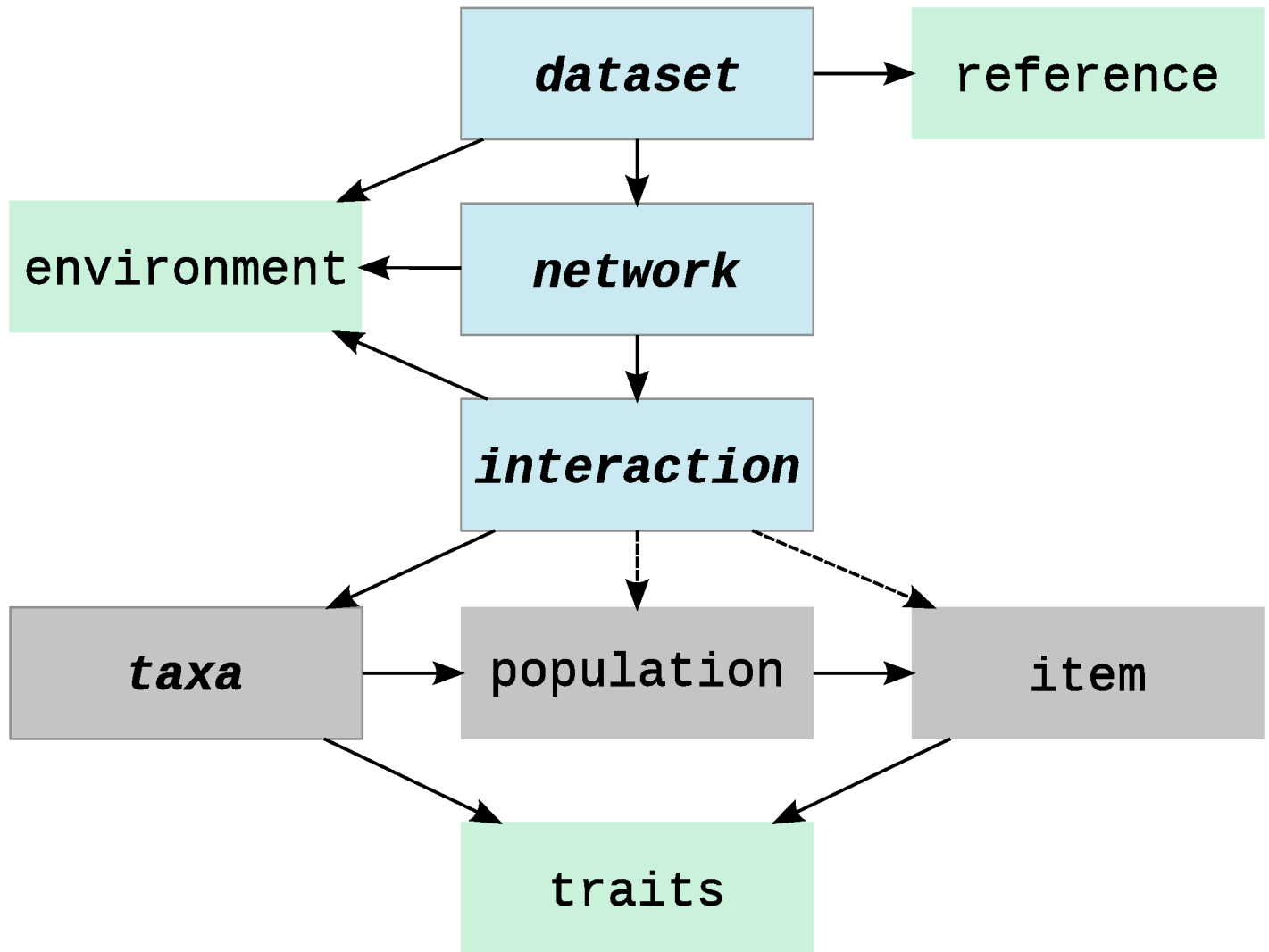


Figure 1: An overview of the data specification, and the hierarchy between objects. Each box correspond to a level of the data specification. Grey boxes are nodes, blue boxes are interactions and networks, and green boxes are metadata. The **bold** boxes (*dataset*, *network*, *interaction*, *taxa*) are the minimal elements needed to represent a network.

4 We propose JSON, a format equivalent to XML, as an efficient way to uniformise data representation for two main reasons.
5 First, it has emerged as a *de facto* standard for web platform serving data, and accepting data from users. Second, it allows
6 *validation* of the data: a JSON file can be matched against a scheme, and one can verify that it is correctly formatted (this
7 includes the possibility that not all fields are filled, as will depend on available data). Finally, JSON objects are easily
8 and cheaply (memory-wise) parsed in the most common programming languages, notably R (equivalent to `list`) and
9 python (equivalent to `dict`). For most users, the format in which data are transmitted is unimportant, as the interaction
10 happens within R – as such, knowing how JSON objects are organized is only useful for those who want to interact with
11 the API directly. The `rmangal` package takes care of converting the data into the correct JSON format to upload them in

1 the database.

2 **Node information**

3 **Taxa**

4 Taxa are a taxonomic entity of any level, identified by their name, vernacular name, and their identifiers in a variety of
5 taxonomic services. Associating the identifiers of each taxa is important to leverage the power of the new generation
6 of open data tools, such as `taxize` [chamberlain_taxize:_2013]. The data specification currently has fields for `ncbi`,
7 `gbif`, `itis`, `eol` and `bold` identifiers. We also provide the taxonomic status, *i.e.* whether a taxa is a true taxonomic entity,
8 a “trophic species”, or a morphospecies.

9 **Population**

10 A population is one observed instance of a taxa object. If your experimental design is replicated through space, then
11 each taxa have a population object corresponding to each locality. Populations do not have associated meta-data, but
12 serve as “containers” for item objects.

13 **Item**

14 An item is an instance of a population. Items have a `level` argument, which can be either `individual` or `population`;
15 this allows to represent both individual-level networks (*i.e.* there are as many items attached to a population than there
16 were individuals of this population sampled), and population-level networks. When item represents a population, it
17 is possible to give a measure of the size of this population. The notion of item is particularly useful for time-replicated
18 designs: each observation of a population at a time-point is an item with associated `trait` values, and possibly population
19 size.

20 **Network information**

21 **Interaction**

22 An interaction links, *a minima*, two taxa objects (but can also link pairs of populations or items). The most
23 important attributes of interactions are the type of interaction (of which we provide a list of possible values, see *Supp.*
24 *Mat. 1*), and its nature, *i.e.* how it was observed. This field help differentiate direct observations, text mining, and

1 inference. Note that the nature field can also take absence as a value; this is useful for, *e.g.*, “cafeteria” experiments in
2 which there is high confidence that the interaction did not happen.

3 **Network**

4 A network is a series of interaction object, along with (i) informations on its spatial position (provided at the latitude
5 and longitude), (ii) the date of sampling, and (iii) references to measures of environmental conditions.

6 **Dataset**

7 A dataset is a collection of one or several network(s). Datasets also have a field for data and papers, both of which
8 are references to bibliographic or web resources describing, respectively, the source of the data, and the papers in which
9 these data have been significantly used. Datasets are the preferred entry point in the resources.

10 **Meta-data**

11 **Trait value**

12 Objects of type `item` can have associated trait values. These consist in the description of the trait being measured, the
13 value, and the units in which the measure was taken.

14 **Environmental condition**

15 Environmental conditions are associated to datasets, networks, and interactions objects, to allow for both macro and micro
16 environmental conditions. These are defined by the environmental property measured, its value, and the units.

17 **References**

18 References are associated to datasets. They accommodate the DOI, JSON or PubMed identifiers, or a URL. When
19 possible, the DOI should be preferred as it offers more potential to interact with other on-line tools, such as the *CrossRef*
20 API.

1 Use cases

2 In this section, we present use cases using the `rmangal` package for R, to interact with a database implementing this data
3 specification, and serving data through an API (<http://mangal.uqar.ca/api/v1/>). It is possible for users to deposit
4 data into this database, through the R package. Data are made available under a *CC-0 Waiver* (Poisot *et al.* 2013). Detailed
5 informations about how to upload data are given in the vignettes and manual of the `rmangal` package. So as to save room
6 in the manuscript, we source each example; the complete `r` files to reproduce the examples of this section are attached as
7 *Suppl. Mat.*. In addition, the `rmangal` package comes with vignettes explaining how users can upload their data into the
8 database, through R.

9 The data we use for this example come from Ricciardi *et al.* (2010). These were previously available on the *Interaction-*
10 *Web DataBase* as a single `xls` file. We uploaded them in the `mangal` database at <http://mangal.uqar.ca/api/v1/dataset/1>.

11 Link-species relationships

12 In the first example, we visualize the relationship between the number of species and the number of interactions, which
13 Martinez (1992) propose to be linear (in food webs).

```
source("usecases/1_ls.r")
```

14 Producing this figure requires less than 10 lines of code. The only information needed is the identifier of the network
15 or dataset, which we suggest should be reported in publications as: “These data were deposited in the `mangal` format
16 at `<URL>/api/v1/dataset/<ID>`”, possibly in the acknowledgements. So as to encourage data sharing, we encourage
17 users of the database to cite the original dataset or publication.

18 Network beta-diversity

19 In the second example, we use the framework of network β -diversity (Poisot *et al.* 2012) to measure the extent to which
20 networks that are far apart in space have different interactions. Each network in the dataset has a latitude and longitude,
21 meaning that it is possible to measure the geographic distance between two networks.

22 For each pair of network, we measure the geographic distance (in km.), the species dissimilarity (β_S), the network dissim-
23 ilarity when all species are present (β_{WN}), and finally, the network dissimilarity when only shared species are considered
24 (β_{OS}).

```
source("usecases/2_beta.r")
```

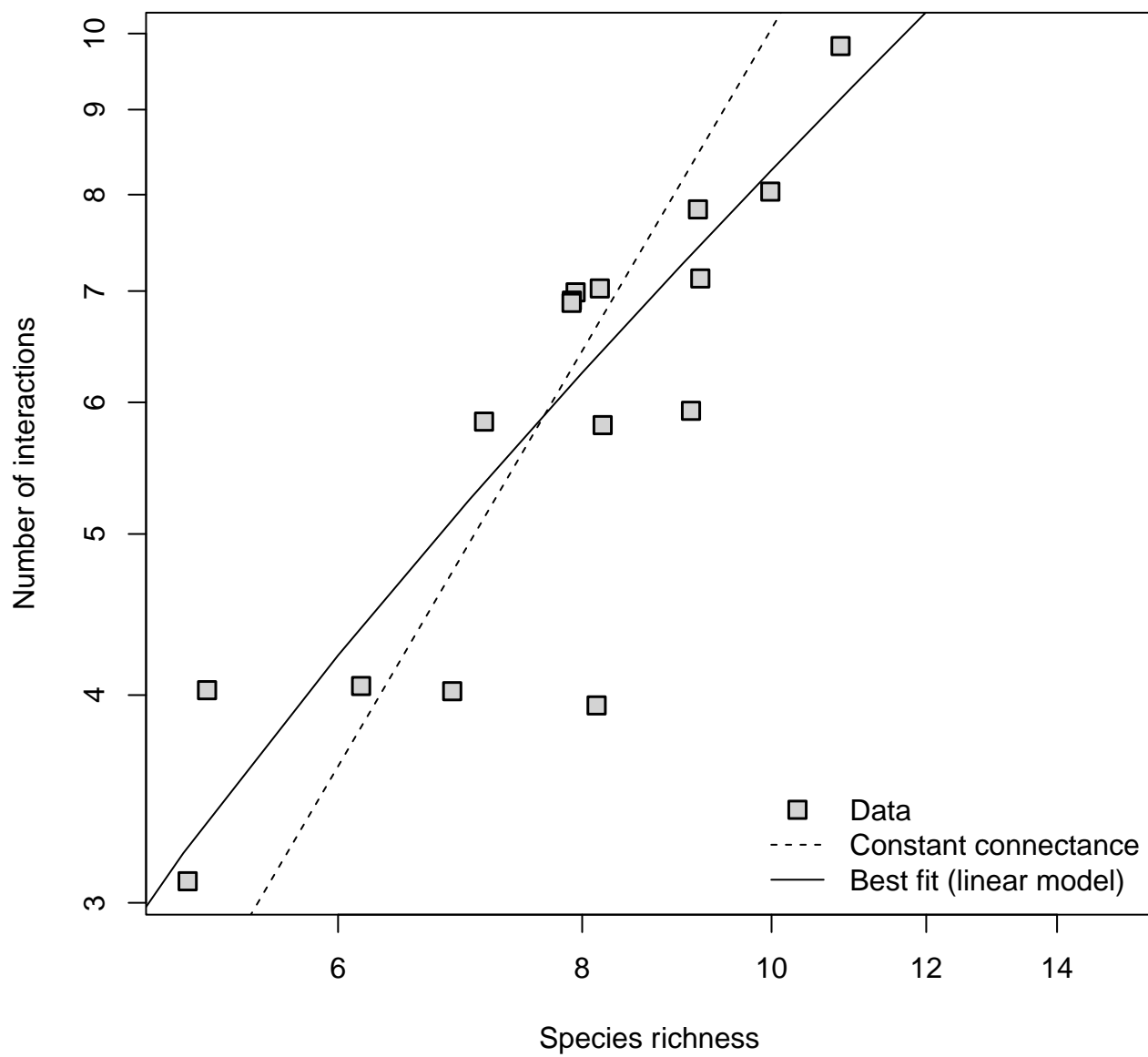


Figure 2: Relationship between the number of species and number of interactions in the anemonefish-fish dataset.

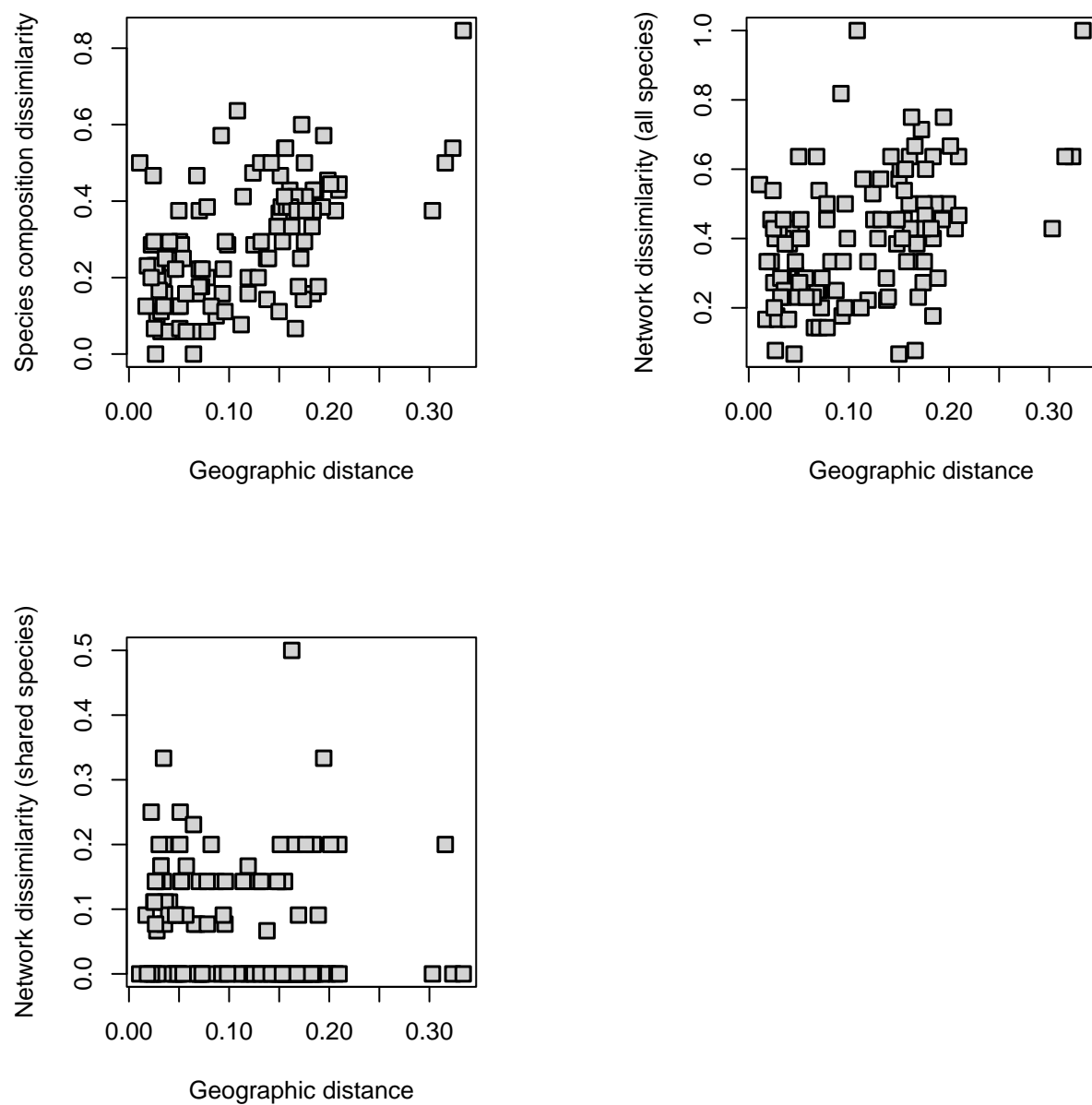


Figure 3: Relationships between the geographic distance between two sites, and the species dissimilarity, network dissimilarity with all, and only shared, species.

1 As shown in *Fig. XX*, while species dissimilarity and overall network dissimilarity increase when two networks are far
2 apart, this is not the case for the way common species interact. This suggests that in this system, network dissimilarity
3 over space is primarily driven by species turnover. The ease to gather both raw interaction data and associated meta-data
4 make producing this analysis extremely straightforward.

5 **Spatial visualization of networks**

6 Bascompte (2009) uses an interesting visualization for spatial networks, in which each species is laid out on a map at the
7 center of mass of its distribution; interactions are then drawn between species to show how species distribution determines
8 biotic interactions. In this final use case, we propose to reproduce a similar figure.

```
source("usecases/3_spatial.r")
```

9 **Conclusions**

10 In this contribution, we presented *mangal*, a data format for the exchange of ecological networks and associated meta-
11 data. We deployed an online database with an associated API, relying on this data specification. Finally, we introduced
12 *rmangal*, a R package designed to interact with APIs using the *mangal* format. We expect that the data specification
13 will evolve based on the needs of the community. At the moment, users are welcome to propose such changes on the
14 project issue page: <https://github.com/mangal-wg/mangal-schemes/issues>. A python wrapper for the API is
15 also available at <http://github.com/mangal-wg/pymangal/>.

16 **References**

- 17 Bascompte, J. (2009). Disentangling the web of life. *Science (New York, N.Y.)*, **325**, 416–9.
- 18 Bastolla, U., Fortuna, M.A., Pascual-García, A., Ferrera, A., Luque, B. & Bascompte, J. (2009). The architecture of
19 mutualistic networks minimizes competition and increases biodiversity. *Nature*, **458**, 1018–1020.
- 20 Dalsgaard, B., Trøjelsgaard, K., González, A.M.M., Nogués-Bravo, D., Ollerton, J., Petanidou, T., Sandel, B., Schleuning,
21 M., Wang, Z., Rahbek, C., Sutherland, W.J., Svenning, J.-C. & Olesen, J.M. (2013). Historical climate-change influences
22 modularity and nestedness of pollination networks. *Ecography*, no–no. Retrieved May 14, 2013,
- 23 Dunne, J.A., Williams, R.J. & Martinez, N.D. (2002). Network structure and biodiversity loss in food webs: robustness
24 increases with connectance. *Ecology Letters*, **5**, 558–567.

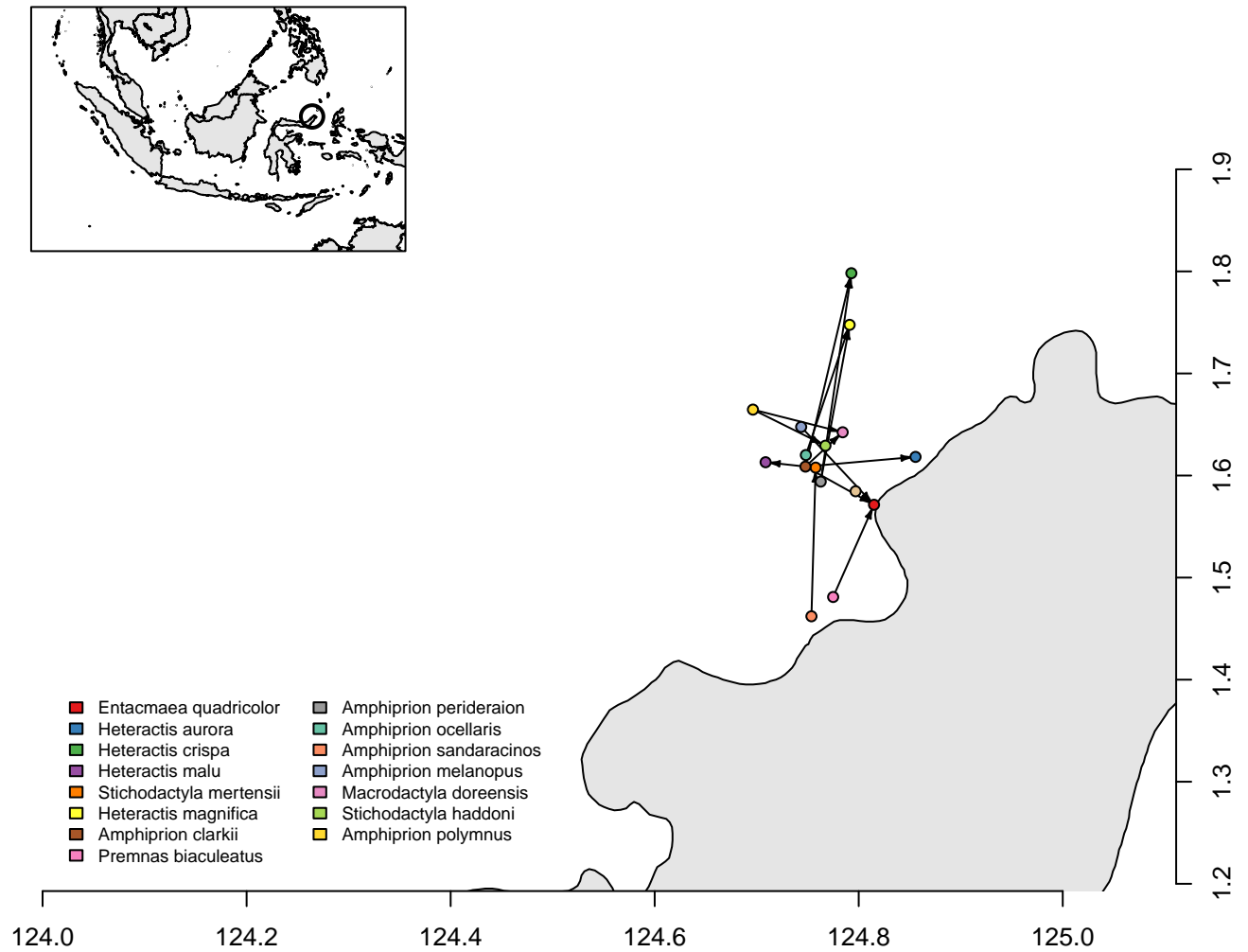


Figure 4: Spatial plot of a network, using the `maps` and `rmangal` packages. The circle in the inset map show the location of the sites. Each dot in the main map represents a species, with interactions drawn between them.

- 1 Gravel, D., Poisot, T., Albouy, C., Velez, L. & Mouillot, D. (2013). Inferring food web structure from predator-prey body
2 size relationships. *Methods in Ecology and Evolution*.
- 3 Martinez, N.D. (1992). Constant connectance in community food webs. *The American Naturalist*, **139**, 1208–1218.
- 4 Pimm, S.L., Lawton, J.H. & Cohen, J.E. (1991). Food web patterns and their consequences. *Nature*, **350**, 669–674.
- 5 Piwowar, H.A. & Vision, T.J. (2013). Data reuse and the open data citation advantage. *PeerJ*, **1**. Retrieved October 05,
6 2013,
- 7 Piwowar, H.A., Day, R.S. & Fridsma, D.B. (2007). Sharing detailed research data is associated with increased citation
8 rate. (J. Ioannidis, Ed.). *PloS one*, **2**, e308.
- 9 Poisot, T., Canard, E., Mouillot, D., Mouquet, N. & Gravel, D. (2012). The dissimilarity of species interaction networks.
10 *Ecology Letters*, **15**, 1353–1361.
- 11 Poisot, T., Mounce, R. & Gravel, D. (2013). Moving toward a sustainable ecological science: don't let data go to waste!
- 12 Ricciardi, F., Boyer, M. & Ollerton, J. (2010). Assemblage and interaction structure of the anemonefish-anemone mutual-
13 ism across the Manado region of Sulawesi, Indonesia. *Environmental Biology of Fishes*, **87**, 333–347. Retrieved January
14 10, 2014,
- 15 Rohr, R.P., Scherer, H., Kehrli, P., Mazza, C. & Bersier, L.-F. (2010). Modeling food webs: exploring unexplained
16 structure using latent traits. *The American naturalist*, **176**, 170–7.
- 17 Schleuning, M., Blüthgen, N., Flörchinger, M., Braun, J., Schaefer, H.M. & Böhning-Gaese, K. (2011). Specialization
18 and interaction strength in a tropical plant-frugivore network differ among forest strata. *Ecology*, **92**, 26–36.
- 19 Yodzis, P. (1981). The stability of real ecosystems. *Nature*, **289**, 674–676.