ESTIMATING ECOLOGICAL TREND: WHICH MODEL SHOULD I USE?

Brian Dennis, University of Idaho

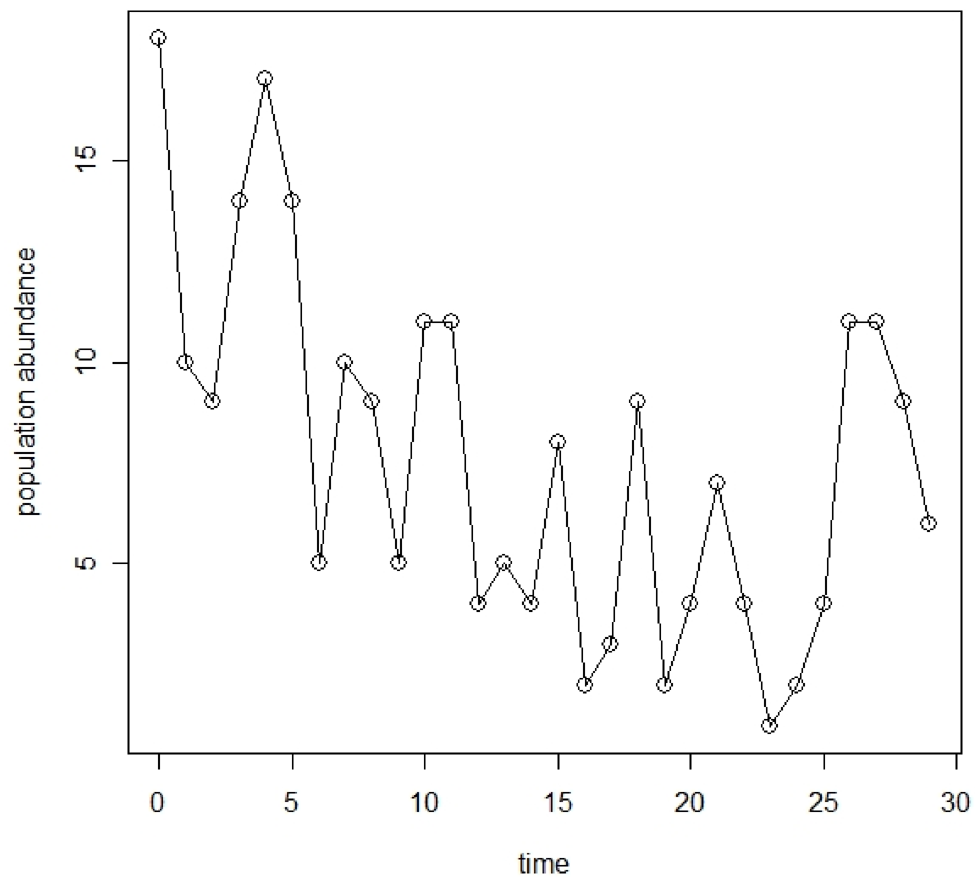
Jean-Yves Humbert, AR-T Research Station, Zurich

L. Scott Mills, University of Montana

Jon S. Horne, University of Idaho

DoD EPA DOE SERDP
Strategic Environmental Research and Development Program

THE PROBLEM: Is this population decreasing?



Exponential growth:

$$n(t) = n_0 \lambda^t ,$$

where $n(t)$ is population abundance at time t, $n_0 = n(0)$ is initial population abundance, and λ is the growth parameter.

Logarithmic scale: $x(t) = \ln n(t)$

$$x(t) = \ln n_0 + (\ln \lambda) t$$

Trend parameter: $\mu = \ln \lambda \begin{cases} > 0, \text{population increasing} \\ < 0, \text{population decreasing} \end{cases}$

Data:

$n_0 \quad n_1 \quad n_2 \quad \cdots \quad n_q$ abundances

$t_0 \quad t_1 \quad t_2 \quad \cdots \quad t_q$ times (can be unequally spaced, and usually $t_0 = 0$)

Temptation: regression of log-abundances on times, with μ estimated as slope.

Caution: a statistical *method* generally is based on a statistical *model* of how the variability in the data arises. In the regression method, one is actually building a stochastic population model!

Estimation: different *stochastic population models* lead to different estimates of trend μ.

THREE MODELS OF STOCHASTIC EXPONENTIAL GROWTH

Model 1: exponential growth, observation error (EGOE). Variability is from observation error only.

$Y(t_i)$ = observed or estimated value of $x(t_i)$:

$$x(t) = \ln n_0 + (\ln \lambda)t$$

$$Y(t_i) = x(t_i) + F_i \ ,$$

where $F_i \sim$ normal(0, τ^2) & F_0, F_1, F_3, ... independent

In the EGOE model, the population abundance follows *deterministic* exponential growth, and the variability in the data is entirely due to sampling or observation error (modeled as lognormally distributed on the original abundance scale).

The parameter μ is the mean change in $Y(t)$ over one time unit:

$$\mathsf{E}[Y(t) - Y(t-1)] = \mu \, .$$

Data:

$$\begin{array}{ccccc} y_0 & y_1 & y_2 & \cdots & y_q \\ 0 & t_1 & t_2 & \cdots & t_q \end{array}$$

Likelihood: connects data with unknown model parameters.

$$L(x_0, \mu, \tau^2) =$$

$$\prod_{i=0}^{q} (\tau^2 2\pi)^{-\frac{1}{2}} \text{exp} \left\{ - \frac{[y_i - (x_0 + \mu t_i)]^2}{2\tau^2} \right\} .$$

(product of normal density functions)

This is identical to the likelihood for a *linear regression* with y_is as response variable and t_is as predictor variable.

Maximum likelihood (ML) estimate, $\widehat{\mu}$, is the least-squares slope estimate in the regression.

($\widehat{x}_0$: intercept estimate, $\widehat{\tau}^2$: MSE)

Confidence intervals from the regression analysis are valid.

Diagnostics: usual plots & tests based on regression residuals.

Model 2: exponential growth, process noise (EGPN). Variability is due to fluctuations in the *population itself* (process noise). Observation error is negligible.

$$x(t) = x_0 + \mu t$$

or

$$dx(t) = \mu\, dt$$ deterministic DE

Add an independent random kick to each growth increment:

$$dX(t) = \mu\, dt + dB(t)\ ,$$

where $dB(t) \sim$ normal(0, $\sigma^2 dt$).

The noise $dB(t)$ represents the effect on growth rate of environmental variability.

The model is a *stochastic DE* or *diffusion process*. The SDE as written above is a recipe for simple computer simulation of trajectories. The model was proposed as a population model by Capocelli and Ricciardi (1974 TPB), and estimation of model parameters and PVA quantities was studied by Dennis et al. (1991 Ecol Monogr).

Markov property: if the value of $X(t)$ is fixed at x, then $X(t+s)$ has a normal($x+\mu s$, $\sigma^2 s$) distribution.

The parameter μ is the mean change in $X(t)$ over one time unit (given that the process starts at x_0):

$$\text{E}[X(t) - X(t-1)] = \mu$$

Data: log-abundances (no sampling error), and time intervals $s_j = t_j - t_{j-1}$.

x_0	x_1	x_2	$\cdots$	x_{q-1}	x_q	
0	t_1	t_2	$\cdots$	t_{q-1}	t_q	
s_1	s_2		$\cdots$	s_q		time intervals

Likelihood:

$$L(\mu, \sigma^2) =$$

$$\prod_{j=1}^{q} (\sigma^2 s_j 2\pi)^{-\frac{1}{2}} \text{exp}\left\{ -\frac{[x_j - (x_{j-1} + \mu s_j)]^2}{2\sigma^2 s_j} \right\}.$$

This is identical to the likelihood for a *linear regression* with values of $(x_j - x_{j-1})/\sqrt{s_j}$ as the response variable, values of $\sqrt{s_j}$ as the predictor variable, and with intercept forced through zero.

Maximum likelihood (ML) estimate, $\widehat{\mu}$, is the least-squares slope estimate in the regression.

($\widehat{\sigma}^2$: MSE)

Confidence intervals from the regression analysis are valid.

Diagnostics: usual plots & tests based on regression residuals.

Model 3: exponential growth, state space (EGSS). Both observation error and process noise are components of variability in the data. The EGSS model combines the EGOE and the EGPN models:

$$dX(t) = \mu\, dt + dB(t) \text{ ,}$$

$$Y(t_i) = X(t_i) + F_i \text{ ,}$$

where $dB(t) \sim$ normal(0, $\sigma^2 dt$), and $F_i \sim$ normal(0, τ^2), with no auto- or cross- correlations in the noises.

This is a "state space model", a type of hierarchical statistical model. The log-abundance $X(t)$ is unobserved. The random values $Y(0)$, $Y(t_1)$, $Y(t_2)$, ..., $Y(t_q)$ are the observed values and can be shown to have a *multivariate normal distribution*. They do not have the Markov property.

The model, with equal time intervals, was proposed by Holmes (2001 PNAS). ML and REML estimation was studied by Staples et al. (2004 Ecology) and Dennis et al (2006 Ecol Monogr). Unequal time intervals are introduced here.

The parameter μ is the mean change in $X(t)$, as well as $Y(t)$, over one time unit (given that the process starts at x_0):

$$\mathsf{E}[X(t) - X(t-1)]$$
$$= \mathsf{E}[Y(t) - Y(t-1)] = \mu$$

Data: estimated log-abundances

$$\begin{array}{cccccc} y_0 & y_1 & y_2 & \cdots & y_{q-1} & y_q \\ 0 & t_1 & t_2 & \cdots & t_{q-1} & t_q \end{array}$$

$s_1 \quad s_2 \quad \cdots \quad s_q$ time intervals

Likelihood is a joint pdf of a multivariate normal distribution:

$$L(x_0, \mu, \sigma^2, \tau^2) =$$

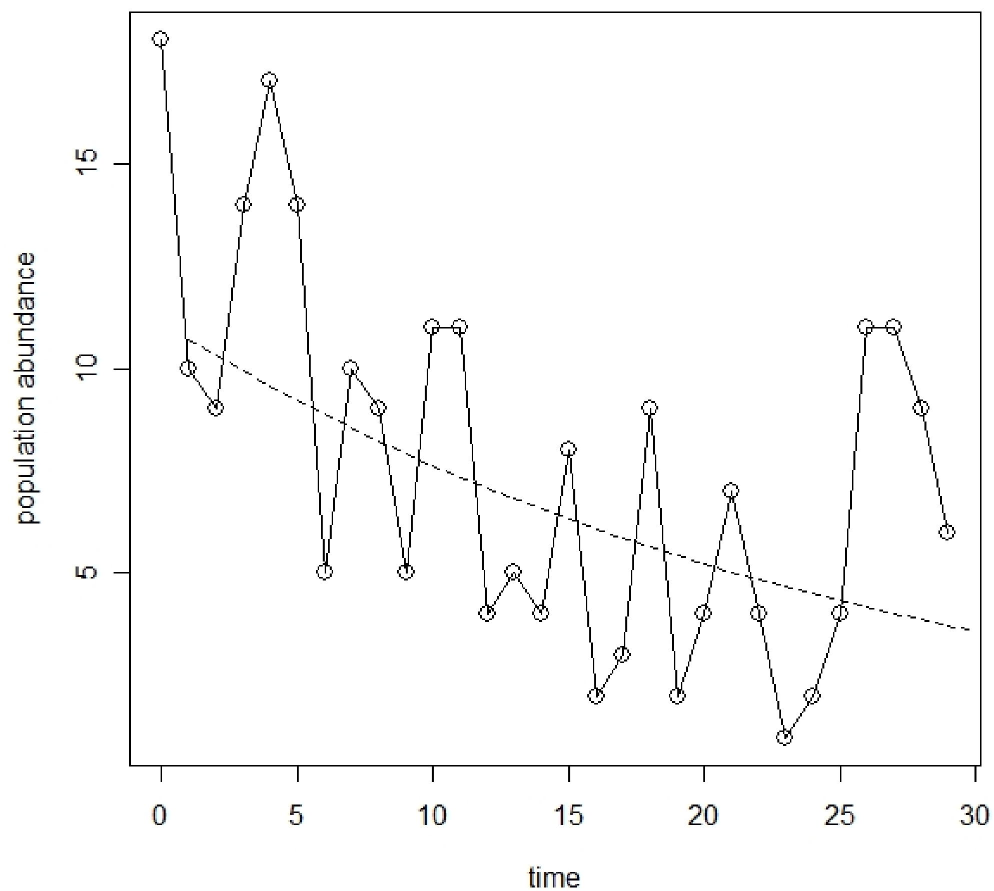
$$(2\pi)^{-\frac{q+1}{2}} |\boldsymbol{V}|^{-\frac{1}{2}} \text{exp} \left[-(\boldsymbol{y} - \boldsymbol{m})' \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{m}) \right]$$

$\boldsymbol{y}$ is the column vector of y_i values, $\boldsymbol{m}$ is the mean vector (column) with elements in the form $x_0 + \mu t_i$, and $\boldsymbol{V}$ is the variance-covariance matrix with variances in the form $\sigma^2 t_i + \tau^2$ and the (i,j)th and (j,i)th covariances both in the form $\sigma^2 t_i$ ($t_i < t_j$).

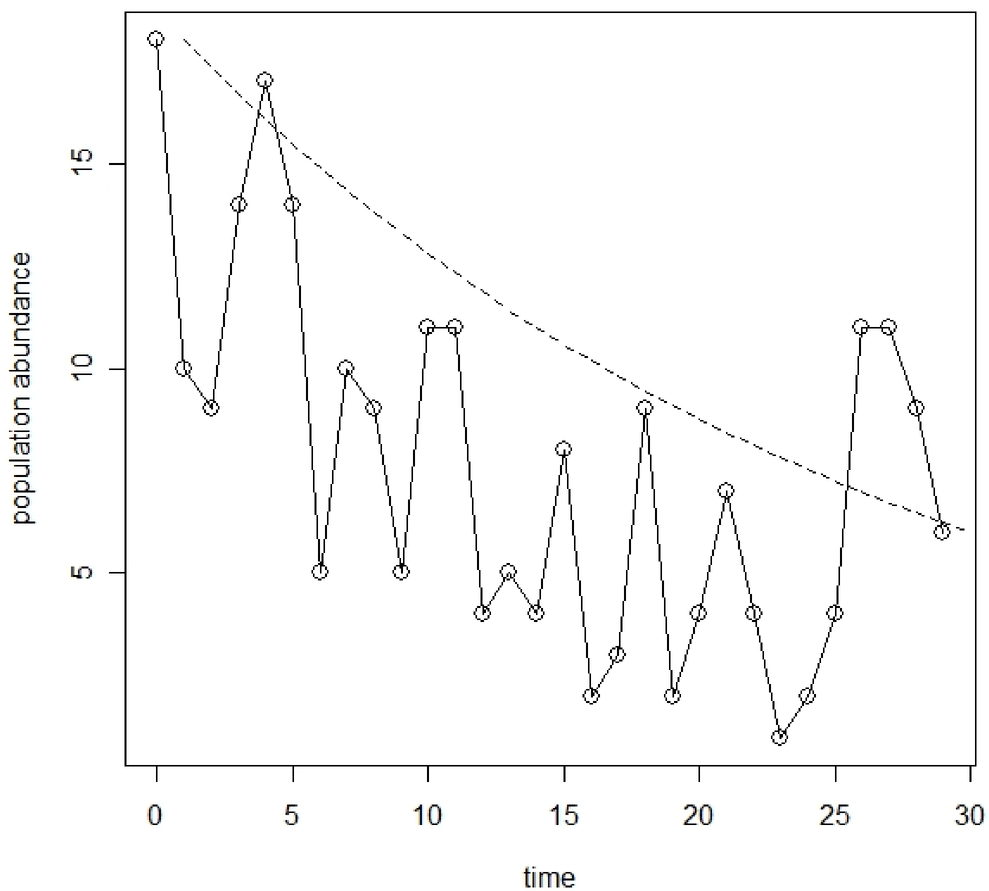
ML estimates jointly maximize L. For the EGSS model, the estimates do not reduce to regression, but must be obtained by numerical optimization.

REML estimates use the multivariate normal distribution of the second differences of the y_i values. REML estimates have improved statistical properties. Details in Staples et al. (2004 Ecology) and Dennis et al. (2006 Ecol. Monogr.). REML estimates for unequal time intervals are derived in our manuscript.

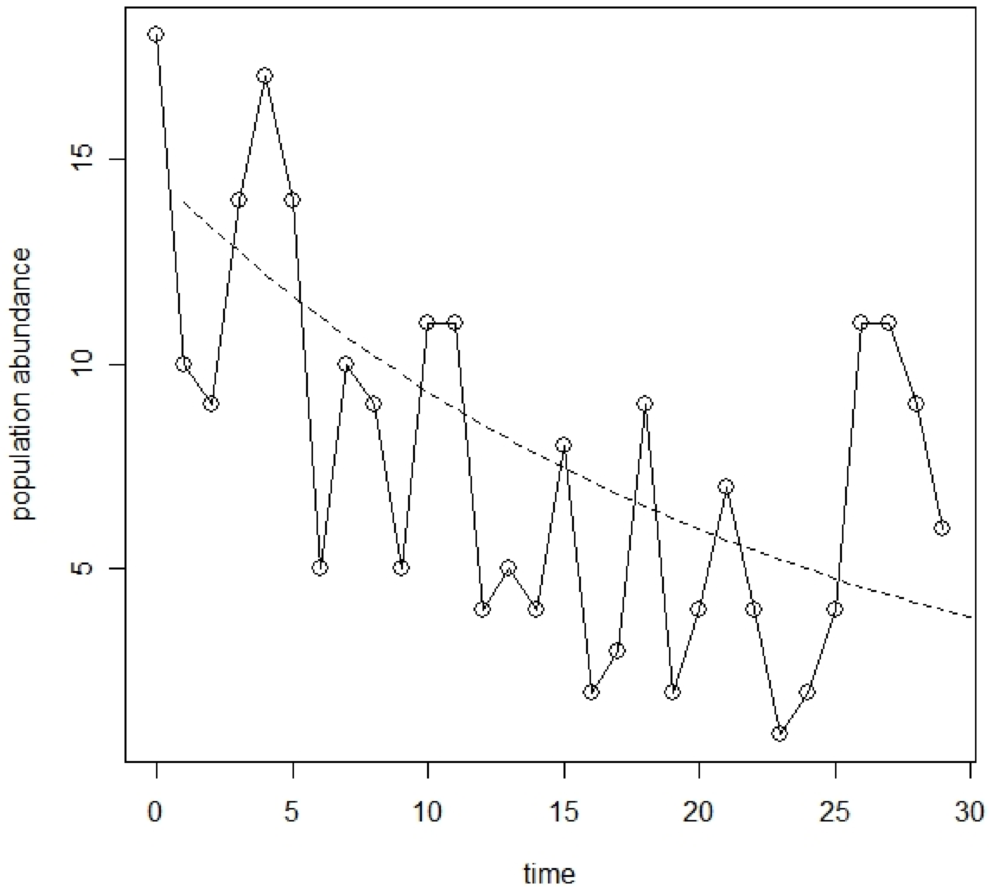
The SERDP project group will soon make available free software for these analyses.



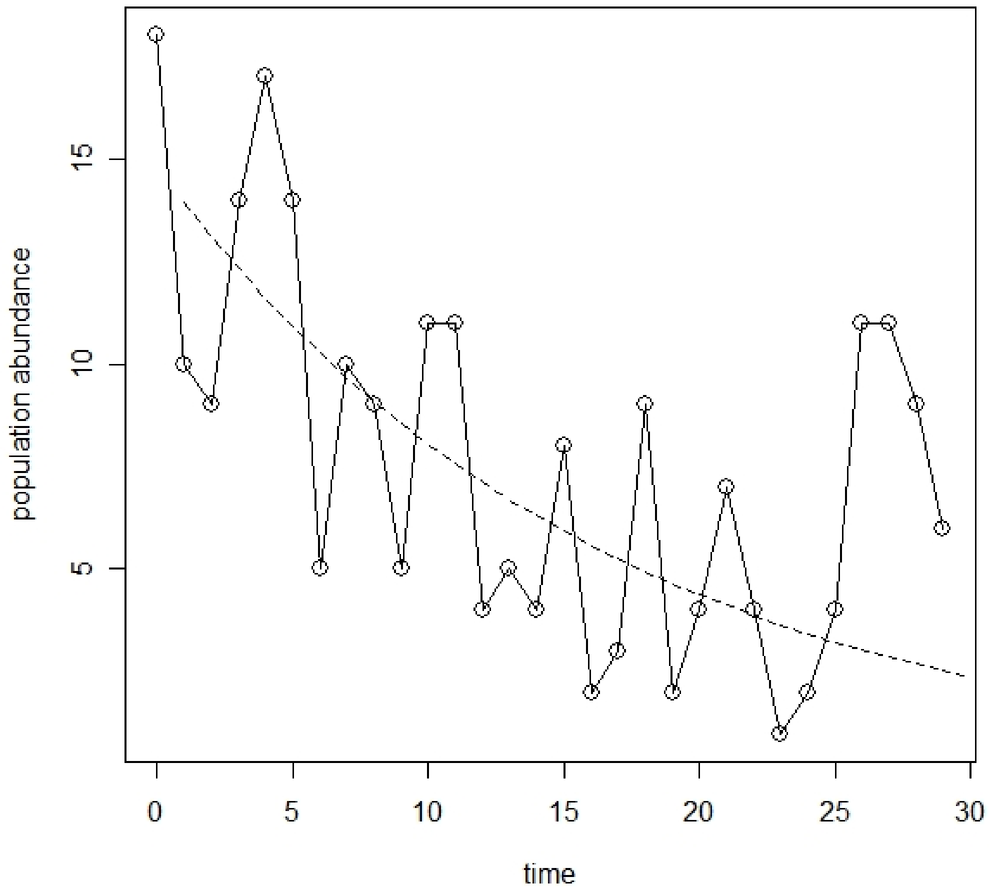
EGOE



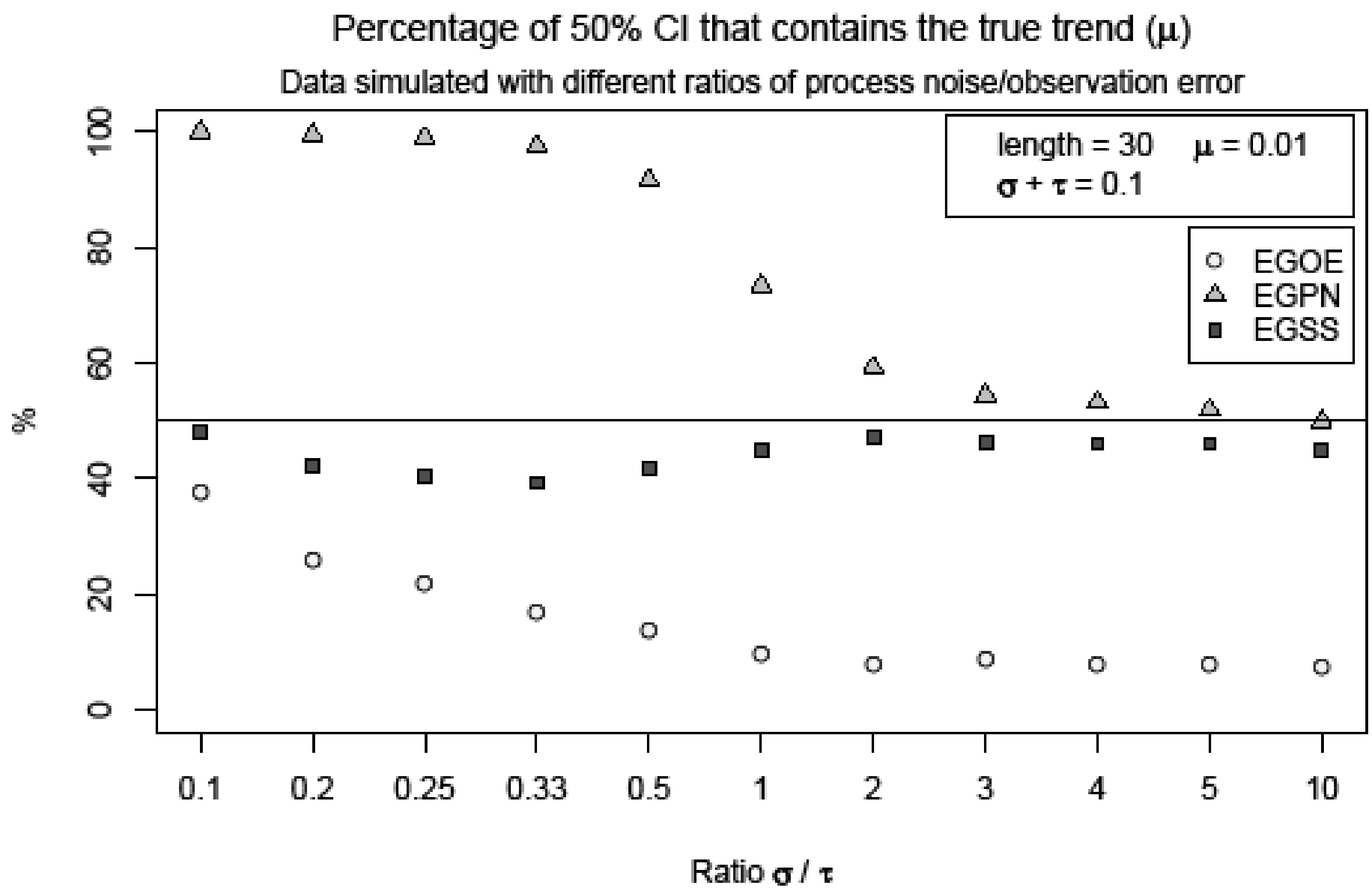
EGPN



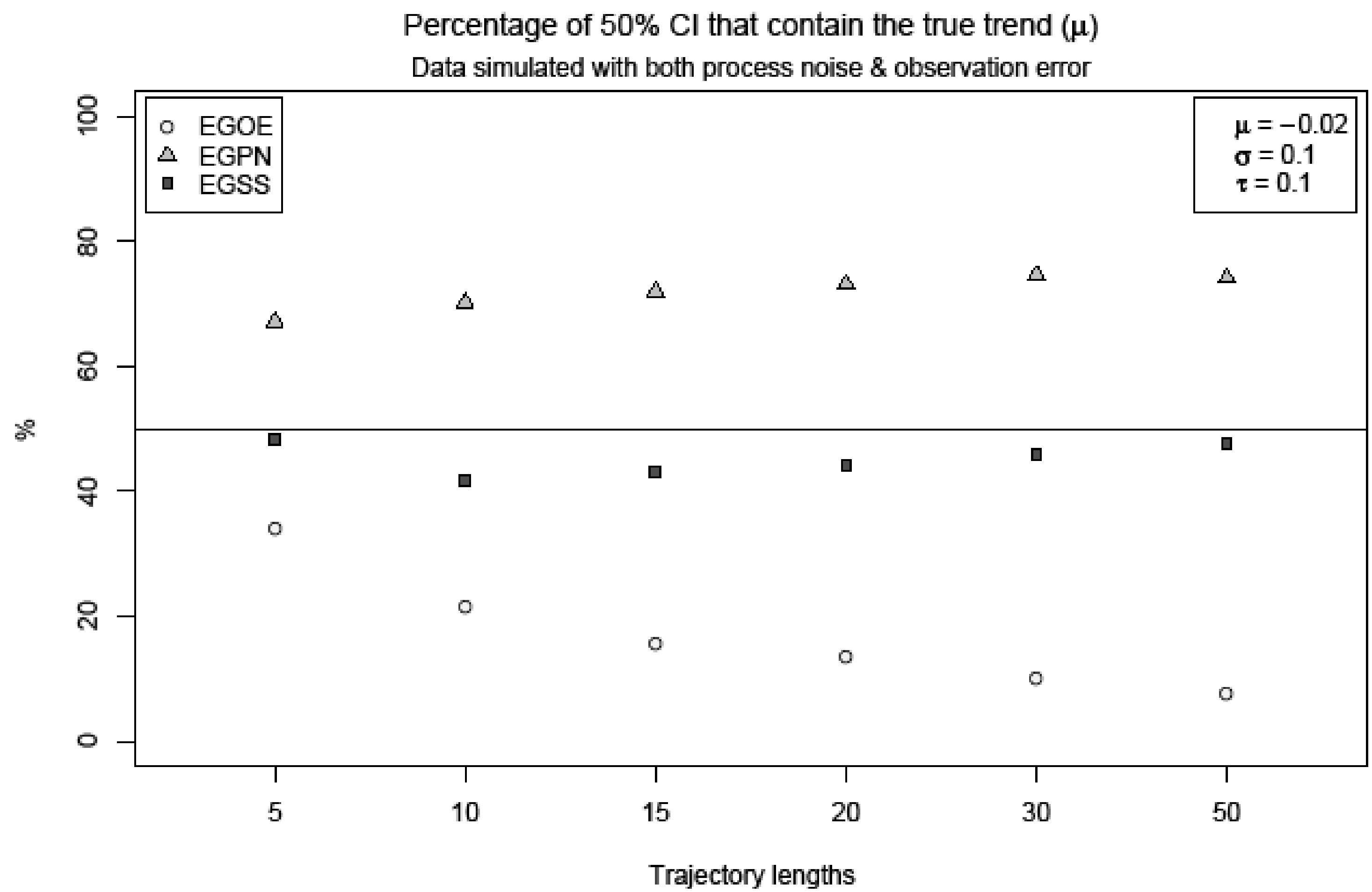
EGSS: ML estimates



EGSS: REML estimates



Percentage of 50% CI that contains the true trend (μ)
Data simulated with different ratios of process noise/observation error
length = 30 μ = 0.01
σ + τ = 0.1
EGOE
EGPN
EGSS
%
0
20
40
60
80
100
0.1
0.2
0.25
0.33
0.5
1
2
3
4
5
10
Ratio σ / τ



Percentage of 50% CI that contain the true trend (μ)
Data simulated with both process noise & observation error
EGOE
EGPN
EGSS
μ = −0.02
σ = 0.1
τ = 0.1
%
0
20
40
60
80
100
5
10
15
20
30
50
Trajectory lengths

Percentage of 50% CI that contains the true trend (μ)
Data simulated with process noise & observation error
EGOE
EGPN
EGSS
length = 30
μ = −0.02
σ = 0.1
τ = 0.1
%
0
20
40
60
80
100
0
5
10
15
20
25
Number of missing data

brian@uidaho.edu

http://www.webpages.uidaho.edu/~brian/reprints/BDennis_reprint_list.htm