ISME

www.nature.com/ismecomms

ARTICLE OPEN

Check for updates

NetGAM: Using generalized additive models to improve the predictive power of ecological network analyses constructed using time-series data

Samantha J. Gleich [1]✉, Jacob A. Cram [2], Jake L. Weissman [1] and David A. Caron [1]

Ecological network analyses are used to identify potential biotic interactions between microorganisms from species abundance data. These analyses are often carried out using time-series data; however, time-series networks have unique statistical challenges. Time-dependent species abundance data can lead to species co-occurrence patterns that are not a result of direct, biotic associations and may therefore result in inaccurate network predictions. Here, we describe a generalize additive model (GAM)-based data transformation that removes time-series signals from species abundance data prior to running network analyses. Validation of the transformation was carried out by generating mock, time-series datasets, with an underlying covariance structure, running network analyses on these datasets with and without our GAM transformation, and comparing the network outputs to the known covariance structure of the simulated data. The results revealed that seasonal abundance patterns substantially decreased the accuracy of the inferred networks. In addition, the GAM transformation increased the predictive power (F1 score) of inferred ecological networks on average and improved the ability of network inference methods to capture important features of network structure. This study underscores the importance of considering temporal features when carrying out network analyses and describes a simple, effective tool that can be used to improve results.

ISME Communications; https://doi.org/10.1038/s43705-022-00106-7

INTRODUCTION

Communities of microorganisms exist in virtually all natural environments on the planet and are shaped by complex interactions. Species-species interactions are diverse and may benefit both species involved (e.g., mutualism), only one (e.g., commensalism, parasitism, predation), or hurt both (e.g., competition) [1, 2]. Ecological network analyses are increasingly used by microbial ecologists to identify potential biotic interactions between organisms, and to form hypotheses regarding microbial community structure and function [1, 3, 4]. Commonly employed network-inference methods include pairwise correlation-based, regression-based, and probabilistic graphical methods [5, 6]. All of these methods leverage microbial abundance measurements to identify co-occurrence patterns between organisms [1, 7, 8]. Biotic interactions between organisms are then predicted based on the species' co-occurrence patterns, resulting in a list of nodes (organisms) connected by edges (associations) [1, 7–9].

Abundance data that are used as input for ecological network analyses are often obtained through monthly time-series sampling efforts [10–15]. Time-series datasets are valuable because documenting changes in microbial community structure over long timeframes can provide information on the monthly, annual, or interannual variability in species abundances [16, 17]. For example, ~11 years of monthly sampling at the San Pedro Ocean Time-Series site revealed that 23% of bacterial operational taxonomic units demonstrated predictable, seasonal abundance patterns in surface waters [13]. Another time-series study that took place over ~6 years in the Western English Channel revealed that the month of year could explain over half of the variability in bacterial community composition [12]. Such studies highlight how long-term time-series datasets may be used to identify predictable changes in microbial community composition over time.

Time-series datasets can provide information on microbial community composition and structure, but ecological networks inferred from them should be built and interpreted with caution. There are statistical challenges associated with time-series network analyses because the samples are not independent over time [12, 18, 19]. This inherent time-dependence may be influenced in part by seasonally reoccurring patterns in the abiotic environment (e.g., seasonal mixing or upwelling events) or long-term changes in the environment over-time (e.g., rising seawater temperature) [17, 20, 21]. As a result of this time-dependence, species abundance patterns can lead to co-occurrence patterns that yield spurious network predictions [4, 22]. For instance, two species may both attain maximal abundances during spring nutrient upwelling events, even if no interactions occur between them. Their shared periodicity in this case may manifest itself as a false "association". Time-dependence may also confound species co-occurrence patterns through the effects of Simpson's paradox [23, 24]. If a mutualistic relationship

[1] Department of Biological Sciences, University of Southern California, 3616 Trousdale Parkway, AHF, Los Angeles, CA 90089-0371, USA. [2] Horn Point Laboratory, University of Maryland Center for Environmental Science, 2020 Horns Point Road, Cambridge, MD 21613, USA. ✉email: gleich@usc.edu

Received: 28 October 2021 Revised: 10 February 2022 Accepted: 17 February 2022
Published online: 10 March 2022

exists between two species, one might expect a positive correlation between the abundances of the two species across samples. However, if the species respond differently to a third variable (e.g., month of year), then the positive association between the two species may be offset or reversed as a result of this time-dependence [24]. Such inaccurate associations indicate that caution should be exercised when carrying out network analyses on time-series datasets [7, 24].

Here, we propose and validate a generalized additive model (GAM)-based data transformation that corrects for potentially confounding time-series signals that are prevalent in microbial relative abundance data. The GAM transformation was conducted prior to carrying out ecological network analyses, in order to remove seasonal, long-term, and autocorrelative trends, thereby allowing researchers to focus on the residual statistical variability of the microbial abundance data. We contend that the residual variability is likely more indicative of true biotic associations than are untransformed data. We used GAMs in this data transformation method, as they are versatile, and commonly used to capture non-linear trends typical of time-series data [25]. Generalized additive models have been used to model both seasonal patterns and long-term trends in time-series data [25–27] and have also been used to capture autocorrelative signals [17, 28]. The GAM-based data transformation presented here has the potential to capture seasonal, long-term, and autocorrelative trends in time-series datasets, thus minimizing the influence that temporal signals have on inferred microbial co-occurrence patterns and increasing the F1 score of commonly employed networking methods.

MATERIALS AND METHODS

Our general strategy was to compare the performance of four approaches for inferring microbial associations from abundance data with overlying time-series signals. The approaches were (1) pairwise spearman correlation analysis (SCC) [1, 29], (2) Graphical lasso analysis (Glasso) [30, 31], (3) pairwise SCC analysis with a pre-processing step where seasonal and long-term splines were fit to and subtracted from each variable using a GAM (GAM-SCC), and (4) Glasso with the same GAM subtraction approach (GAM-Glasso). Our validation strategy for the GAM transformation consisted of generating mock datasets with underlying associations, masking those associations by adding seasonal and long-term signals to the abundance data, and comparing the predicted associations obtained from each network inference method to the true species-species associations.

Data simulation: generating mock abundance data with time-series properties

We generated mock abundance datasets that had a predetermined, underlying network structure and contained long-term and seasonal species abundance patterns. First, a covariance matrix was generated to describe the relationships between species in a mock dataset (Fig. S1, Panel 1). The covariance matrices were constructed with underlying network structures that followed either a scale-free Barabási-Albert model, a random Erdős-Rényi model, or a model of network topology based on a real microbial dataset (American Gut dataset; Fig. S1) [32, 33]. The Erdős-Rényi and Barabási-Albert model datasets were generated so that each dataset contained 400 species and 200 samples, and the American Gut datasets were created so that each dataset contained 127 species and 200 samples. A random Bernoulli distribution was used to simulate the covariance matrix for the Erdős-Rényi networks. We set the probability of interactions occurring between species in a given Erdős-Rényi network to 1%. The Barabási-Albert networks were generated using the "sample_pa" function in the igraph package [34]. The "graph2prec" function in the SpiecEasi package was used to predict the covariance matrix of the American Gut dataset [33]. The covariance between species in a dataset was considered "high" or "low" when the true associations in the covariance matrix were set to 100 or 10 respectively (Fig. S1, Panel 1). These covariance matrices describe the "real", underlying species interactions in our mock datasets.

After generating a covariance matrix, the mean abundance for each species was generated from a normal distribution with a mean of 10 and a variance of 1. These mean abundance values and the covariance matrix were used to parameterize a multivariate normal distribution from which species abundance values for all 200 samples in a dataset were drawn (Fig. S1, Panel 2). The values generated from this multivariate normal distribution were the species abundance values without time-series features confounding the relationship between two associated species (Fig. S1, Panel 2).

"Gradual" or "abrupt" seasonal trends were added to 0%, 25%, 50% or 100% of the species in each mock dataset. The gradual seasonal trend increased over 5 months, peaked at a specific month, and decreased over 5 months. Conversely, the abrupt seasonal signal increased over 2 months, peaked at a specific month, and decreased over 2 months (Fig. S1, Panel 3). These seasonal signals were generated by plugging a vector of consecutive integers of length 200 (N_t) into the gradual (Eq. (1)) or abrupt (Eq. (2)) seasonal equations (Fig. S1, Panel 3)…

$$Gradual : S_t = \left(\frac{\cos\left(N_t * 2 * \frac{\pi}{12}\right)}{2}\right) + 0.5 \tag{1}$$

$$Abrupt : \ S_t = \left(\left(\frac{\cos\left(N_t * 2 * \frac{\pi}{12}\right)}{2}\right) + 0.5\right)^{10} \tag{2}$$

where N is the random vector of consecutive integers, S is the output seasonal vector, and t is the index of vectors N and S. The starting value of vector N_t was drawn at random for each species to allow the seasonal peaks to be centered at different months. Each element in the seasonal vector (S_t) was then multiplied by the corresponding element in the abundance vector (X_t) of a specific species to obtain mock species abundance values with a gradual or abrupt seasonal trend (Fig. S1, Panel 3).

A long-term time-series trend was added to the abundance values of 0% or 50% of the species in the mock datasets (Fig. S1, Panel 4). When a long-term signal was applied to 50% of the species in a dataset, half of the species were randomly selected to have this long-term trend. Then, a vector of linear values was generated following Eq. (3) such that…

$$Long - term\ trend : \ L_t = \pm m(L_{t-1}) + 0.01 \tag{3}$$

where L_t is the point in the line at the next time point and m is the slope of the line. The slope parameter (m) was generated from a random normal distribution with a mean of 0.01 and a variance of 0.01. The slope parameter (m) was also multiplied by -1 half of the time to ensure that half of the long-term trends increased over time and half decreased over time (Fig. S1, Panel 4). After generating the vector of linear values (L_t), each element of this vector was added to each element of the abundance vector (X_t) of a specific species to simulate long-term time-series trends (Fig. S1, Panel 4).

Time-series predictor columns were added to each dataset after applying monthly and long-term abundance trends to a portion of the species in the mock datasets. The predictors that were used in the downstream GAM-based data transformation were the month of the year (i.e., 1–12) and the day of the time-series (i.e., 1–200). In total, we generated 100 mock datasets for every combination of conditions (84 combinations total; Table S1), resulting in 8400 mock time-series datasets that were used in the downstream count data transformation, GAM subtraction, and network analysis procedures.

Data simulation: Simulating count data from abundance values

The 8400 time-series datasets that were generated using the methods described above were transformed to make the abundance values resemble high-throughput sequencing data because microbial time-series sampling efforts are often processed using such molecular methods (e.g., tag-sequencing, meta-omics). Analysis of high-throughput sequencing data is complicated by the compositional (i.e., relative) nature of the data and by the high number of zeros that may be prevalent in a dataset (i.e., zero-inflation; see Supplementary Information) [35, 36]. Relative abundances of different species in natural communities are also highly skewed, so that relatively few species constitute most of the organisms in a sample although many rare species are also present [37, 38]. Therefore, species abundances were first exponentiated to increase the prevalence of abundant species and to decrease the prevalence of rare species (Fig. S1, Panel 5). The exponentiated species abundance values were then converted to relative abundance values by dividing each species count by the sum of all species counts in a sample (Fig. S1, Panel 6). The resulting relative abundance values and time-series predictor variables were used in data normalization and GAM-transformation steps prior to carrying out the network analyses.

Network inference: Count data normalization and GAM transformation

Several steps were taken to back out the species-species relationships in the mock datasets. We advocate these steps to infer network structure from a real time-series dataset. A centered log-ratio (CLR) transformation was first applied to the species relative abundance values to normalize the mock species abundance data across samples using the "clr" function in the compositions package in R (Fig. 1) [39]. This transformation step is important to avoid spurious inferences induced by the inherent compositionality of relative abundance data [31, 33, 36]. In addition to the CLR transformation used in our main network iterations, we carried out additional network iterations using the modified CLR [40], cumulative sum scaling [41], and total sum scaling [42] transformations (see Supplementary Information). In all cases, the normalized dataset was copied, with one copy subjected to a subsequent GAM transformation, and the other one not GAM-transformed.

The GAM transformation was carried out by fitting GAMs to each individual species in the dataset to remove monthly signals, long-term trends, and autocorrelation from the species abundance data. These GAMs were fit using the "gamm" function in the mgcv package in R [43, 44]. The GAMs that were used included the "month of year" parameter as a cyclical spline predictor and the "day of time-series" parameter as a penalized thin-plate spline predictor ("ts" in the mgcv package; Fig. 1), which given our one-dimensional data is analogous to a natural cubic spline. In addition, the first GAM included a continuous AR1 ("corCAR1" in the mgcv package) correlation structure term in the model. This corCAR1 model was revised for specific species when the GAM could not be resolved or when significant autocorrelation was detected in the GAM residuals (Fig. 1). The GAM revision step fit 4 new GAMs with different correlation structure terms (i.e., "AR1", "CompSymm", "Exp", and "Gaus") to the species that could not be fit using the corCAR1 model or that contained significant autocorrelation in the corCAR1 GAM residuals. Then, the correlation structure term that addressed these issues for the largest number of individuals was used as the GAM model for this group of species. After fitting a GAM to all of the species in the input dataset, the residuals of each GAM were extracted and were used as the new, GAM-transformed abundance values (Fig. 1). These GAM residuals represent species abundance values with a reduced influence of time (Fig. 2) and were used as input in the downstream GAM-SCC and GAM-Glasso network analyses.

Network inference: Network runs and statistical analyses

The pre-processed species abundance data with and without the GAM-removal of time-series signals were used in SCC and Glasso networks in order to compare the outputs of the SCC, GAM-SCC, Glasso, and GAM-Glasso network inference approaches (Fig. 1). Additional network iterations were also carried out using the CCLasso [45] and SPRING [40] network inference approaches (see Supplementary Information). For the SCC and Glasso network iterations, a nonparanormal transformation was applied to the species abundance datasets with and without the GAM transformation using the "huge.npn" function in the huge package in R [46]. Spearman correlation networks were then constructed by calculating the correlation between every pair of species in the mock abundance datasets. A Bonferroni-corrected p value of 0.01 was used as a cutoff to identify edges in these SCC networks. The Glasso networks were constructed by testing 30 regularization parameter values (i.e., lambdas) in each network using the "batch.pulsar" (criterion = "stars"; rep.num = 50) function in the pulsar package in R [47]. The lambda that resulted in the most stable network output was selected using the StARS method [48]. Finally, the graph that resulted from the StARS output was used to obtain a species adjacency matrix for the Glasso networks.

The species-species associations predicted by the SCC, GAM-SCC, Glasso, and GAM-Glasso networks were compared to the true species-species associations and the F1 scores of the network predictions were calculated. The F1 score is a measure of classification performance (presence or absence of an edge) that accounts for uneven classes, which is essential when dealing with sparse networks. The F1 scores of the GAM-transformed networks were compared to the networks that did not undergo GAM transformation using paired Wilcoxon tests with Bonferroni correction. An adjusted p value of 0.01 was used as a cutoff to identify under what circumstances the GAM significantly improved the F1 score of a Glasso or SCC network.

Network inference: Comparison of predicted network structures

Additional networks were generated using the methods described above to compare the predicted network structures obtained from the GAM-Glasso, Glasso, GAM-SCC, and SCC approaches to the real network structures. These additional networks were constructed using smaller mock datasets to allow for better visualization of the network outputs and contained species with a gradual seasonal signal and high species-species covariance (see Supplementary Information). The average clustering coefficient and the degree distribution of these additional network outputs were calculated and used for the network structure comparisons. The average clustering coefficient of a network describes the likelihood that two species that are both associated with a third species are also associated with each other [49], and in a sense describes the "clumpiness" of a network. The network degree distributions describe the probability distribution of the number of interactions per node in a network [50].

RESULTS

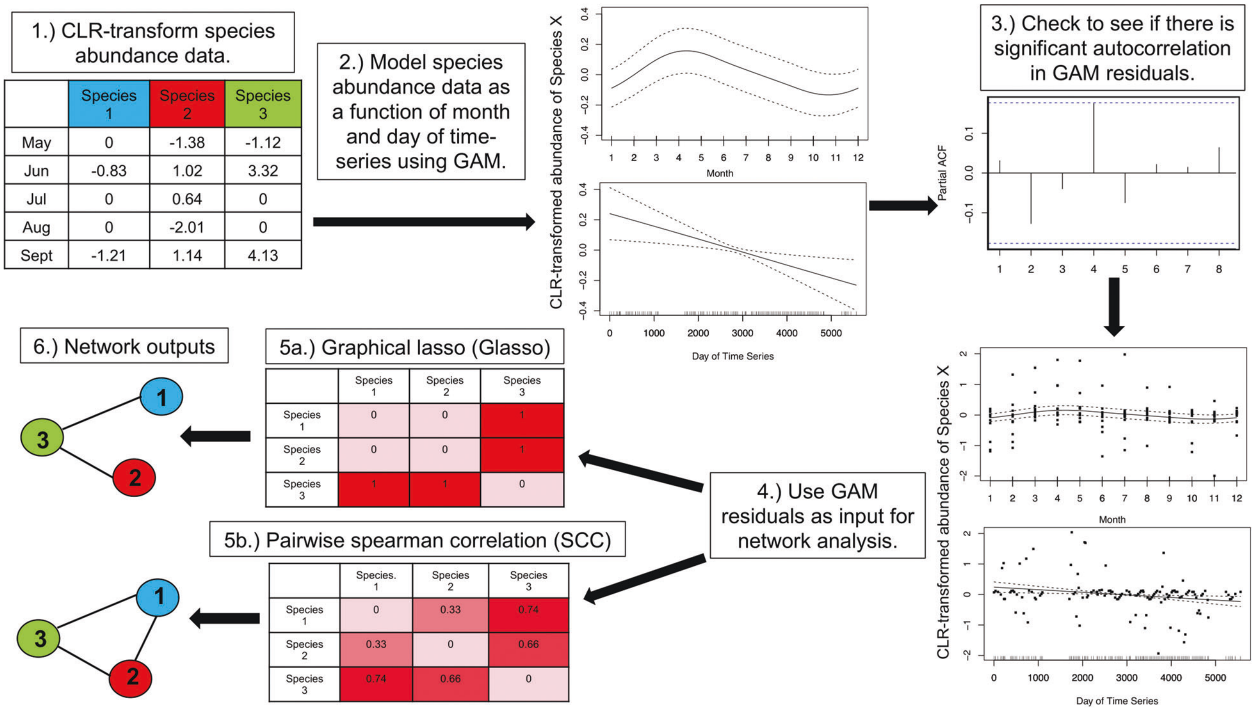
Seasonal abundance patterns decreased the performance of network inference methods

The F1 scores of the reconstructed network outputs generally decreased as the proportion of species in the mock dataset with a seasonal abundance pattern increased. The decreases in network F1 scores with increases in the percent of species with a seasonal abundance pattern were always prevalent in the Glasso and SCC networks when the GAM transformation was not applied and when some percentage of species (>0%) in the mock datasets had a seasonal abundance pattern (Fig. 3; Tables S2–S4). In general, the highest F1 scores were associated with networks that did not contain any species with an underlying seasonal signal (0%), while the lowest F1 scores were typically associated with networks in which all of the species had a seasonal abundance pattern (100%; Fig. 3; Tables S2–S4).

The general decline in network F1 score with a greater percentage of species exhibiting seasonality was often less pronounced when the mock datasets were GAM-transformed prior to carrying out the network analyses. For example, when species-species covariance was high, the GAM-SCC method tended to perform similarly regardless of whether 25%, 50%, or 100% of the species in a Barabási-Albert or American Gut dataset network had a gradual seasonal abundance pattern (Figs. S5 and S7, Panels B and F). The decline in network F1 score with increases in the percentage of species exhibiting seasonality was also less pronounced in GAM-Glasso networks when there was high covariance between species and when some of the species in the input dataset had a gradual, seasonal abundance pattern (Figs. S2–S4, Panels B and F).

GAM transformation improved network inference on average

The GAM transformation increased the F1 score of the Glasso networks in 82.3% of all network runs (Fig. 3, Panels A–C; most points fall below the 1:1 line). Specifically, the GAM transformation significantly increased the mean F1 score of the Glasso networks when a gradual seasonal signal (Fig. S1, Panel 3) was applied to some fraction of the species in the input dataset (Figs. S2–S4, Panels B, D, F, and H; Tables S2–S4). The GAM-Glasso networks also had significantly higher F1 scores when 50% of species in the input dataset had an abrupt, seasonal abundance pattern (Figs. S2–S4, Panels A, C, E, and G; Tables S2–S4). For the Barabási-Albert models, the F1 scores of the GAM-Glasso networks were significantly greater than those of the Glasso networks when 50% of the species in the input dataset had a long-term trend with



1.) CLR-transform species abundance data.

	Species 1	Species 2	Species 3
May	0	-1.38	-1.12
Jun	-0.83	1.02	3.32
Jul	0	0.64	0
Aug	0	-2.01	0
Sept	-1.21	1.14	4.13

2.) Model species abundance data as a function of month and day of time-series using GAM.

3.) Check to see if there is significant autocorrelation in GAM residuals.

4.) Use GAM residuals as input for network analysis.

5a.) Graphical lasso (Glasso)

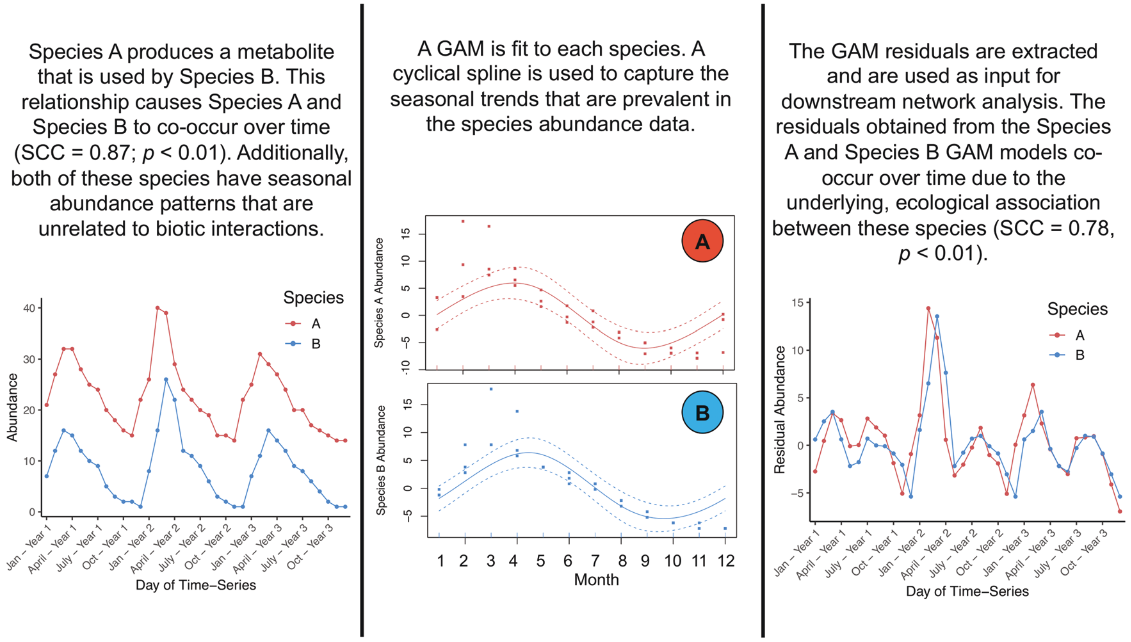
	Species 1	Species 2	Species 3
Species 1	0	0	1
Species 2	0	0	1
Species 3	1	1	0

5b.) Pairwise spearman correlation (SCC)

	Species. 1	Species 2	Species 3
Species 1	0	0.33	0.74
Species 2	0.33	0	0.66
Species 3	0.74	0.66	0

6.) Network outputs

Fig. 1 Steps used to carry out the GAM-based transformation of time-series species abundance data prior to carrying out pairwise spearman correlation (SCC) and graphical lasso (Glasso) ecological network analyses. The raw, species abundance data were first CLR-transformed (1). Generalized additive models (GAMs) were then fit to each species in the dataset (2) and the residuals of each GAM were checked for significant autocorrelation (3). The residuals of each GAM were extracted (4) and were used as input in the SCC and Glasso network analysis methods (5). Finally, the GAM-transformed network outputs were obtained (6; see text for additional details).



Species A produces a metabolite that is used by Species B. This relationship causes Species A and Species B to co-occur over time (SCC = 0.87; $p < 0.01$). Additionally, both of these species have seasonal abundance patterns that are unrelated to biotic interactions.

A GAM is fit to each species. A cyclical spline is used to capture the seasonal trends that are prevalent in the species abundance data.

The GAM residuals are extracted and are used as input for downstream network analysis. The residuals obtained from the Species A and Species B GAM models co-occur over time due to the underlying, ecological association between these species (SCC = 0.78, $p < 0.01$).

Fig. 2 A conceptual figure that demonstrates how the GAM transformation can remove seasonal signals while preserving ecologically relevant species co-occurrence patterns. In this example, the co-occurrence pattern between Species A and Species B persists even after the seasonal signals are removed by the GAM transformation.

no (0%) seasonality (Fig. S2, Panels E–H; Table S2). Similar increases in F1 score were noted in the Erdős-Rényi and American Gut dataset, GAM-Glasso networks when the covariance between species was low and when 50% of the species in the dataset had long-term changes in abundance with no (0%) seasonality (Figs. S3–S4, Panels G and H; Tables S3–S4).

The GAM transformation also led to substantial increases in the F1 score of the SCC networks. Overall, the GAM transformation increased the F1 score of SCC networks in 78.2% of all network runs (Fig. 3, Panels D–F; most points fall below the 1:1 line). The average F1 scores of the networks were significantly greater when the data were GAM-transformed prior to carrying out a SCC

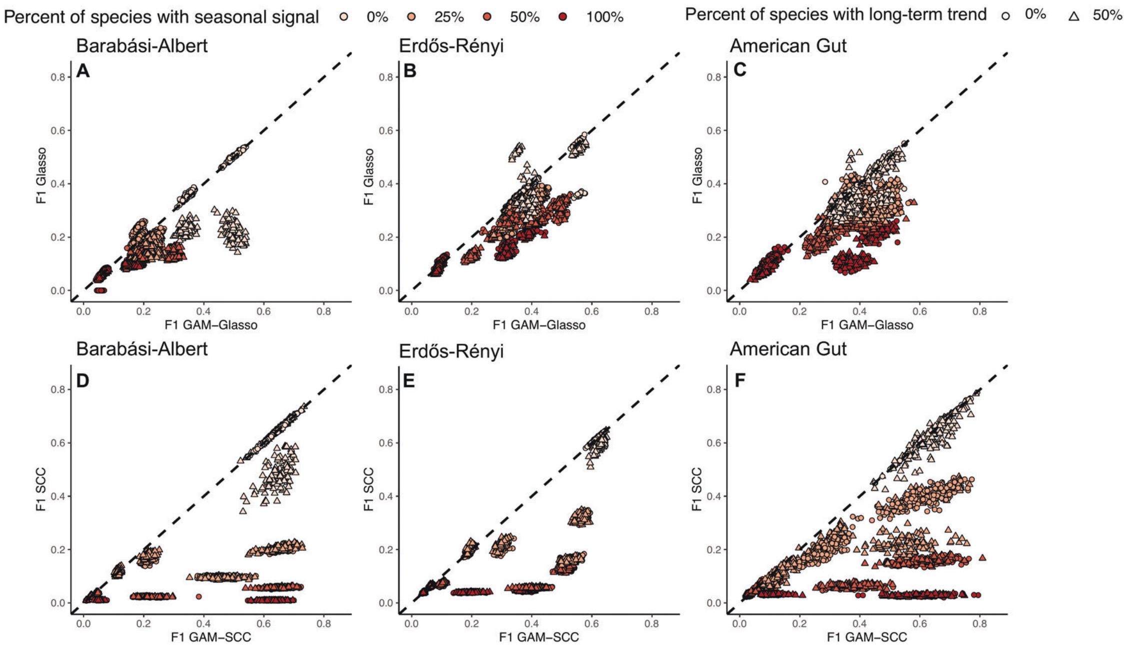


Fig. 3 F1 scores of the networks constructed without GAM transformation (*y*-axis) plotted against the F1 scores of the GAM-transformed networks (*x*-axis) for all of the mock time-series datasets that were simulated. A–C Show the comparison between the Glasso and the GAM-Glasso networks, while (**D**–**F**) show the comparison between the SCC and the GAM-SCC networks. The dashed, black lines show the 1:1 relationship. Data points below the 1:1 line depict network outputs that had a higher F1 score after applying the GAM-based data transformation, while data points that fall above the 1:1 line depict network runs that had a higher F1 score without applying the GAM-based data transformation prior to network construction.

network analysis when a gradual seasonal signal (Fig. S1, Panel 3) was applied to some fraction of the species in the input dataset (Figs. S5–S7, Panels B, D, F, and H; Tables S2–S4). The mean F1 scores of all GAM-SCC networks were also significantly greater than those of the SCC networks when there was high covariance between species and when an abrupt seasonal signal (Fig. S1, Panel 3) was applied to 25% or 50% of the species in the input dataset (Figs. S5–S7, Panels A and E; Tables S2–S4).

GAM transformation improved the ability of Glasso and SCC networks to capture real network structure

The GAM-transformation improved the ability of the Glasso and SCC methods to capture the underlying structure of the Barabási-Albert and Erdős-Rényi networks (Figs. 4–5). The real network degree distributions were more similar to the GAM-SCC and GAM-Glasso degree distributions than they were to the degree distributions of the SCC and Glasso networks without GAM transformation (Figs. 4–5). The GAM-SCC approach was the most successful in capturing the real, scale-free Barabási-Albert network degree distribution and had the highest average precision, recall, and F1 score of the four methods tested (Fig. 4; Panel C). Conversely, the GAM-Glasso approach did the best job of capturing the real, Erdős-Rényi network structure, as the SCC and GAM-SCC approaches predicted a number of high-degree nodes that were not present in the true network structures (Fig. 5; long right tail in Panels C and D). Some high-degree nodes were also predicted in the Glasso and GAM-Glasso, Erdős-Rényi networks (Fig. 5; long right tail in Panels A and B) but in general were less pronounced than those noted in the SCC and GAM-SCC degree distributions.

The average clustering coefficients of the GAM-Glasso networks were the most similar to the average clustering coefficients of the real networks, while the average clustering coefficients of the GAM-SCC networks were the least similar to those of the real networks (Figs. 4–5). The exaggerated average clustering coefficients of the SCC and GAM-SCC networks were to be expected, given the transitive nature of correlative relationships between variables. In general, the average clustering coefficient values that were obtained from the four network prediction approaches were substantially higher than the average clustering coefficient values of the real networks. These high, average clustering coefficient values implied that the outputs obtained from the network inference approaches resulted in networks that were “clumpier” than the real networks.

Effectiveness of the GAM transformation decreased when seasonal abundance patterns were abrupt

The GAM transformation slightly decreased the predictive power of a small subset of the Glasso networks when abrupt seasonal abundance patterns (Fig. S1, Panel 3) were prevalent in the input dataset (Figs. S2–S4, Panels A, C, E, and G; Tables S2–S4). Specifically, when the covariance between species was high and all of the species (100%) in the input dataset had an abrupt seasonal abundance pattern, the F1 scores of the Glasso networks were significantly greater than those of the GAM-Glasso networks, though the magnitude of the differences in performance were minor (Figs. S2–S4, Panel A; Tables S2–S4). The GAM transformation also decreased or did not alter the F1 score for a small number of the SCC networks when abrupt seasonal abundance patterns were noted in the input dataset. There were some statistically significant, but very small, decreases in the GAM-SCC F1 scores relative to the SCC F1 scores when 50% of the species in the input dataset had long-term increases or decreases in abundance without having seasonal abundance patterns (0%; Figs. S5 and S7, Panels E–F).

In sum, there was never a substantial decrease, and there was often a substantial increase, in network F1 score when applying the GAM transformation to the data before network inference.

DISCUSSION

Ecological co-occurrence networks can be useful for capturing complex, biotic interactions when applied to high-throughput

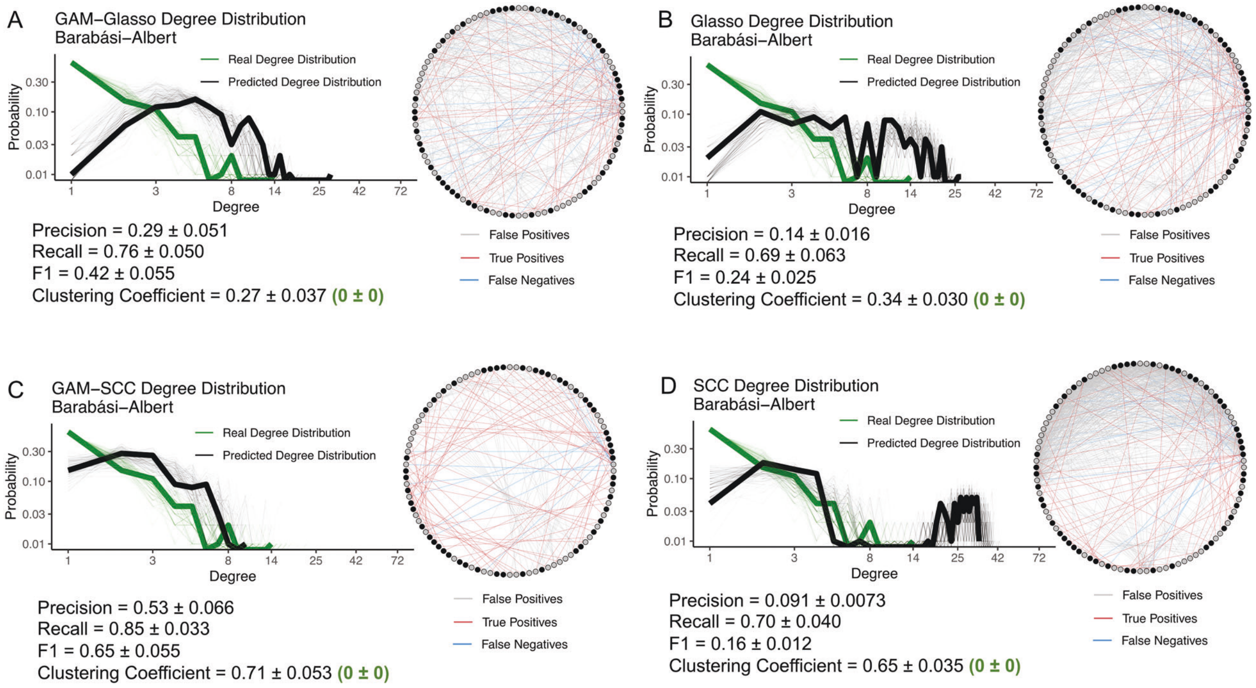


Fig. 4 The GAM-SCC networks did the best job of capturing the real, Barabási-Albert network degree distribution. The degree distributions and network outputs of 100 GAM-Glasso (**A**), Glasso (**B**), GAM-SCC (**C**), and SCC (**D**) time-series networks that contained 100 species. The networks depicted were constructed with mock species abundance data that had an underlying Barabási-Albert network structure and that contained 50 species with a gradual, seasonal abundance pattern. The fine green lines on the log-log plots show the real degree distributions and the fine black lines show the network-predicted degree distributions. The bolded lines show the degree distributions of the representative networks that are depicted. On the representative network images, the red edges show those edges that are true positive associations, the blue edges show those edges that are false negative associations, and the gray edges show those edges that are false positive associations. The black nodes in the network images represent the species that have a seasonal abundance pattern, while the gray nodes represent those species that do not have a seasonal abundance pattern.

sequencing datasets of microbial communities. However, networks can yield inaccurate associations if time-series properties are prevalent in the dataset [4, 22]. The performance of the SCC and Glasso methods used in this study tended to decrease as the number of species with a seasonal abundance pattern in a dataset increased (Fig. 3; Tables S2–S4). This finding indicates that time-series networks can result in a higher number of false positive and false negative associations than networks that are constructed with datasets that lack time-series features. It is likely that the false positive associations that were detected in our time-series network outputs were indirect associations that resulted from the shared periodicity of two or more species over time. In addition, true species-species associations may have been missed by our network runs (false negatives) if seasonal and long-term signals overpowered the influence that other organisms had on species abundance patterns.

The GAM transformation carried out prior to network construction in this study improved network inference on average for all of the networking methods tested (Fig. 3; Fig. S10). The generally higher F1 scores that were noted following GAM transformation suggest that the GAM model was able to successfully capture and remove many of the seasonal and long-term signals that were prevalent in the mock communities. Previous efforts have been made to account for indirect or time-dependent associations in network analyses. For example, the EnDED program (environmentally driven edge detection) uses environmental variables (e.g., temperature, salinity, etc.) to predict and remove indirect, environmentally driven associations in an ecological network after network inference has already been performed, provided that environmental metadata are available [51]. Time-ordered networks have also been used by researchers to account for time when conducting network analyses [22, 52]. In time-ordered networks, a node is created for each species at each time point in a dataset, thus allowing those networks to capture associations between species at specific points in time [52]. However, this approach makes no correction for seasonal or long-term trends. The GAM-based data transformation proposed here provides an effective tool that can be used to account for multiple time-series features prior to carrying out ecological network analyses, can be tailored to a wide variety of time-series datasets, and can be used in conjunction with any downstream networking method.

The SCC and Glasso network comparisons carried out in this study demonstrated that both of these methods have unique strengths and weaknesses and that method selection may depend on the research question(s) being asked. With Barabási-Albert networks, the GAM transformation improved the F1 score and the degree distribution of the SCC networks more than those of the Glasso networks (Fig. 4). It is known that correlation-based networks tend to capture both direct and indirect associations in an input dataset [4, 9, 17], while Glasso networks can avoid capturing indirect associations [31]. The notably higher F1 scores and improvements in the degree distribution plots that were observed in the GAM-SCC networks relative to the SCC networks are presumably due to ability of the GAM to remove many of the indirect associations that would otherwise be detected in the SCC network analyses. While the GAM-SCC approach captured the real, Barabási-Albert degree distribution better than the GAM-Glasso approach, the GAM-Glasso approach was better able to capture the real, Erdős-Rényi degree distribution (Fig. 5). In addition, the average clustering coefficients of the GAM-Glasso networks were the most similar to the real, Barabási-Albert and Erdős-Rényi networks (Figs. 4–5), suggesting that the GAM-Glasso networks were more similar to the real networks than the SCC networks in terms of network “clumpiness”.

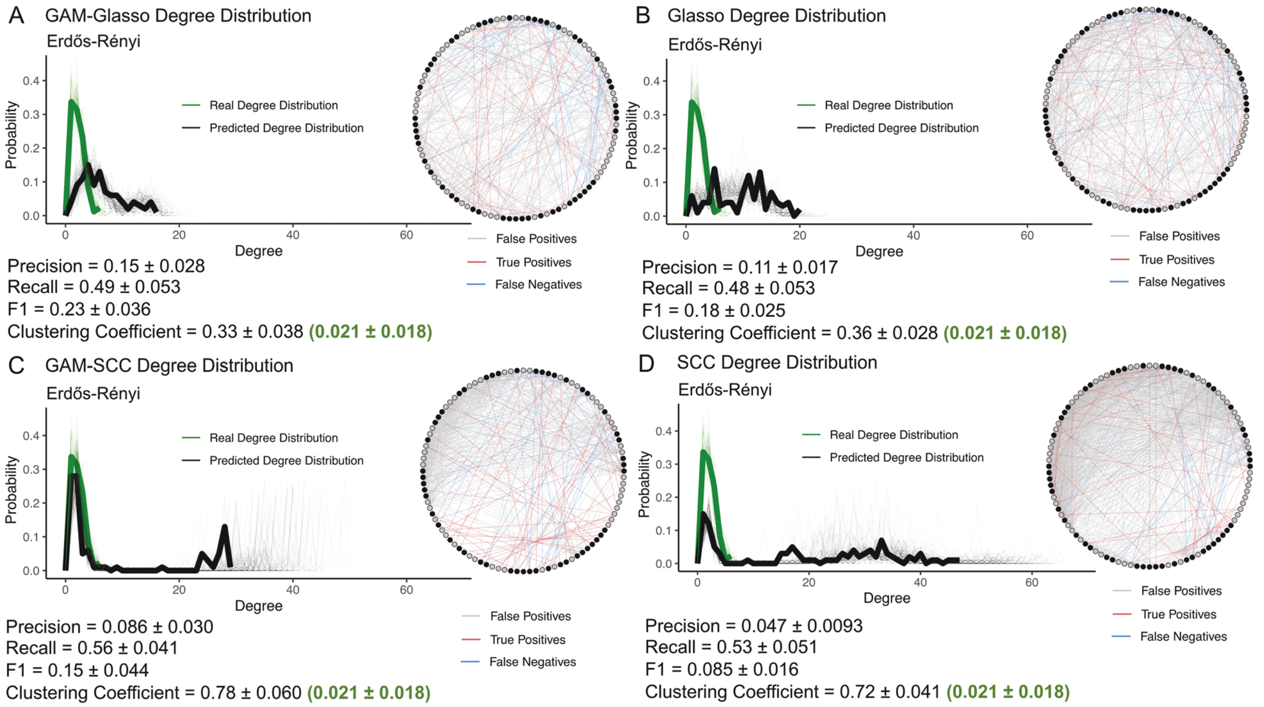


Fig. 5 The GAM-Glasso networks did the best job of capturing the real, Erdős-Rényi network topology. The degree distributions and network outputs of 100 GAM-Glasso (**A**), Glasso (**B**), GAM-SCC (**C**), and SCC (**D**) time-series networks that contained 100 species. The networks depicted were constructed with mock species abundance data that had an underlying Erdős-Rényi network structure and that contained 50 species with a gradual, seasonal abundance pattern. Panels and color coding are the same as described for Fig. 4.

The GAM transformation did not substantively improve network inference under certain zero-inflation scenarios (Fig. S8) and when total sum scaling and cumulative sum scaling normalization methods were used as opposed to a CLR transformation (Fig. S9). The GAM transformation also led to decreases in F1 score at times when abrupt seasonal abundance patterns were prevalent (Figs. S2–-S7). It is possible that the smoothing functions used in the GAMs were unable to capture the some of the periodic spikes in species abundance that were noted in the abrupt, seasonal abundance patterns and therefore did not fully remove these abrupt signals. It is also possible that the GAM transformation inadvertently removed the influence that other species in the dataset had on the abundance pattern of a specific species and therefore decreased the number of true positive associations detected in the networks created using the GAM-transformed data. These scenarios highlight some potential issues that may arise if the GAM model used in the transformation overfits the data or does not accurately capture the seasonal and long-term signals that may be prevalent in the abundance data of a specific organism. If the GAM model overfits the data, then meaningful co-occurrence patterns between organisms may be lost in the GAM residuals. Conversely, if the GAM model fails to capture the time-series signals that may be influencing species abundance patterns, then the network predictions will still be influenced by temporal trends as opposed to true species-species associations. The observation that the GAM method did not always improve model performance when seasonal abundance patterns were abrupt suggests an opportunity for future improvement. Generalized additive models are good at fitting smooth trends in the data, but other methods might be better at removing abrupt seasonal signals. In any event, decreases in the F1 scores under these specific conditions (Figs. S2–S7) were marginal relative to the benefits obtained using the GAM transformation, suggesting that the GAM models used in our analyses were generally effective at removing time-series signals.

It may be beneficial to explore whether our GAM transformation improves the performance of other network analysis tools. Extended local Similarity Analysis (eLSA) which identifies time-lagged associations [53] and Liquid Analysis (LA) which explores interactions between trios of variables [54, 55] would likely be improved by removing seasonal signals. Future efforts may also be aimed at incorporating batch effects into our GAM framework to account for some of the additional non-biological factors that influence species abundance patterns over time. Generating a method that can identify sparse networks, time-lagged associations and three-way interactions, while removing seasonal signals and other sources of non-biological variability would be a clear future direction for a robust and flexible analysis of high-throughput data.

CONCLUSION

The results of this study highlight the importance of considering temporal features when carrying out ecological network analyses with time-series data, given that time-dependent species abundance patterns may confound network predictions. The GAM-based data transformation presented here (NetGAM) provides a simple, yet effective tool that can be used to reduce the influence that time-series properties have on microbial abundance data prior to network construction. We published our method in a publicly available R package (https://github.com/sgleich/NetGAM) so that this data transformation can be used by other researchers in future time-series network analysis efforts. Accounting for seasonal abundance patterns, long-term trends, and autocorrelation in time-series datasets using our GAM method may substantially improve network inference. We recommend that future networking studies account for time-dependent species abundance patterns that may be prevalent in an input dataset in order to reduce the number of false positive and false negative associations that are detected through time-series network analyses.

REFERENCES

1. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10:538–50.
2. Moënne-Loccoz Y, Mavingui P, Combes C, Normand P, Steinberg C. Microorganisms and biotic interactions. In: Bertrand JC, Caumette P, Lebaron P, Matheron R, Normand P, Sime-Ngando T, editors. Environmental microbiology: fundamentals and applications. Dordrecht: Springer; 2015. p. 395–444.
3. Chaffron S, Rehrauer H, Pernthaler J, Von, Mering C. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res. 2010;20:947–59.
4. Carr A, Diener C, Baliga NS, Gibbons SM. Use and abuse of correlation analyses in microbial ecology. ISME J. 2019;13:2647–55.
5. Layeghifard M, Hwang DM, Guttman DS. Disentangling interactions in the microbiome: a network perspective. Trends Microbiol. 2017;25:217–28.
6. Matchado MS, Lauber M, Reitmeirer S, Kacprowski T, Baumbach J, Haller D, et al. Network analysis methods for studying microbial communities: a mini review. Comput Struct Biotechnol. 2021;19:2687–2698.
7. Barner AK, Coblentz KE, Hacker SD, Menge BA. Fundamental contradictions among observational and experimental estimates of non-trophic species interactions. Ecology. 2018;99:557–66.
8. Freilich MA, Wieters E, Broitman BR, Marquet PA, Navarrete SA. Species co-occurrence networks: can they reveal trophic and non-trophic interactions in ecological communities? Ecology. 2018;99:690–9.
9. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. 2016;10:1669–81.
10. Eiler A, Hayakawa DH, Rappé MS. Non-random assembly of bacterioplankton communities in the subtropical North Pacific Ocean. Front Microbiol. 2011;2:140.
11. Steele JA, Countway PD, Xia L, Vigil PD, Beman JM, Kim DY, et al. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. ISME J. 2011;5:1414–25.
12. Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B, et al. Defining seasonal marine microbial community dynamics. ISME J. 2012;6:298–308.
13. Chow CET, Sachdeva R, Cram JA, Steele JA, Needham DM, Patel A, et al. Temporal variability and coherence of euphotic zone bacterial communities over a decade in the Southern California Bight. ISME J. 2013;7:2259–73.
14. Cram JA, Xia LC, Needham DM, Sachdeva R, Sun F, Fuhrman JA. Cross-depth analysis of marine bacterial networks suggests downward propagation of temporal changes. ISME J. 2015;9:2573–86.
15. Deutschmann I, Krabberød A, Benites L, Latorre F, Delage E, Marrasé C, et al. Disentangling temporal associations in marine microbial networks. Preprint at https://doi.org/10.21203/rs.3.rs-404332/v1 (2021).
16. Karl DM, Church MJ. Microbial oceanography and the Hawaii Ocean Time-series programme. Nat Rev Microbiol. 2014;12:699–713.
17. Fuhrman JA, Cram JA, Needham DM. Marine microbial community dynamics and their ecological interpretation. Nat Rev Microbiol. 2015;13:133–46.
18. Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S. Annually reoccurring bacterial communities are predictable from ocean conditions. PNAS. 2006;103:13104–9.
19. Martin-Platero AM, Cleary B, Kauffman K, Preheim SP, McGillicuddy DJ, Alm EJ, et al. High resolution time series reveals cohesive but short-lived communities in coastal plankton. Nat Commun. 2018;9:1–11.
20. Comeau AM, Li WK, Tremblay J-É, Carmack EC, Lovejoy C. Arctic Ocean microbial community structure before and after the 2007 record sea ice minimum. PLoS ONE. 2011;6:e27492.
21. Giovannoni SJ, Vergin KL. Seasonality in ocean microbial communities. Science. 2012;335:671–6.
22. Blonder B, Wey TW, Dornhaus A, James R, Sih A. Temporal dynamics and network analysis. Methods Ecol Evol. 2012;3:958–72.
23. Scheiner SM, Cox SB, Mittelbach GG, Osenberg C, Kaspari M. Species richness, species–area curves and Simpson's paradox. Evol Ecol Res. 2000;2:791–802.
24. Armitage DW, Jones SE. How sample heterogeneity can obscure the signal of microbial interactions. ISME J. 2019;13:2639–46.
25. Ravindra K, Rattan P, Mor S, Aggarwal AN. Generalized additive models: Building evidence of air pollution, climate change and human health. Environ Int. 2019;132:104987.
26. Cram JA, Chow CET, Sachdeva R, Needham DM, Parada AE, Steele JA, et al. Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. ISME J. 2015;9:563–80.
27. Murphy RR, Perry E, Harcum J, Keisman J. A generalized additive model approach to evaluating water quality: Chesapeake Bay case study. Environ Model Softw. 2019;118:1–13.
28. Otto SA, Kadin M, Casini M, Torres MA, Blenckner T. A quantitative framework for selecting and validating food web indicators. Ecol Indic. 2018;84:619–31.
29. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS ONE. 2012;8:e1002687.
30. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9:432–41.
31. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol. 2015;11:e1004226.
32. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al. American Gut: an open platform for citizen science microbiome research. mSystems. 2018;3:e00031–18.
33. Kurtz ZD, Müller CL, Miraldi E, Bonneau R. SpiecEasi: Sparse inverse covariance for ecological statistical inference. R package version 1.0.7. 2020;1.
34. Csardi G, Nepusz T. The igraph software package for complex network research. Int J Complex. 2006;1695:1–9.
35. Pan AY. Statistical analysis of microbiome data: the challenge of sparsity. Curr Opin Endocr Metab. 2021;19:35–40.
36. Espinoza JL, Shah N, Singh S, Nelson KE, Dupont CL. Applications of weighted association networks applied to compositional data in biology. Environ Microbiol. 2020;22:3020–38.
37. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere". PNAS. 2006;103:12115–20.
38. Caron DA, Countway PD. Hypotheses on the role of the protistan rare biosphere in a changing world. Aquat Microb Ecol. 2009;57:227–38.
39. Van den Boogaart KG, Tolosana-Delgado R. "Compositions": a unified R package to analyze compositional data. Comput Geosci. 2008;34:320–38.
40. Yoon G, Gaynanova I, Müller CL. Microbial Networks in SPRING - Semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. Front Genet. 2019;10:516.
41. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat. Methods. 2013;10:1200–2.
42. Peschel S, Müller CL, von Mutius E, Boulesteix A-L, Depner M. NetCoMi: network construction and comparison for microbiome data in R. Brief Bioinform. 2020;22:1–18.
43. Wood SN. Stable and efficient multiple smoothing parameter estimation for generalized additive models. J Am Stat Assoc. 2004;99:673–86.
44. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc: Series B (Statistical Methodology). 2011;73:3–36.
45. Fang H, Huang C, Zhao H, Deng M. CCLasso: correlation inference for compositional data through Lasso. Bioinformatics. 2015;31:3172–80.
46. Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L. The huge package for high-dimensional undirected graph estimation in R. J Mach Learn Res. 2012;13:1059–62.
47. Müller CL, Bonneau R, Kurtz Z. Generalized stability approach for regularized graphical models. arXiv preprint *arXiv*:160507072. Retrieved from http://arxiv.org/abs/1605.07072 (2016).
48. Liu H, Roeder K, Wasserman L. Stability approach to regularization selection (stars) for high dimensional graphical models. Adv Neural Inf Process Syst. 2010;24:1432.
49. Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A. The architecture of complex weighted networks. PNAS. 2004;101:3747–52.
50. Poisot T, Gravel D. When is an ecological network complex? Connectance drives degree distribution and emerging network properties. PeerJ. 2014;2:e251.
51. Deutschmann IM, Lima-Mendez G, Krabberød AK, Raes J, Vallina SM, Faust K, et al. Disentangling environmental effects in microbial association networks. Microbiome. 2021;9:232.
52. Blonder B, Dornhaus A. Time-ordered networks reveal limitations to information flow in ant colonies. PLoS ONE. 2011;6:e20298.
53. Xia LC, Steele JA, Cram JA, Cardon ZG, Simmons SL, Vallino JJ, et al. Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. BMC Syst Biol. 2011;5:1–12.
54. Li K-C. Genome-wide coexpression dynamics: theory and application. PNAS. 2002;99:16875–80.
55. Ai D, Li X, Pan H, Chen J, Cram JA, Xia LC. Explore mediated co-varying dynamics in microbial community using integrated local similarity and liquid association analysis. BMC Genom. 2019;20:117–28.

ACKNOWLEDGEMENTS

This work and SJG were supported by the Simons Foundation Grant P49802 (to DAC). JLW was supported by a postdoctoral fellowship Simons Foundation Award 653212. We thank Ben Tully and the members of the Bioinformatics Virtual Coordination Network (BVCN) for creating and contributing to the online platform that inspired this project.

AUTHOR CONTRIBUTIONS

SJG, JAC, and JLW developed and validated the data transformation approach. SJG wrote the paper with guidance from DAC, JAC, and JLW. All authors contributed to the final paper.

COMPETING INTERESTS

This work was supported by Simons Foundation Grant P49802 (to DAC).

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43705-022-00106-7.

Correspondence and requests for materials should be addressed to Samantha J. Gleich.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.