Appendix B: Parameter Distributions

Distribution of the $\hat{\sigma}^2_{slp}$ estimate, Eq. 2

Consider for the moment, an idealized $\hat{\sigma}^2_{slp}$ estimate using subsampled data to eliminate overlap in the N_{t+t}/N_t ratios and L=1. Let's call it $\hat{\sigma}^2_{islp}$. We can derive the distribution of $\hat{\sigma}^2_{islp} / \sigma^2_p$ by observing that the slope of var(ln(N_{t+t})-ln(N_t)) versus t (t = 1,2,…, t') is basically

$$\hat{\sigma}^2_{islp} = \left(\hat{\sigma}^2_{t'} - \hat{\sigma}^2_1\right)/(t'-1)$$
$$\text{where } \hat{\sigma}^2_i = \text{var}(\ln(N_{t+i}) - \ln(N_t)) \text{ for } t = 1, i+2, 2i+3, \ldots, ni+n+1$$

since the var(ln(N_{t+t})-ln(N_t)) versus t line is generally a straight. Using $\frac{\hat{\sigma}^2_t}{\sigma^2_p} \sim \frac{\sigma^2_t}{df_t \sigma^2_p} \chi^2_{df_t} \sim g\left(\frac{df_t}{2}, \frac{\sigma^2_t}{\sigma^2_p}\frac{2}{df_{t'}}\right)$ where $g(\alpha, \beta)$ is a gamma distribution with shape α and scale β, the distribution of $\hat{\sigma}^2_{islp} / \sigma^2_p$ is straight-forward to derive.

$$\frac{\hat{\sigma}^2_{islp}}{\sigma^2_p} = \frac{\left(\hat{\sigma}^2_{t'} - \hat{\sigma}^2_1\right)/(t'-1)}{\sigma^2_p} \sim g\left(\frac{df_{t'}}{2}, \frac{2\sigma^2_{np} + t'\sigma^2_p}{(t'-1)\sigma^2_p}\frac{2}{df_{t'}}\right) - g\left(\frac{df_1}{2}, \frac{2\sigma^2_{np} + \sigma^2_p}{(t'-1)\sigma^2_p}\frac{2}{df_1}\right)$$
$$\text{where } \hat{\sigma}^2_i = \text{var}(\ln(N_{t+i} / N_t)) \text{ for } t = 1, i+2, 2i+3, \ldots, ni+n+1$$
$$\text{and } \text{E}(\hat{\sigma}^2_i) = 2\sigma^2_{np} + i\sigma^2_p \qquad \text{Eq. A1}$$
$$df_{t'} = \text{number of } \ln(N_{t+t'} / N_t) \text{ ratios minus 1}$$
$$df_1 = \text{number of } \ln(N_{t+1} / N_t) \text{ ratios minus 1}$$

Note that the sequential $N_{t+t'} / N_t$ ratios are chosen so that there is no overlap thus each ratio is independent. Assume for the moment, that the two gamma distributions are independent -- which they are not. In this case, we can show that the limiting distribution of Eq. A1 as $df_{t'}$ and df_1 become large is χ^2 with $\gamma\, df_{t'}$ degrees of freedom:

The moment generating function of $g(\alpha_1, \beta_1) - g(\alpha_2, \beta_2)$ is $(1-\beta_1 t)^{-\alpha_1}(1+\beta_2 t)^{-\alpha_2}$. Thus the moment generating function for the distribution in Eq. A1 is

$$\text{mgf} = \left(1-\frac{4V+2\boldsymbol{t}'}{(\boldsymbol{t}'-1)df_{t'}}t\right)^{-df_{t'}/2}\left(1+\frac{4V+2}{(\boldsymbol{t}'-1)df_1}t\right)^{-df_1/2} \quad \text{where } V=\boldsymbol{s}^2_{np}/\boldsymbol{s}^2_p.$$

Take the natural log of this to get,

$$\ln(\text{mgf}) = -\frac{df_{t'}}{2}\ln\left(1-\frac{4V+2\boldsymbol{t}'}{(\boldsymbol{t}'-1)df_{t'}}t\right)-\frac{df_1}{2}\ln\left(1+\frac{4V+2}{(\boldsymbol{t}'-1)df_1}t\right).$$

Using the Taylor expansion for ln(1+x) and multiplying the second element by $df_{t'}/df_{t'}$,

$$\begin{aligned}\ln(\text{mgf}) &\approx -\frac{df_{t'}}{2}\left(-\frac{4V+2\boldsymbol{t}'}{(\boldsymbol{t}'-1)df_t}t-\frac{(4V+2\boldsymbol{t}')^2}{2(\boldsymbol{t}'-1)^2df_{t'}^2}t^2+O(1/df_{t'}^{3+})\right)- \\ &\quad -\frac{df_{t'}}{2}\left(\frac{4V+2}{(\boldsymbol{t}'-1)df_{t'}}t-\frac{(4V+2)^2(df_{t'}/df_1)}{2(\boldsymbol{t}'-1)^2df_{t'}^2}t^2+O(1/df_{t'}^{3+})\right) \\ &= -\frac{df_{t'}}{2}\left(-\frac{2}{df_{t'}}t-\frac{(4V+2\boldsymbol{t}')^2+(4V+2)^2(df_{t'}/df_1)}{2(\boldsymbol{t}'-1)^2df_t^2}t^2+O(1/df_{t'}^{3+})\right).\end{aligned}$$

Ignoring higher order terms, the ln(mgf) has the form:

.

$$-\frac{\boldsymbol{y}df_{t'}}{2}\left(-\frac{2}{\boldsymbol{y}df_{t'}}t-\frac{1}{2}\frac{4}{\boldsymbol{y}^2df_{t'}^2}t^2\right)\approx-\frac{\boldsymbol{y}df_{t'}}{2}\ln\left(1-\frac{2}{\boldsymbol{y}df_{t'}}t\right)\text{for } df_{t'} \text{ big.}$$

which is the ln(mgf) of the following χ^2 distribution:

$$\frac{1}{\boldsymbol{y}df_{t'}}\boldsymbol{c}^2_{\boldsymbol{y}df_{t'}} \text{ where } \quad \boldsymbol{y}=\frac{4(\boldsymbol{t}'-1)^2}{(4V+2\boldsymbol{t}')^2+(4V+2)^2(df_{t'}/df_1)}.$$

As noted, the gamma distributions for the variances of ln($N_{t+t'}/N_t$) and var(ln(N_{t+1}/N_t)) are actually correlated. The effect of the correlation, as seen from numerical experiments, is to cause the distribution in Eq. A1 to approach the limiting distribution faster (i.e. when the *df*s in the gamma distributions are smaller).

The $\hat{\boldsymbol{s}}^2_{slp}$ used in the Dennis-Holmes method is somewhat different than the idealized $\hat{\boldsymbol{s}}^2_{slp}$ used in this derivation. First, the $N_{t+t'}/N_t$ ratios cannot generally be subsampled due to short

time series. This means the ratios are correlated and $df_{t'}$ is substantially less than the number of ratios minus one; additionally the lack of subsampling makes $\hat{s}^2_{slp}$ biased. The data are running sum transformed (L>1); this leads to further bias. These are trade-offs that improve estimation for short corrupted time series by reducing the number of negative variance estimates (percent errors column in Table B1). Despite the differences, understanding the limiting distribution for the idealized $\hat{s}^2_{islp}$ helps us understand why when we estimated a non-idealized $\hat{s}^2_{slp}$ (L > 1 and data not subsampled) from simulated data, we observed that $\hat{s}^2_{slp}$ / mean($\hat{s}^2_{slp}$) showed a distribution of the form $(1/df)\chi^2_{df}$ for a wide range of time series lengths, non-process to process error ratios, and filter lengths (Table B1).

Monte Carlo estimation was used to numerically estimate the χ^2 distributions for $\hat{s}^2_{slp}$ estimates used in the Dennis-Holmes method (= the slope of ln(R_{t+t}/R_t) versus t for t = 1,2,3,4). Monte Carlo estimation uses parameter estimates from samples of data generated with simulations to calculate the distribution of the parameter estimate (this is akin to parametric bootstrapping). We generated 5000 time series of length n using the model, $N_{t+1} = N_t \exp(\mu + \varepsilon_p)$, $O_t = N_t \exp(\varepsilon_{np})$ where the process error, $\varepsilon_p \sim$ Normal(0, σ_p), and the non-process error, $\varepsilon_{np} \sim$ Normal(0, σ_{np}). Let mean($\hat{s}^2_{slp}$) denote the mean of all 5000 $\hat{s}^2_{slp}$ estimates. For each simulation, we calculated the statistic $\hat{s}^2_{slp}$ / mean($\hat{s}^2_{slp}$) = η. We then found the best fitting df_{slp} parameter such that

$$\eta \sim \frac{1}{df_{slp}} \chi^2_{df_{slp}}$$

We did this by finding the df_{slp} that maximized the p-value from a Kolmogorov-Smirnov goodness of fit test. We repeated the fitting process for different time series lengths (n), filter lengths (L), ratios of process to non-process error ratios (σ_p/σ_{np}) and μ and σ_p. The best fitting df_{slp} values for

different n, L and ($\sigma_{p/}\sigma_{np}$) are given in Table B1 with the p-values for the fitted distribution. The observed bias and γ parameters from the simulations are given in Table B2. The degrees of freedom depended mainly on the length of the time series, n, and the length of the filter, L. There was an approximately linear relationship between n, L and the df_{slp} values in Table B1. The following formula gives a close approximation of the numerically calculated df_{slp}:

$$df_{slp} = 0.333 + 0.212\, n - 0.387\, L \quad \text{for} \quad n > 15.$$

Variance of $\hat{m}_R$, Eq. 3

Given n observations, $O_1, O_2, O_3 \ldots O_n$, of the true population size, $N_1, N_2, N_3 \ldots N_n$, the O_t series is transformed into a running sum, $R_1, R_2, R_3 \ldots R_r$ where $r = n\text{-}L+1$ and $R_t = \sum_{i=t}^{t+L-1} O_t$.

$$\hat{m}_R = \text{sample mean of } \ln(R_{t+1}/R_t) = \frac{1}{n-L}(\ln(R_r) - \ln(R_1))$$

$$\text{var}(\hat{m}_R) = \text{var}\left(\frac{1}{n-L}(\ln(R_r) - \ln(R_1))\right)$$

$$= \frac{1}{(n-L)^2}\text{var}(\ln(R_r) - \ln(R_1)) = \frac{1}{(n-L)^2}\text{var}\left(\ln\left(\sum_{i=r}^{n} O_i\right) - \ln\left(\sum_{i=1}^{L-1} O_i\right)\right)$$

Denote $\overline{N}_t$ as the mean of the N's that comprise the running sum, R_t: $\overline{N}_t = \frac{1}{L}\sum_{i=t}^{t+L-1} N_i$, and recall $O_t = \varepsilon_{np,t}\, N_t$.

Thus $\text{var}(\hat{m}_R) = \frac{1}{(n-L)^2}\text{var}\left(\ln\left(\sum_{i=r}^{n} \varepsilon_{np,i}\delta_{i,r}\overline{N}_r\right) - \ln\left(\sum_{i=1}^{L-1} \varepsilon_{np,i}\delta_{i,1}\overline{N}_1\right)\right)$ where $\delta_{i,t} = \frac{N_i}{\overline{N}_t}$

$$\approx \frac{1}{(n-L)^2}\,\mathrm{var}\left(\ln\left(\sum_{i=r}^{n}\boldsymbol{e}_{np,i}\right)-\ln\left(\sum_{i=1}^{L-1}\boldsymbol{e}_{np,i}\right)+\ln\left(\overline{N}_r\right)-\ln\left(\overline{N}_1\right)\right)$$

for L small so that $\boldsymbol{d}_{i,t}=\frac{N_i}{\overline{N}_t}\approx 1$

$$\text{and } \mathrm{var}(\hat{m}_R)\approx \frac{1}{(n-L)^2}\left(\frac{2\boldsymbol{s}_{np}^2}{L}+(n-L)\boldsymbol{s}_p^2\right)$$

for $\boldsymbol{s}_{np}^2<1$ so that $\mathrm{var}\left(\ln\left(\sum_{i=r}^{k}\boldsymbol{\mathring{a}}_{np,i}\right)\right)\approx\frac{\boldsymbol{s}_{np}^2}{L}$ and L small so that $\ln\left(\frac{\overline{N}_r}{\overline{N}_1}\right)\approx(n-L)\boldsymbol{s}_p^2$

Note that $\hat{m}_R$ is the mean of the ln(R_{t+1}/R_t) ratios from the time series; however, for corrupted time series, the variance of the $\hat{m}_R$ is not 1/(n-L) times the variance of the ln(R_{t+1}/R_t) ratios, as it would be the case for uncorrupted time series:

$$\frac{1}{n-L}\mathrm{var}(\ln(R_{t+1}/R_t))=\frac{1}{n-L}\left(\frac{2}{L}\boldsymbol{s}_{np}^2+\boldsymbol{s}_p^2\right)\neq\mathrm{var}(\hat{m}_R)=\frac{1}{(n-L)^2}\left(\frac{2}{L}\boldsymbol{s}_{np}^2+(n-L)\boldsymbol{s}_p^2\right)$$

Using the variance of the ln(R_{t+1}/R_t) ratios would lead to high overestimation of the variance of $\hat{m}_R$. This overestimation is greater for smaller L.

Estimate of the distribution $\hat{m}_R$ from data, Eq. 4

If $\boldsymbol{s}_p^2$ and $\boldsymbol{s}_{np}^2$ were known, it would be straight-forward to specify the distribution of $\hat{m}_R$, (i.e, normal($\hat{m}_R, \boldsymbol{s}_{m,R}^2$)) however, instead we have to use estimates of $\boldsymbol{s}_p^2$ and $\boldsymbol{s}_{np}^2$ which themselves have some distribution. Below is outlined an estimate of the distribution of $\hat{m}_R$ which uses only $\hat{\boldsymbol{s}}_{slp}^2$. Deriving a distribution based on both $\hat{\boldsymbol{s}}_{slp}^2$ and $\hat{\boldsymbol{s}}_{np}^2$ appears problematic given the nature of the distribution of $\hat{\boldsymbol{s}}_{np}^2$ (see below) and given that the $\hat{\boldsymbol{s}}_{np}^2$ estimate is not independent of $\hat{\boldsymbol{s}}_{slp}^2$.

By simple algebra, we can rewrite $\frac{\hat{m}_R - m}{\sqrt{\hat{s}^2_{slp}/(n-L)}}$ as

$$\frac{\hat{m}_R - m}{\sqrt{\hat{s}^2_{slp}/(n-L)}} = \frac{\hat{m}_R - m}{\sqrt{(n-L)s^2_{m,R}/(n-L)}} \frac{1}{\sqrt{s^2_{slp}/(n-L)s^2_{m,R}}} \frac{1}{\sqrt{\hat{s}^2_{slp}/s^2_{slp}}}$$

$$= \frac{1}{\sqrt{\hat{s}^2_{slp}/(n-L)s^2_{m,R}}} \frac{\hat{m}_R - m}{s_{m,R}} \frac{1}{\sqrt{\hat{s}^2_{slp}/s^2_{slp}}}$$

$$\text{Thus, } \frac{\hat{m}_R - m}{\sqrt{\hat{s}^2_{slp}/(n-L)}} \sim \frac{1}{\sqrt{g}} \text{normal}(0,1) \frac{1}{\sqrt{c^2_{df_{slp}}/df_{slp}}} \sim \frac{1}{\sqrt{g}} t_{df_{slp}}$$

$$\text{where } g = \frac{s^2_{slp}}{(n-L)s^2_{m,R}} = \frac{s^2_{slp}}{\left(\frac{2}{L(n-L)}s^2_{np} + s^2_p\right)} \text{ for } s^2_{np} \text{ small}$$

Point estimate of s^2_{np}

A point estimate of s^2_{np} can be calculated by noting that $\text{var}\left(\ln\left(O_{t+1}/O_t\right)\right) = 2s^2_{np} + s^2_p$,

thus $\hat{s}^2_{np} = \frac{1}{2}\left(\hat{s}^2_O - \hat{s}^2_p\right) \approx \frac{1}{2}\left(\hat{s}^2_O - \hat{s}^2_{slp}\right)$ and $\hat{s}^2_{np} \sim \frac{1}{2}s^2_O c^2_{n-1} - \frac{1}{2}s^2_{slp} c^2_{df_{slp}}$ where

$\hat{s}^2_O = \text{var}(\ln(O_{t+1}/O_t))$ from the data.