Quantifying Impacts of an Environmental Intervention Using Environmental DNA: Supplemental Text 2

Elizabeth Andruszkiewicz Allan, Ryan P. Kelly, Erin D'Agnese, Maya Garber-Yonts, Megan Shaffer, Zachary Gold, Andrew O. Shelton

2022

Our analysis depends upon a set of quantitative models, each linking our observations of metabarcoding reads or qPCR cycle-threshold values to an underlying concentration of target-species DNA in water samples.

In summary, we (1) use a mock community with a known composition to calibrate our environmental metabarcoding data as described in Shelton et al. 2022. The result is a set of estimated proportions of DNA from each species in each sample. We then (2) relate qPCR cycle-threshold values for a reference species (here, *O. clarkii*) from the same set of samples to a standard curve to yield quantitative estimates of the concentration of our reference species in each sample. We (3) use these absolute estimates of DNA concentration to expand the metabarcoding-derived proportion data into a complete set of quantitative estimates of DNA concentrations for each species in each sample. Finally, we (4) construct a time-series model for these species-specific concentrations, sharing information across creeks and time-points. This allows us to interpolate unobserved data points and more important, to compare our observations to the (counterfactual) expecations for species' DNA concentrations in the absence of a construction project. We detail the statistical details of these steps below.

Calibration with a Mock Community

See Shelton et al. 2022; McLaren et al; Silverman et al

qPCR Calibration

See (Shelton et al. 2019) and (McCall et al. 2014) for similar analyses.

For all samples i, on qPCR plates j, we either observe ($z_{i,j} = 0$ or do not observe $z_{i,j} = 1$) amplification; we omit the subscripts i and j from the following description except where necessary for clarity. We assume an intercept of zero.

We model the probability of detection $P(z = 1))$ as a linear function of concentration and slope parameter ϕ, ($P(z = 1) = \theta = c\phi$), with a logit transform to constrain the inferred probability to between 0 and 1.

For those samples that amplify ($z = 1$), we model the observed Ct value (y) as a linear function of our parameter of interest, the log-concentration of target-species DNA under analysis (c). We treat y as drawn from a normal distribution $y \sim N(\mu_{i,j}, \sigma_{i,j}))$, where each triplicate sample on each qPCR plate has its own estimated mean and standard deviation. The means are estimated as a straightforward linear model, $mu = \beta_{0,j} + \beta_{1,j}c$, but we allow the standard deviation to vary as a linear function of log-concentration so as to accurately capture decreasing precision with decreasing concentration: $\sigma = e^{\gamma_0 + \gamma_{1,j}c}$; we estimate these parameters as an exponent to constrain $\sigma > 0$.

Samples with known concentrations (i.e., standards) were fit jointly with unknown samples (i.e., environmental samples); because qPCR plate identity was shared among all environmental samples and standards within a plate, this has the effect of applying plate-specific slope and intercept values for the standard curve to each of the environmental samples on the plate.

We apply moderately informative priors that make use of background information in hand. For example, because qPCR standard curves of all kinds have slopes near -3, this slope becomes our background expectation as embodied in the prior on β_1, but the standard deviation of that prior leaves plenty of room for this background to be overwhelmed by the observed data. The same logic applies to the intercept of the standard curve, which in qPCR (for any given species) generally falls near 39 cycles, an expectation that we formalize by having β_0 drawn from a normal distribution with $\mu = 39$ and $\sigma = 3$.

Taken together with priors, the model is:

$$z_{i,j} \sim bernoulli(\theta_{i,j})$$

$$\theta_{i,j} = logit^{-1}(\phi * c_{i,j})$$

$$y_{i,j} \sim normal(\mu_{i,j}, \sigma_{i,j}) \text{ if } z_{i,j} = 1$$

$$\mu_{i,j} = \beta_{0,j} + \beta_{1,j} * c_{i,j}$$

$$\sigma_{i,j} = e^{\gamma_0 + \gamma_{1,j} * c_{i,j}}$$

$$\beta_0 \sim normal(39, 3)$$

$$\beta_1 \sim normal(-3, 1)$$

$$\gamma_1 \sim normal(0, 5)$$

$$\gamma_0 \sim normal(-2, 1)$$

Model Diagnostics: 3 chains, 2500 iterations, for all parameters, $\hat{R} \leq 1.002$.

Expanding Proportions into Absolute Abundances

As described in the main text, calibrated metabarcoding analysis yielded quantitative estimates of the proportions of species' DNA in environmental samples prior to PCR.

We then converted these proportions into absolute abundances by expansion, in light of the qPCR results for our reference species *O. clarkii*. We estimated the total amplifiable salmonid DNA in environmental sample i as $DNA_{salmonid_i} = \frac{[qPCR_{reference_i}]}{Proportion_{reference_i}}$, and then expanded species' proportions into absolute concentrations by multiplying these sample-specific total concentrations by individual species' proportions, such that for species j in sample i, $DNA_{i,j} = DNA_{salmonid_i} * Proportion_{i,j}$.

See Pont et al. 2022; McClaren 2022 pre-print

Time-Series Model

At a given station in a given creek, there is some distribution of DNA concentration for a species. For simplicity, we focus on a single species and a single station (downstream or upstream) for the moment.

The (log) DNA concentration in creek i at time t is distributed as $Y_{i,t} \sim \mathcal{N}(\mu_{i,t}, \sigma^2)$. We may choose to let σ vary across creeks, time points, or with a covariate such as creek flow.

We are interested in how the DNA concentration changes over time, so we assert that the expected value of DNA in a creek at time t, $\mu_{i,t}$, depends upon its value in the previous time step $t-1$, in some way. Further, we can let $\mu_{i,t}$ in, say, our focal Padden creek, depend upon the observations in other creeks (i.e., where creek $i \neq$ Padden) if we think that similar environmental and demographic forces are affecting all creeks in similar ways. We can use these inferences to model data we cannot observe directly – namely, a counterfactual scenario in which a human intervention did not occur – to estimate the effect of that intervention.

We use a simple, first-order autoregressive (AR(1)) model with $\mu_{i,t}$ as a linear function of $\mu_{i,t-1}$ with slope β and intercept α. Here, β reflects the degree of autocorrelation between time steps t and $t-1$; a stationary model requires $|\beta| \leq 1$. α estimates the shift in the mean, after accounting for autocorrelation, at a given creek and timepoint.

To share information across creeks, we can assert a constant β for all creeks within a timepoint – that is, the abundance of each species' DNA at a given timepoint is similarly dependent upon its abundance at the prior timepoint. Our model would then look like this:

$$Y_{i,t} \sim \mathcal{N}(\mu_{i,t}, \sigma^2)$$

$$\mu_{i,t} = \alpha_{i,t} + \beta_t \mu_{i,t-1}$$

where the slope term, β, is shared across creeks for a given time point.

We can add observations from many species and from the two stations per creek – upstream and downstream of the culvert – simply by adding subscripts to the model. If we let d be a subscript indicating station ($d = 1$ if downstream, $d = 2$ if upstream), and let j be a subscript indicating species across the same set of samples, we have a single overall model of the change in eDNA concentration among species, creeks, timepoints, and stations.

We then add a term, γ, to explicitly estimate the effect of culvert removal. We index γ with an index r reflecting the state of a creek as either being in its undisturbed state ($r = 1$) or else subject to restoration ($r = 2$; only Padden Creek has this designation, and only after October 2021). We estimate γ for each species j and each timepoint t.

Finally, we add a term, η, to capture the additional variation in DNA concentration not otherwise explained by the autocorrelation element of the model. Differences between η for upstream and downstream stations within a set of time/creek/species observations reflect a combination of differences due to the culvert itself and random process variation.

We complete the model by specifying the prior distributions from which each parameter is drawn, selecting weakly informative priors for each parameter.

$$
\begin{aligned}
Y_{i,t,d,j} &\sim \mathcal{N}(\mu_{i,t,d,j},\, \sigma^2) \\
\mu_{i,t,d,j} &= \alpha_{i,t,j} + \beta_j \mu_{i,t-1,d,j} + \gamma_{t,j,r} + \eta_{i,t,d,j} \\
\alpha_{i,j,t} &\sim \mathcal{N}(\mu_{\alpha_j}, \sigma_\alpha) \\
\beta &\sim \mathcal{N}(0, 5) \\
\gamma &\sim \mathcal{N}(0, 5) \\
\sigma &\sim gamma(1, 1) \\
\eta &\sim \mathcal{N}(0, \sigma_\eta) \\
\sigma_\eta &\sim gamma(1, 1) \\
\mu_\alpha &\sim \mathcal{N}(0, 5)
\end{aligned}
$$

To reflect (in part) the hierarchical structure of our data, we let our intercept terms, α be drawn from species-specific distributions, where each species has a different mean, but all species share a common variance.

The η terms are all drawn from a common distribution, representing variation among triplicate biological observations at a creek/time/station.

Note that the time-series model treats DNA concentrations at time zero, μ_0, as a parameter to be estimated freely from the observed data; all subsequent concentrations are a function of the concentration at the previous timestep. Accordingly, we assign a weakly informative prior on μ_0 as well, $\mu_0 \sim \mathcal{N}(0, 5)$. This prior reflects the prior belief that the DNA concentration for each species is between $4.5 * 10^-5$ and $2.2 * 10^4$ copies/L with 95% probability.

The η term gives us a way of estimating the effects of the culverts themselves on each species, after subtracting out the effects of autocorrelation, and other modeled parameters. The difference between upstream and downstream values of *eta* for a given species/creek/time yields our estimate of this effect.

This model shares enough information across time points (within a creek) and across creeks (within a time point) that we can use it to infer DNA concentrations that we do not actually observe – we treat the temporal/spatial points to be inferred as missing data, parameters to be estimated by the larger model.

[insert example figure]

Model Diagnostics: 3 chains, 2500 iterations, for all parameters,

References

McCall, Matthew N., Helene R. McMurray, Hartmut Land, and Anthony Almudevar. 2014. "On Non-Detects in qPCR Data." *Bioinformatics* 30 (16): 2310–16. https://doi.org/10.1093/bioinformatics/btu239.

Shelton, Andrew Olaf, Ryan P. Kelly, James L. O'Donnell, Linda Park, Piper Schwenke, Correigh Greene, Richard A. Henderson, and Eric M. Beamer. 2019. "Environmental DNA Provides Quantitative Estimates of a Threatened Salmon Species." *Biological Conservation* 237 (September): 383–91. https://doi.org/10.1016/j.biocon.2019.07.003.