Preregistration

LDP Mock Preregistration

Kirsten Bevandick[1], Kate Colson[1], Stefano Mezzini[2], William Ou[1], Julee Stewart[3]

[1] University of British Columbia

[2] University of British Columbia (Okanagan campus)

[3] University of Regina

2021-10-06

Citations in parentheses: (Gushulak *et al.*, 2021), (but also, see Gushulak *et al.*, 2021)

Citation out of parentheses: Gushulak *et al.* (2021) state that…

See what references you can add by opening the .bib file in a RStudio and using the name after `@{article` (e.g. `@article{gushulak_effects_2021`), or by running `citr:::insert_citation()` (requires installing `citr`).

NOTE: to knit using a bib file, you will need to install a different version of `prereg` using `remotes::install_github("crsh/prereg@issue-16").` not sure why this happens. – Stefano

Roles

1. **Research question + hypothesis** (William)
    - clearly identify the research question of interest in the replication?
    - include at least 1 testable hypothesis?

2. **Data** (Julee, Kate)

   - Description of existing data and/or data collection procedures? (If existing data is included, is the reference(s) to the original data source included?)
   - A description of the variables included in the dataset and/or to be included in the analysis
   - A study design plan?

3. **Analysis**

   - Does the pre-registration include at least 1 example statistical analysis using simulated/dummy data?
   - Are the simulated data informed by published data?

4. **Figure** (William)

   - Does the pre-registration include a figure?
   - summarize/present the key variables in the analysis (with appropriate response and predictor variables)?
   - Include properly labelled axes and a legend (if applicable)?
   - Include a figure caption

5. **Literature** (Julee)

   - Does the pre-registration include in-text citations and a bibliography of all studies mentioned?

Study Information

Title LDP Mock Preregistration

Description Enter your response here.

Hypotheses

Enter your response here.

Design Plan

Study type

Experiment. A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

Observational Study. Data is collected from study subjects that are not randomly assigned to a treatment. This includes surveys, natural experiments, and regression discontinuity designs.

Meta-Analysis. A systematic review of published studies.

Other. Please explain.

Blinding

No blinding is involved in this study.

For studies that involve human subjects, they will not know the treatment group to which they have been assigned.

Personnel who interact directly with the study subjects (either human or non-human subjects) will not be aware of the assigned treatments.

Personnel who analyze the data collected from the study are not aware of the treatment applied to any given group.

Study design

Enter your response here.

Randomization

Enter your response here.

Sampling Plan

Existing data

Registration prior to creation of data. As of the date of submission of this research plan for preregistration, the data have not yet been collected, created, or realized.

Explanation of existing data

Enter your response here.

Data collection procedures

Enter your response here.

Sample size

Enter your response here.

Sample size rationale

Enter your response here.

Stopping rule

Enter your response here.

Variables

Manipulated variables

Depth

For phytobenthos analyses, we will manipulate the depth that surface sediment samples are collected. Three sediment samples will be collected for every 1-m depth interval of Gall Lake. Samples for this categorical variable will range from a 1-m to 16-m depth and will be determined with the use of a depth sounder. In R scripts, this manipulated variable will be named "depth_m."

Measured variables

Stable Isotopic Analysis

The single outcome variables for the stable isotopic ratio of nitrogen and carbon will be measured from freeze-dried sediment subsamples of Gall Lake. Samples will be placed in a Thermo Finnigan Delta V isotope ratio mass spectrometer that has a ConFlow IV dilution inlet system as described by @gushulak_effects_2021. Following the calibration procedure for laboratory standards explained in Bunting et al. (2010) and Savage et al. (2004), the isotope values will be analyzed with the use of atmospheric gas. The standard notation for the stable isotopic ratio of nitrogen and carbon are 15N and 13C, respectively. In R scripts, these single outcome variables will be named "d15n" and "d13c" for nitrogen and carbon, respectively. *how do we use symbols in R markdown?*

Indices

Nitrogen and Carbon Content

Using generalized additive models, we will be determining the percent content of nitrogen and carbon in the depth intervals of Gall lake by manipulating the stable isotope content of these elements. Specifically, gamma distributions as described by Mushet et al. (2019) will be used to determine the percent content. In R scripts, the manipulated variables will be named "perc_n" and "perc_c" for nitrogen and carbon, respectively. *equation?*

Analysis Plan

Statistical models

Similarly to Gushulak *et al.* (2021), the isotope data was analyzed using Generalized Additive Models (GAMs) via the `mgcv` package (Wood, 2011, 2017). The amount of ^{15}N and ^{1}3C (`d15n` and `d13c` in the `R` scripts, respectively) were modeled using a Gaussian conditional distribution with an *identity* link function since both parameters take any real value (i.e. they can be both positive and negative). The percentages of nitrogen and carbon in the samples (`perc_n` and `perc_c`, respectively) were converted to proportions (`frac_n` and `frac_c`) so they coould be modeled with a beta distribution with a *logit* link function. (There is no distribution in the `mgcv` package for numbers between 0 and 100). Finally, the proportion of carbon to nitrogen (C:N, `c_n_ratio` in the scripts) were modeled using a GAM with a gamma conditional distribution and a *log* link function. A gamma distribution was most appropriate since the ratio of two positive (non-zero) numbers is strictly greater than zero. The mean effects and their Bayesian credible intervals (CIs) were estimated on the link scale using normal approximation (±1.96 standard deviations) and back-transformed to response values using the inverse-link function. For example, the mean C:N ratios and their CIs was estimated on the *log* scale and then back-transformed to ratios by exponentiating the estimate. (See the following section for more information on GAMs and transformations.)

All five models accounted for a shared global trend between replicates and trends within-replicate. For simplicity, the smoothness parameter was assumed to be the same between replicates (see model "GS" in Pedersen *et al.*, 2019).

```
# stable nitrogen isotope
m_d15n <- gam(d15n ~
                s(depth_m, k = 15) + # average effect of depth
                s(depth_m, replicate, k = 10, bs = 'fs'), # effect of replicate
              family = gaussian(link = 'identity'), # I(k) = k
              data = isotopes,
              method = 'REML') # optimiziation method for the smoothness parameter

# stable carbon isotope
m_d13c <- gam(d13c ~
                s(depth_m, k = 15) +
                s(depth_m, replicate, k = 10, bs = 'fs'),
              family = gaussian(link = 'identity'),
              data = isotopes,
              method = 'REML')

# fraction of carbon in the samples
m_frac_c <- gam(frac_c ~
                  s(depth_m, k = 15) +
                  s(depth_m, replicate, k = 10, bs = 'fs'),
                family = betar(link = 'logit'), # logit(0) = -Inf, logit(1) = Inf
                data = isotopes,
```

```r
                method = 'REML')

# fraction of nitrogen in the samples
m_frac_n <- gam(frac_n ~
                  s(depth_m, k = 15) +
                  s(depth_m, replicate, k = 10, bs = 'fs'),
                family = betar(link = 'logit'),
                data = isotopes,
                method = 'REML')

#C:N ratio in the samples
m_c_n <- gam(c_n_ratio ~
               s(depth_m, k = 15) +
               s(depth_m, replicate, k = 10, bs = 'fs'),
             family = Gamma(link = 'log'), # log(0) = -Inf, log(Inf) = Inf
             data = isotopes,
             method = 'REML')
```

Thus, each model had two predictors: a predictor which accounted for the mean effect of lake depth (`s(depth_m, k = 15)`) and one which used factor smooths (`bs = 'fs'`) for each replicate. Both of the predictors used thin plate regression splines (the default for the `s()` function in `mgcv`). The number of knots (`k`) for each smooth was allowed to be reasonably high, since the models were fit via penalized maximum likelihood and thus over-fitting the data was unlikely (Wood, 2011; Simpson, 2018). The smoothness parameters were optimized using Restricted Marginal Likelihood (`method = 'REML'`), rather than the default Generalized Cross Validation (GCV), since REML does not over-fit as often as GCV (Reiss & Ogden, 2009; Wood, 2011).

Although credible intervals were estimated, this study was performed with a purely Bayesian approach, so there was no interest in statistical significance and Frequentist null-hypothesis testing. Rather, the models were used to compare similarities between our estimates and the estimates by Gushulak *et al.* (2021).

Transformations

The only data transformation that was performed was the conversion of percent carbon and nitrogen (between 0% and 100%) to proportions (between 0 and 1). Note that since the transformation $a = \frac{b}{100}$ is a linear transformation, Jensen's inequality does not apply here (Jensen, 1906).

Modeling the data with GAMs removes the need for transforming data which violates the normality assumptions which linear models depend on: GAMs estimate transformed mean responses ($g[\mathbb{E}(Y)]$, where $g(\cdot)$ is the link function) rather than the mean transformed response ($\mathbb{E}[g(Y)]$), so Jensen's inequality (Jensen, 1906) does not apply here. ## Inference criteria

Data exclusion

Enter your response here.

Missing data

Enter your response here.

Exploratory analyses (optional)

N/A

Other

Other (Optional)

Enter your response here.

References

Gushulak C.A.C., Haig H.A., Kingsbury M.V., Wissel B., Cumming B.F. & Leavitt P.R. (2021). Effects of spatial variation in benthic phototrophs along a depth gradient on assessments of whole-lake processes. *Freshwater Biology*, fwb.13820. https://doi.org/10.1111/fwb.13820

Jensen J.L.W.V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica* **30**, 175–193. https://doi.org/10.1007/BF02418571

Pedersen E.J., Miller D.L., Simpson G.L. & Ross N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ* **7**, e6876. https://doi.org/10.7717/peerj.6876

Reiss P.T. & Ogden T.R. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 505–523. https://doi.org/10.1111/j.1467-9868.2008.00695.x

Simpson G.L. (2018). Modelling Palaeoecological Time Series Using Generalised Additive Models. *Frontiers in Ecology and Evolution* **6**, 149. https://doi.org/10.3389/fevo.2018.00149

Wood S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* **73**, 3–36

Wood S.N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd edn. Chapman; Hall/CRC.