THINKING OUTSIDE THE BOX - PREDICTING BIOTIC INTERACTIONS IN DATA-POOR ENVIRONMENTS

DAVID BEAUCHESNE[1*], *PHILIPPE DESJARDINS-PROULX*[2], *PHILIPPE ARCHAMBAULT*[3], and *DOMINIQUE GRAVEL*[2]

[*] *email: david.beauchesne@uqar.ca*
[1] *Université du Québec à Rimouski*
[2] *Université de Sherbrooke*
[3] *Université Laval*

September 14, 2016

1 Abstract

Large networks of ecological interactions, such as food webs, are complex to characterize, be it empirically or theoretically. The former requires exhaustive observations, while the latter generally requires ample data to be validated. We therefore wondered whether readily available data, namely empirically described interactions in a variety of ecosystems and taxonomy, could be combined to predict species interactions in data deficient ecosystems. To test this, we built a biotic interactions catalogue from a collection of *94* empirical food webs, detailed predator-prey interaction databases and interactions from the Global Biotic Interactions (GloBI) database. We used an unsupervised machine learning method to predict interactions between any given set of taxa, given pairwise taxonomic proximity and known consumer and resource sets found in the interaction catalogue. Initial results suggest that pairwise interactions can be predicted with high accuracy. Although results are seemingly dependent on the comprehensiveness of the catalogue rather than the taxonomic similarity between taxa, taxonomy was found to complement well the catalogue and to allow the algorithm to perform well when the amount of information available through the catalogue decreased. Given it's high accuracy, this methodology could democratize the use of food webs and network level descriptors in remote location where empirical data is hard to gather. Network characteristics could then be efficiently evaluated and correlated to levels of environmental stressors in order to improve vulnerability assessments of ecosystems to global changes, opening promising avenues for further research and for management initiatives.

2 Introduction

Large networks of ecological interactions, such as food webs, are complex to characterize, be it empirically or theoretically. The former requires exhaustive observations, while the latter generally requires ample data to be validated. For this reason, studies focusing on communities of interacting species remain understudied even though we acknowledge the importance of considering the reticulated nature of complex networks (ref). When time is of the essence, the long term studies required quickly become impractical and the use of network level approaches relegated to the sideline (ref).

Alternatively, a currently evolving approach is to predict interactions using proxies such as functional traits, phylogenies and spatial distributions (e.g. Gravel et al., 2013; Morales-Castilla et al., 2015). For example, multiple traits can play a significant role in community dynamics and influence the presence and intensity of biotic interactions, like the influence of body size on predator-prey interactions, a literal take on *big fish eats small fish* (Cohen et al., 2003; Brose et al., 2006; Gravel et al. 2013). However, the time required to gather the necessary data to apply those methods may still be restrictive, or the data be unavailable altogether, so much so that other methods have been developed to fill the gaps in knowledge (e.g. Schrodt et al. 2016).

We therefore wondered whether more readily available data could be used to infer interactions in data deficient ecosystems. There is an increasing amount of data available describing worldwide species interactions, some freely available through the Global Biotic Interactions (GloBI) database (Poelen et al. 2014). Another readily available piece of information on species is their taxonomy, through initiatives like the World Register of Marine Species (WoRMS; Bailly et al. 2016). More than simple nomenclature, evolutionary processes are thought to influence consumer-resource relationships so that taxonomically related species would be more likely to share similar types of both consumers and resources (Eklof et al. 2012; Morales-Castilla et al. 2015). Based on that assumption, taxonomy might be useful in predicting interactions for species lacking detailed information, but which have a taxonomically related species for which such information is available. The objective of this work is thus to combine empirical biotic interactions originating from a variety of ecosystems with taxonomic relatedness to predict interactions in data deficient ecosystems.

3 Methods

The objective of the methodology presented is to predict the interactions between all taxa within an arbitrary set *S1* using a set of taxa *S0* with empirically described interactions from which we can extract sets of consumers and resources and their taxonomy. We couple the use of empirical data with an unsupervised machine learning method to achieve this.

3.1 Biotic interaction catalogue

We built a biotic interaction catalogue to serve as a set of taxa *S0* with empirically described interactions. The empirical data used to construct the interaction catalogue was gathered in two successive steps. The first consisted of gathering data from a collection of 94 empirical food webs in marine and coastal ecosystems from which we extracted pairwise taxa interactions (Brose et al. 2005; Kortsch et al. 2015; GlobalWeb database). We also used a detailed predator-prey interaction database describing trophic relationships between *XX* predators and their prey (Barnes et al. 2008). From these datasets, only interactions between taxa at the taxonomic scale of the family or higher were selected for inclusion in the catalogue.

As empirical food webs are vastly dominated by non-interactions, these datasets yielded a highly skewed distribution of interactions vs non-interactions. To counterbalance this, the second step of data compilation consisted of extracting observed interactions from the Global Biotic Interaction (GloBI) database (ref), which describes binary interactions for a wide range of taxa worldwide. We extracted all interactions available on GloBI for species belonging to the families of taxa identified through step 1. Interactions were extracted using the rGloBI package in R (ref). As for step 1, only interactions between taxa at the taxonomic scale of the family or higher were retained

The nomenclature used between datasets and food webs varied substantially. Taxa names thus had to be verified, modified according to the scientific nomenclature and validated. This process was performed using the Taxize package in R (ref) and manually verified for errors. The same package was used to extract the taxonomy of all taxa for which interactions were obtained in previous steps. The complete R code and data used for the catalogue is available at https://github.com/david-beauchesne/Interaction_catalog.

3.2 Unsupervised machine learning

We use the K-nearest neighbor (KNN) algorithm (**ref**) to predict pairwise interactions for a set of taxa S. The KNN algorithm predicts missing entries or proposes additional entries by a majority vote based on the K nearest (*i.e.* most similar) entries. In this case, taxa are described by a set of resources when considered as a consumer, a set of consumers when considered as a resource and their taxonomy (*i.e.* kingdom, phylum, class, order, family, genus, species). Similarity between taxa was evaluated using the Tanimoto similarity measure (**ref**), which compares two vectors with i elements based on the number of elements they share and contain:

$$tanimoto(\mathbf{x}, \mathbf{y}) = \frac{\sum_i x_i \wedge y_i}{\sum_i x_i \vee y_i}, \tag{1}$$

where $\wedge$ is bitwise *and*, while $\vee$ is the bitwise *or* operators. Adding a weighing scheme, we can measure the similarity using two different sets of vectors with i and j elements, respectively.

$$tanimoto_t(\mathbf{x}, \mathbf{y}, w_t) = w_t tanimoto(\mathbf{x_i}, \mathbf{y_i}) + (1 - w_t) tanimoto(\mathbf{x_j}, \mathbf{y_j}), \tag{2}$$

where w_t is the weight given to vector i, $\mathbf{x_i}$, $\mathbf{y_i}$ are the resource or consumer sets of the two taxa and $\mathbf{x_j}$ and $\mathbf{y_j}$ are the vectors for the taxonomy of two taxa. When $w_t = 0$ only resource or consumer sets are used to compute similarity, while $w_t = 1$ solely uses taxonomy.

3.3 Predicting interactions, Biotic predictor algorithm, Two-way Tanimoto algorithm, Feng shui name algorithm, Find a name for the algorithm

The XXX algorithm is built on a series of logical steps that ultimately predicts a candidate resources list C_R for each taxon in *S1* (Figure 1). For all consumer taxa T_C in *S1*, the algorithm first verify whether it has empirical resources T_R listed in the catalogue. When it does, if T_R are also in *S1*, they are added as predicted resources for T_C. This corresponds to what we refer to as the catalogue contribution to resource predictions. Two taxa in *S1* that are known to interact through the catalogue are automatically assumed to interact in *S1*.

Otherwise, the algorithm passes to what we refer to as the predictive contribution to resource predictions, with candidate resources for T_C identified with the KNN algorithm. If T_R are also in *S1*, K most similar resource $T_{R'}$ are identified in *S1* to add to C_R. Then for all T_C in *S1*, the algorithm identifies K most similar consumer $T_{C'}$ in *S0* and extracts their resource sets. As before, if those resources are found in *S1* they are added to C_R, otherwise K most similar resources $T_{R'}$ are identified in *S1* to add to C_R. A simple working example is presented at Figure **??**. Other parameters are used in the algorithm, but not presented here for the sake of message clarity. A more comprehensive mathematical description of the algorithm and the parameters used is however available through Figure 1 and the complete R code and data used for the algorithm is available at https://github.com/david-beauchesne/Predict_interactions.

3.4 Algorithm prediction accuracy

We used the most extensive and taxonomically detailed datasets included in the catalogue (**ref**) to assess the prediction accuracy of the algorithm. Testing accuracy of a particular dataset was done by first removing from the catalogue all pairwise interacting originating from that dataset. Accuracy was evaluated using three different statistics:

1. $Score_y$ is the fraction of interactions correctly predicted:

$$Score_y = \frac{a}{a+c} \tag{3}$$

2. $Score_{\neg y}$ is the fraction of non-interactions correctly predicted:

$$Score_{\neg y} = \frac{d}{b+d} \tag{4}$$

3. TSS, The True Skilled Statistics (TSS) evaluated prediction success by considering both true and false predictions, returning a value ranging from 1 (prefect predictions) to -1 (inverted predictions; **ref**):

$$TSS = \frac{(ad-bc)}{(a+c)(b+d)} \tag{5}$$

where a is the number of links predicted (1) and observed (1), b is the number predicted (1) but not observed (0), c is the number predicted absent (0) but observed (1) and d is the number of predicted absent (0) and observed absent (0). These three statistics give a different perspective on prediction accuracy, focusing in turn on true interactions and non-interactions, and on both true and false predictions.

We evaluated the three statistics for the complete algorithm and for the catalogue and the predictions individually to evaluate their respective contribution to the algorithm predictive accuracy. Multiple w_t values were also tested to

evaluate whether taxa similarity measured as a function of resource/consumer sets or taxonomy contributed more significantly towards increased predictive accuracy. The same was done with multiple K values.

Finally, we evaluated the influence of the comprehensiveness of the catalogue on prediction accuracy. We selected the arctic food web from Kortsch et al. (2015). This food web was selected as it is highly detailed taxonomically and because empirical data remains available for most of its taxa after its exclusion from the catalogue. We iteratively and randomly ($n = 50$) removed a percentage of empirical data describing the food web taxa from the catalogue before generating new predictions with the algorithm. We also tested w_t values of 0.5 and 1 to evaluate whether taxonomic similarity could support predictive accuracy in cases when empirical data for species in *S1* in the catalogue is unavailable.

3.5 Southern Gulf of Saint Lawrence

As an example, we used the XXX algorithm to predict interactions in the Southern Gulf of Saint Lawrence (SGSL) in eastern Canada. The empirical data and taxa list come from Savenkoff et al. (2004). They present a list of 29 functional groups for a total of 80 taxa. As their analysis was performed on the functional groups rather than the taxa themselves, we used the algorithm to predict interactions between all 80 taxa, which we aggregated back to their original functional groups to compare with interactions presented in Savenkoff et al. (2004).

4 Results

4.1 Biotic interaction catalogue

The data compilation process allowed us to build an interaction catalogue composed of 276708 pairwise interactions (interactions = 72110; non-interactions = 204598). A total of 9712 taxa (Superfamily = 15; Family = 591; Subfamily = 29; Tribe = 8; Genus = 1972; Species = 7097) are included in the catalogue, 4159 of which have data as consumers and 4375 as resources.

4.2 Algorithm predictive accuracy

The overall predictive accuracy of the algorithm ranges between 80% to almost 100% in certain cases (Figure 2), except for one food web (*i.e.* **ref**). Both interactions and non-interactions are well predicted by the algorithm. TSS scores are lower than $Score_y$ and $Score_{\neg y}$ due to misclassified interactions and non-interactions. This can also be observed through the effect of varying K values, which increases the number of potential candidate resources for each taxa in the predictive portion of the algorithm. Prediction accuracy increases for interactions, while it decreases for non-interactions, as K values increase.

Similarity being predominantly measured with resource/consumer sets (w_t closer to 0) yielded better predictions than when measured with taxonomy (w_t

closer to 1; Figure 2). Resource/consumer sets therefore appears to serve as a better predictor of similarity between taxa for interactions predictions. It is nonetheless interesting to note that although the predictive contribution of the algorithm decreases as w_t increases, an increased mean and decreased variability values for the TSS and $Score_y$ statistics is also observed (Figure 2)). This suggests that while using taxonomy for similarity measurements yields lower predictive accuracy, it may also complement the catalogue contribution by predicting interactions not captured through empirical data, effectively increasing the predictive accuracy of the complete algorithm.

The partitioning of the catalogue and predictive portions of the algorithm shows that it is dependent on the comprehensiveness of the catalogue for high prediction accuracy (Figures 2, 3). As the amount of empirical data available in the catalogue decreases so does the overall accuracy of the algorithm (Figures 3). The predictive contribution of the algorithm however slows down the decrease in the prediction efficiency of the algorithm. Prediction accuracy still remains around 75% with only 40% of *S1* taxa found in the catalogue (Figures 3). Furthermore, the use of taxonomy for similarity measurements is more efficient as empirical data becomes scarcer and no different than resource/consumer sets for the complete algorithm when ample data is available (Figures 3).

4.3 Southern Gulf of Saint Lawrence

In total, there were empirical data available through the catalogue for 53 out of 80 SGSL taxa. The XXX algorithm correctly predicted approximately 75% of interactions (a) and non-interactions (d) extracted from Savenkoff et al. (2004). It however predicted an additional 113 interactions (c) that were not noted in Savenkoff et al. (2004) and failed to predict 44 observed interactions that were (c).

5 Discussion

We show that out of the box interaction inference for a set of taxa with incomplete or unavailable preexisting information can be achieved with high accuracy using a combination of empirical data describing biotic interactions and taxonomic relatedness. Although the efficiency of the algorithm is dependent on the comprehensiveness of the interactions catalogue, taxonomic proximity acts as a complement to increase the number of observed interactions correctly predicted. Taxonomic proximity also supports the efficiency of the algorithm when catalogue comprehensiveness decreases. Overall, we believe that the methods performs well and offers promising avenues for further research and management initiatives.

5.1 Algorithm accuracy

5.2 Interactions classification

The $Score_y$ and $Score_{\neg y}$ are inversely proportional. This means that non-interactions are misclassified as interactions in the process of increasing $Score_y$, consequently decreasing $Score_{\neg y}$. This could either stem from the algorithm poorly predicting non-interactions or from the empirical data itself. Accuracy evaluation assumes that non-interactions from empirical food web are observed data, yet it is usually not the case. Most empirical webs have a strong focus attributed to higher order consumer species and very little attention given to other taxa (**ref**). Furthermore, the methodologies used to obtain consumer-resource data usually relies in gut content analyses, which is efficient at observing interactions, but not so for absence of interactions (**ref**). Misclassified interactions could thus be real, albeit unobserved through empirical data available. It should however be noted that impossible interactions are still predicted by the algorithm, like atlantic cod (*Gadus morhua*) and large demersal feeders such as the atlantic halibut (*Hippoglossus hippoglossus*) consuming whales and birds.

5.3 Southern Gulf of Saint Lawrence

Ok, so the fact that it is not that well predicted, that we predict more

We can break down functional groups for a more thorough understanding of individual components of the network rather than the aggregated version.

With 80 species, it isn't such a big deal, but the whole Saint Lawrence comprises over 1300 species. Grouping species in functional groups makes more sense. Forming functional groups based on their interactions with other species, i.e. describing analytically identified groups, could provide a new way of aggregating species together and at the same time evaluate whether original functional groupings made sense.

Something else that could be done is to reevaluate

5.4 Usefulness of taxonomic relatedness

While we found that taxonomy could be useful as a complement to predictions made using empirical data, the accuracy of predictions made using the KNN algorithm are quite low. Other uses of this machine learning approach have been known to achieve much higher prediction rates (*e.g.* **refs**), which leaves us to believe that taxonomy is not the optimal proxy for predicting interactions. While evolutionary history plays a significant role in influencing consumer-resource trait matching and food web structure (**ref**), phylogenetic constraints does not account efficiently for certain traits such as body size (Eklöf and Stouffer 2015). Including traits like body size and metabolism as an additional component of this algorithm could thus help increasing overall prediction accuracy, especially in cases where the catalogue lacks data on taxa for which interactions have to be predicted. Although promising, such an approach would undermine the premise

under which this method was built and which constitutes its main strength, predicting interactions in data deficient environments using readily available data.

5.5 Perspectives

Given the efficiency and simplicity of the XXX algorithm,

The simplicity of the XXX algorithm could prove useful for a variety of reasons

Given it's high accuracy, this methodology could democratize the use of food webs and network level descriptors in remote location where empirical data is hard to gather. Network characteristics could then be efficiently evaluated and correlated to levels of environmental stressors in order to improve vulnerability assessments of ecosystems to global changes, opening promising avenues for further research and for management initiatives.

Democratizing the use of network level predictors Management Applied research in theoretical ecology Adding interaction strength to predictions

6 Acknowledgements

We thank the Fond de Recherche Québécois Nature et Technologie (FRQNT) and the Natural Science and Engineering Council of Canada (CRSNG) for financial support. This project is also supported by Québec Océan, the Quebec Centre for Biodiversity Science (QCBS), and the Notre Golfe and CHONeII networks. We also wish to thank K. Cazelles for constructive comments and suggestions.

6.1 Box 1

The XXX algorithm follows a series of logical steps to predict resources for all taxa in an arbitrary set of taxa *S1* using a set of taxa *S0* with empirically described interactions from which we can extract sets of consumers and resources and their taxonomy. In this example, we are predicting interactions for a fictitious $S1 = \{T_1, T_9, T_{10}, T_{11}, T_{12}\}$ using *S0* with information on 12 taxa. This catalogue holds information on consumer or resource for 10 taxa and the taxonomy for all 12 taxa in the list.

S0 taxa ID	taxonomy	resource	consumer
T_1	$\{a, b, c\}$	$\{T_2, T_3, T_{12}\}$	$\{T_4\}$
T_2	$\{e, f, g\}$		$\{T_1, T_5\}$
T_3	$\{i, j, k\}$		$\{T_5\}$
T_4	$\{m, n, o\}$	$\{T_1, T_5\}$	
T_5	$\{a, b, d\}$	$\{T_8, T_9\}$	$\{T_4\}$
T_6	$\{i, q, r\}$	$\{T_2, T_8\}$	$\{T_4\}$
T_7	$\{e, f, h\}$		$\{T_1, T_6\}$
T_8	$\{s, t, u\}$		$\{T_5, T_6\}$
T_9	$\{s, t, v\}$		$\{T_5\}$
T_{10}	$\{i, j, l\}$		
T_{11}	$\{m, n, p\}$		
T_{12}	$\{q, r, s\}$		$\{T_1\}$

Similarity between all pairs of taxa in *S0* is measured for consumer, resource and taxonomic proximity using equation 1. The upper triangular matrix represents similarity measured with taxa sets of resources/consumers, while the lower triangular represents taxonomic similarities. For consumer/resource set similarities, values of 0 mean that similarity equals 0 for both similarity measurements.

$tanimoto(T_C x, T_C y)$ / $tanimoto(T_R x, T_R y)$

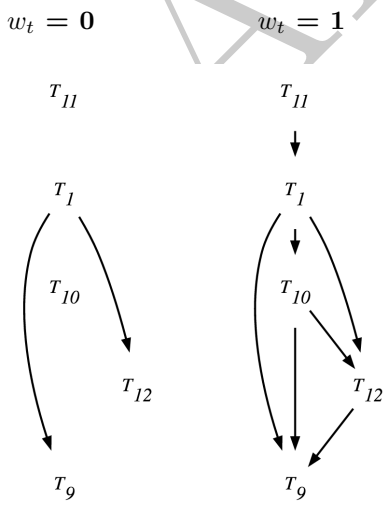
	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	T_{12}
T_1	-	0	0	0	0/1	0.3/1	0	0	0	0	0	0
T_2	0	-	0/0.5	0	0	0	0/0.3	0/0.3	0/0.5	0	0	0/0.5
T_3	0	0	-	0	0	0	0	0/0.5	0/1	0	0	0
T_4	0	0	0	-	0	0	0	0	0	0	0	0
T_5	0.5	0	0	0	-	0.3/1	0	0	0	0	0	0
T_6	0	0	0.2	0	0	-	0	0	0	0	0	0
T_7	0	0.5	0	0	0	0	-	0/0.3	0	0	0	0/0.5
T_8	0	0	0	0	0	0	0	-	0	0	0	0
T_9	0	0	0	0	0	0	0	0.5	-	0	0	0
T_{10}	0	0	0.5	0	0	0.2	0	0	0	-	0	0
T_{11}	0	0	0	0.5	0	0	0	0	0	0	-	0
T_{12}	0	0	0	0	0	0.5	0	0.2	0.2	0	0	-

$tanimoto(T_T x, T_T y)$

From these, the algorithm goes through logical steps to identify a candidate resource list C_R for each taxon in *S1* using either empirical data directly or K most similar taxa with equation 2. Going through the process for T_1, using $K = 1$ and $w_t = 1$:

Steps	Catalogue	Predictions
1) $I(T_1, T_R)$ in *S1*?		
T_2 = no $\rightarrow t(T_2, T_{R'}, w_t)$ = NA	{}	{}
T_3 = no $\rightarrow t(T_2, T_{R'}, w_t) = T_{10} = 0.5$	{}	$\{T_{10}\}$
T_{12} = yes	$\{T_{12}\}$	$\{T_{10}\}$
2) $t(T_1, T_{C'}, w_t) = T_5 = 0.5$		
I(T_5,T_R) in *S1*?		
T_8 = no $\rightarrow t(T_8, T_{R'}, w_t) = T_9 = 0.5$	$\{T_{12}\}$	$\{T_9, T_{10}\}$
T_9 = yes	$\{T_9, T_{12}\}$	$\{T_9, T_{10}\}$

The logical steps allow us to predict a set of resources for $T_1 = T_9$, T_{10}, T_{12}. Doing it for all taxa in *S1* with $w_t = 0$ and 1 predicts the following networks:



6.2 Figures

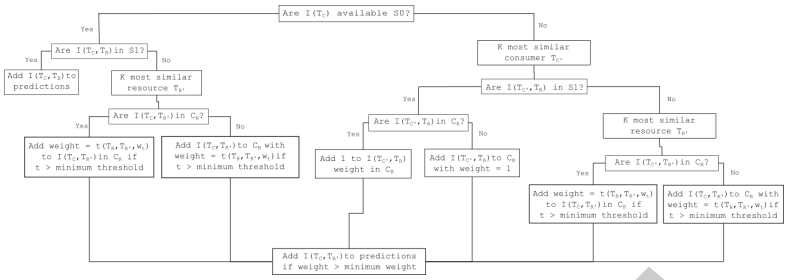


Figure 1: Description of the logical steps used by the algorithm to suggest a list of candidate resources (C_R) for each taxa (T_C) in *S1*

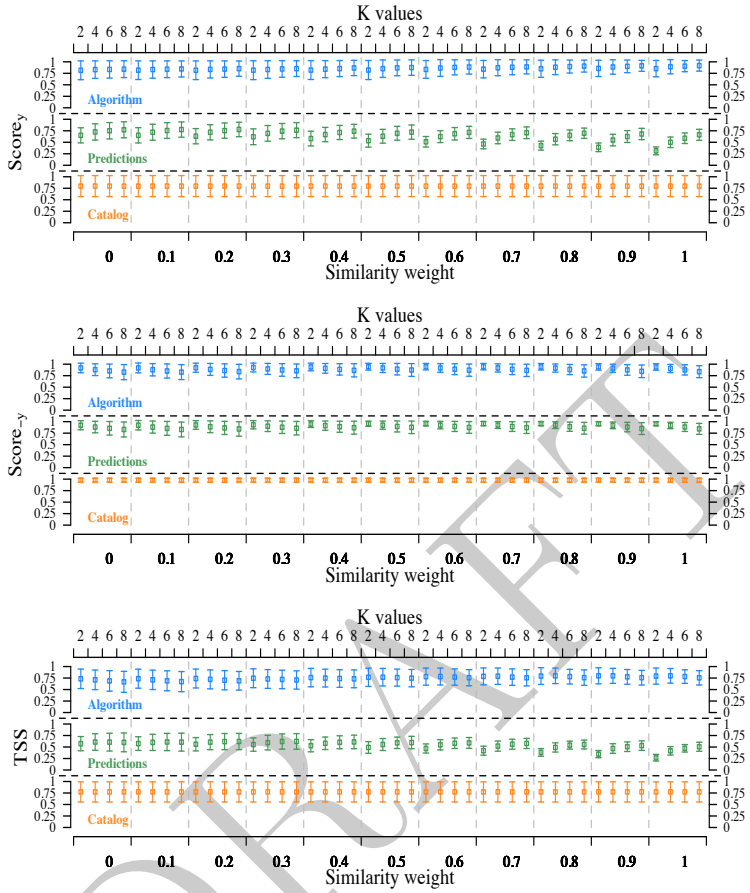


Figure 2: The graph presents the three statistics as a function of trait weight, which varies between 0 and 1. A weight of 0 means that similarity is measured only using set of resources for each taxa, while a weight equal to 1 means that similarity is based solely on taxonomy. We present 6 food webs with over 50 taxa each and the Barnes et al. (2008) dataset.

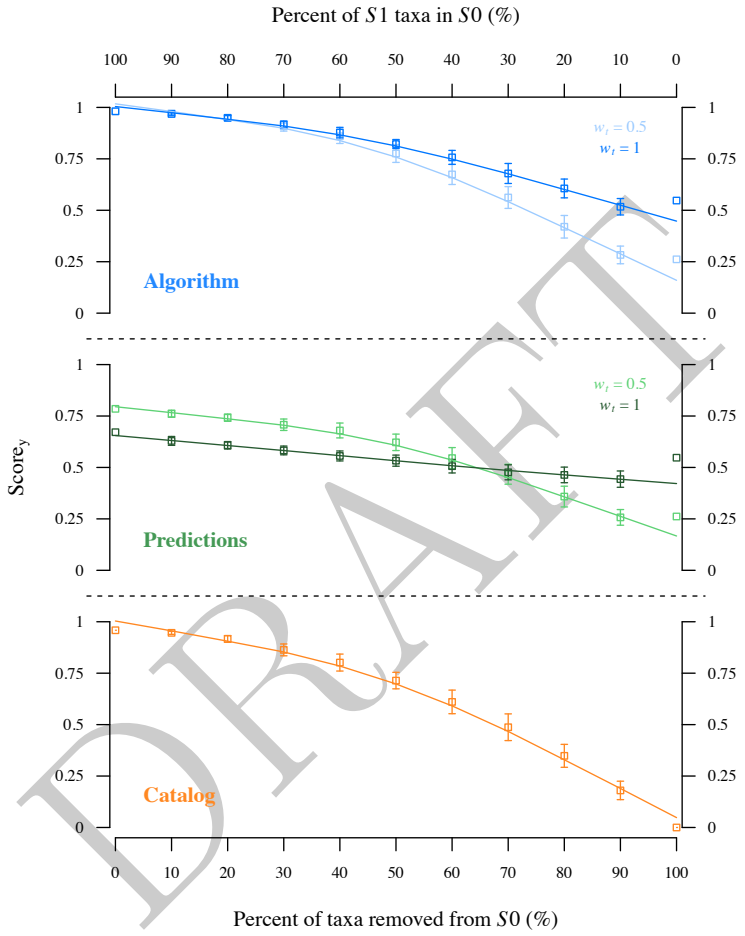


Figure 3: Graph presenting predictive accuracy as a function of the amount of information available in the catalogue. The arctic food web from Kortsch et al. (2015) was used for this, as it is highly detailed and because almost all taxa found in it had information in the catalogue even when not included in the catalogue. A random percentage of taxa in the web was iteratively removed from the catalogue (n = 50) before predicting interactions with the XXX algorithm.