Numerical Integration of Ecological Data in the Presence of Noise

Nina Embleton

School of Mathematics, University of Birmingham

Midlands MESS, 25 June 2013

Ecological Problem

- Insect pests destroy approximately 14% of crops prior to harvesting worldwide.[1]
- Further losses are incurred post-harvest.
- Integrated pest management (IPM) is the amalgamation of various tactics which work cooperatively to minimise crop damage in a sustainable way:

  1. Preventative pest management tactics are implemented
  2. Pest population is monitored
  3. Control action used when the population exceeds some threshold level.

[1] Pimentel, *Pesticides and pest control, In: Integrated Pest Management: Innovation-Development Process*, 2009.

Ecological Problem

- This threshold value can take into consideration a variety of criteria; the most well known depend on economic factors.[1]
- The fundamental principle of IPM is that a control action is only used if and when it is necessary.
- An accurate estimation of the population abundance is important in ensuring that the correct decision is made.

[1] Stern et al., *The integration of chemical and biological control of the spotted alfalfa aphid*, 1959.

Acquiring an Estimate of Pest Abundance

Figure: Trap Layout - Flatworm Monitoring [1]

[1] Petrovskaya, Petrovskii, & Murchie, *Challenges of ecological monitoring:*

Acquiring an Estimate of the Pest Abundance

- An estimate of the pest abundance is typically formulated from the sample mean pest population density: (1)

$$I \approx I_a = A\bar{f}, \quad \bar{f} = \frac{1}{N}\sum_{i=1}^{N} f_i.$$

- Alternatively, an estimate can be formulated by means of numerical integration

$$I = \int\int_D f(x,y)\, dxdy \approx I_a, \quad I_a = \sum_{i=1}^{N} w_i f_i.$$

(1) Davis, *Statistics for describing populations. In: Handbook of Sampling Methods for Arthropods in Agriculture*, 1994.

Accuracy of an Estimate

- Analysing the approximation error provides an assessment of the accuracy of an estimate.
- The relative approximation error is given by

$$E_{rel} = \frac{|I - I_a|}{|I|}.$$

- We require the approximation to be sufficiently accurate:

$$E_{rel} \leq \tau.$$

Measurement Error

- In our previous studies we assumed that the pest population densities provided by trap counts can be converted to the true densities. **This is not realistic**
- A trap catch is measurement of pest density and is subject to *measurement error*
- There are considered to be two components of measurement error: random and systematic. [1]
- The relationship between the measured and true pest population densities is given by:

$$f_i = \tilde{f}_i + \epsilon_{m_i}, \quad \epsilon_{m_i} = \epsilon_{r_i} + \epsilon_{s_i}.$$

[1] BIPM et al. *Guide to the Expression of Uncertainty in Measurements*, 1995

Measurement Error

Random Error

- Random error is the result of measurement noise.
- Noise causes the measured value to either exceed or fall below the true value.
- An overestimate is equally likely to occur as an underestimate.

Example

Measuring a length with a ruler to nearest mm.

Measurement Error

Systematic Error

- Systematic error is a result of bias.
- Bias always affects the measurement in the same way, *i.e.* always causing either an over- or underestimate.

Example

Taking a measurement with faulty or uncalibrated equipment.

Random Error in Pest Monitoring

- We focus only on the affect of noise and ignore the contribution made by systematic error:

$$f_i = \tilde{f}_i + \epsilon_{r_i}$$

- Noise impacts the repeatability of the trapping results.
- A trap may thus catch a number of insects below or above the true mean pest density at the trap location.

Impact of Random Error

- The conventional approach to reducing random error is to make the number of measurements taken sufficiently large.
- In the pest monitoring application, repeating any single measurement $\tilde{f}_i$ multiple times is not possible as the initial conditions cannot be recreated.
- Installing a large number N of traps is the only means of acquiring a large number of measurements.
- As N increases, the under and overestimates 'balance out'.

Impact of Random Error

- Practical limitations imposed by the problem mean that the number of traps that can be installed is restricted.
- Typically, the number of traps installed is $N \sim 10$
- Since we cannot make the number N of traps large, we cannot reasonably dismiss the effects of random error as negligible.
- **Key question**: How significant is the impact of random error on the accuracy of an pest abundance estimate for small N?

Integrating Measured Data

- Recall that the relative error based on exact values is

$$E_{rel} = \frac{|I - I_a|}{|I|}.$$

- Applying a method of numerical integration to measured values yields the following estimate

$$\tilde{I} = \sum_{i=1}^{N} w_i \tilde{f}_i.$$

- The relative error of the estimate based on measured values is

$$\hat{E}_{rel} = \frac{|I - \tilde{I}|}{|I|}.$$

- **Objective**: Establish how $\hat{E}_{rel}$ differs to E_{rel}.

Uncertainty in a Measurement

$$f_i = \tilde{f}_i + \epsilon_{r_i}$$

- The exact value of the pest density f_i at any location i is not known.
- The exact value of the random measurement error ϵ_{r_i} cannot be known either.
- Thus there is an *uncertainty* associated with the measured value $\tilde{f}_i$.

Uncertainty in a Measurement

- We consider each measured value of the pest density $\tilde{f}_i$ to be a realisation of a normally distributed random variable $\tilde{F}_i$.
- $\tilde{F}_i \sim N(\mu_i, \sigma_i^2)$.
- The probability density function is

$$p(\tilde{f}_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp -\frac{1}{2} \left(\frac{\tilde{f}_i - \mu_i}{\sigma_i} \right)^2$$

  where $\mu_i = f_i$.
- The uncertainty in the measured value can be quantified as

$$u(\tilde{f}_i) = \sigma_i.$$

Uncertainty in a Measurement

- Any single measurement $\tilde{f}_i$ lies in the range

$$\tilde{f}_i \in [f_i - z\sigma_i, \tilde{f}_i + z\sigma_i],$$

with probability

$$P(z) = erf\left(\frac{z}{\sqrt{2}}\right)$$

where

$$erf(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}t^2\right) dt.$$

Uncertainty in a Measurement

- We assume the pest population density obtained via a trap count is somewhere within 5-30% of the true density at the trap location

$$\tilde{f}_i \in [f_i - \nu_m f_i, f_i + \nu_m f_i], \qquad \nu_m \in [0.05, 0.3].$$

- Equating the two ranges for $\tilde{f}_i$

$$[f_i - z\sigma_i, f_i + z\sigma_i] = [f_i - \nu_m f_i, f_i + \nu_m f_i]$$

we obtain

$$\sigma_i = \frac{\nu_m f_i}{z}$$

Uncertainty in a Measurement

- Henceforth we set $z = 3$ and work with the 99.73% confidence interval

$$\tilde{f}_i \in [(1 - \nu_m) f_i, (1 + \nu_m) f_i].$$

- The uncertainty associated with the measured pest densities is then

$$u(\tilde{f}_i) = \sigma_i = \frac{\nu_m f_i}{3}.$$

- Uncertainty associated with the measured values will give rise to uncertainty in the estimate based on the measured values.

Uncertainty in the Estimate

Example

- We consider the problem in 1D and let the unit interval $D = [0, 1]$ represent the agricultural field.
- Let the pest population density function be defined by a continuous function $f(x)$.
- We artificially generate the measured values by perturbing the true values

$$\tilde{f}_i = f_i + \gamma\sigma_i,$$

  where γ is a random variable taken from the standard normal distribution and $\sigma_i = (\nu_m f_i)/3$.

$$f(x) = \frac{1}{3}\sin\left(\frac{3\pi x}{2}\right) + \frac{2}{3}$$

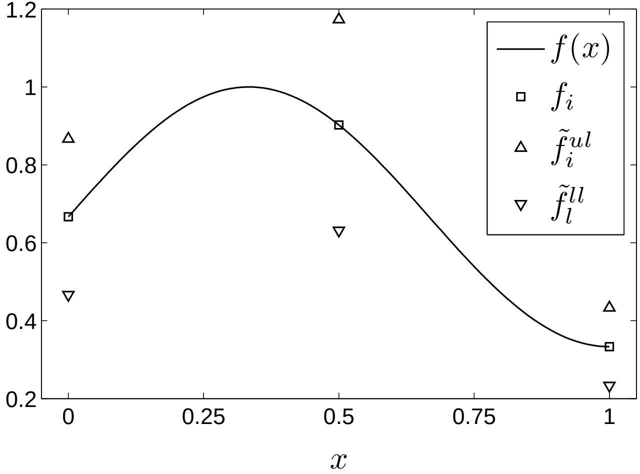


Figure: Random error is introduced to the function values. The measurement tolerance is set as $\nu_m = 0.3$.

Uncertainty in the Estimate

Example

- We generate $n_r = 100000$ sets of measured data $\{\tilde{f}_i\}$.
- Each set of data is integrated to obtain n_r estimates $\tilde{I}$ of the pest abundance.
- We choose to implement the compound trapezoidal rule

$$\tilde{I} = \sum_{i}^{N} w_i \tilde{f}_i$$

where $w_1 = w_N = h/2$ and $w_i = h, i = 2, \ldots N - 1$.

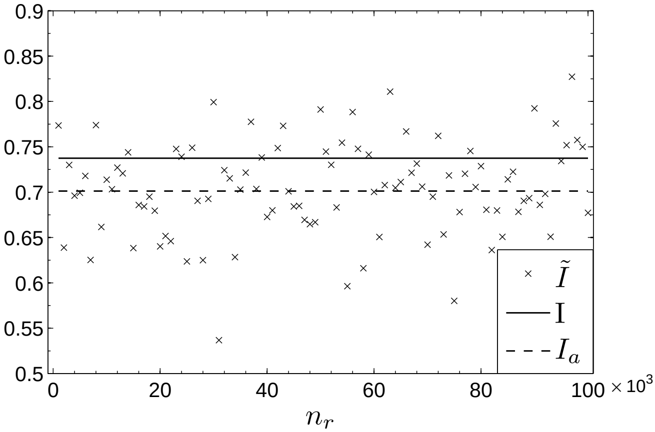


Figure:

- **Key question**: Given that we can quantify the uncertainty in the measured value, can we quantify the uncertainty in the resulting estimate of pest abundance?

Quantifying the Uncertainty

- We consider a measured value $\tilde{f}_i$ to be a realisation of a normally distributed random variable F_i with mean $\mu_i = f_i$ and standard deviation σ_i.
- The estimate $\tilde{I}$ is a linear combination of the measured values:

$$\tilde{I} = \sum_{i=1}^{N} w_i \tilde{f}_i$$

- Therefore we can consider $\tilde{I}$ to be a realisation of a normally distributed random variable $\tilde{I}_F$:

$$\tilde{I}_F = \sum_{i=1}^{N} w_i F_i.$$

- The random variable $\tilde{I}_F$ has mean $\mu_{\tilde{I}} = I_a$ and standard deviation $\sigma_{\tilde{I}}$.

Quantifying the Uncertainty

- Let us consider a function of measured values:

$$z = g(\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_N)$$

- The uncertainty associated with the measured values can be combined to yield the uncertainty associated with the estimate.
- Let us assume the function g is linear with respect to each f_i.
- First order expansion of z about the expectation of the measured values $E(\tilde{f}_i) = \mu_i$ gives

$$z - \mu_z = \sum_{i=1}^{N} \frac{\partial g}{\partial \tilde{f}_i}(\tilde{f}_i - \mu_i)$$

where $\mu_z = g(\mu_1, \mu_2, \ldots \mu_N)$.

Quantifying the Uncertainty

- Squaring both sides we obtain

$$(z-\mu_z)^2 = \sum_{i=1}^{N} \left(\frac{\partial g}{\partial \tilde{f}_i}\right)^2 \left(\tilde{f}_i - \mu_i\right)^2 + $$

$$+2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{\partial g}{\partial \tilde{f}_i}\frac{\partial g}{\partial \tilde{f}_j}(\tilde{f}_i-\mu_i)(\tilde{f}_j-\mu_j)$$

- Taking the expectation gives law of the propagation of uncertainty. [1] :

$$u^2(z) = \sigma_z^2 = \sum_{i=1}^{N}\left(\frac{\partial g}{\partial \tilde{f}_i}\right)^2 \sigma_i^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{\partial g}{\partial \tilde{f}_i}\frac{\partial g}{\partial \tilde{f}_j}\sigma_i\sigma_j\rho_{ij}.$$

[1] BIPM et al. *Guide to the Expression of Uncertainty in Measurements*, 1995

Quantifying the Uncertainty

- The estimate of the pest abundance is a function of the measured value

$$\tilde{I} = \sum_{i=1}^{N} w_i \tilde{f}_i \qquad \tilde{I}_F = \sum_{i=1}^{N} w_i F_i .$$

- We assume the measured data $\tilde{f}_i$ to be uncorrelated.
- From the law of the propagation of uncertainty for uncorrelated data we obtain

$$u(\tilde{I}) = \sqrt{\sum_{i=1}^{N} c_i^2 u^2(\tilde{f}_i)}$$

where the sensitivity coefficients c_i are

$$c_i = \left. \frac{\partial \tilde{I}_F(F)}{\partial F_i} \right|_{F=\tilde{f}} .$$

Quantifying the Uncertainty

- The sensitivity coefficients are simply the weights of the method of numerical integration employed since

$$\left.\frac{\partial \tilde{I}_F(F)}{\partial F_i}\right|_{F=\tilde{f}} = \left.\frac{\partial}{\partial F_i}\left(w_1 F_1 \cdots + w_i F_i \cdots + w_N F_N\right)\right|_{F=\tilde{f}} = w_i.$$

- Thus the uncertainty associated with the estimate is quantified by

$$u(\tilde{I}) = \sqrt{\sum_{i=1}^{N} w_i^2 u^2(\tilde{f}_i)}$$

- The expression for $u(\tilde{I})$ is equivalent to the standard deviation of the normally distributed random variable $\tilde{I}_F$

$$u(\tilde{I}) = \sigma_{\tilde{I}}.$$

- We recall that with 99.73% confidence any measured pest density belongs to the interval

$$\tilde{f}_i \in [(1-\nu_m)\tilde{f}_i, (1+\nu_m)\tilde{f}_i]$$

- We can thus say with 99.73% confidence that an estimate of pest abundance based on measured densities lies in the range

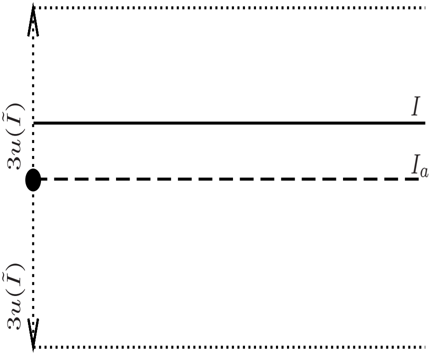
$$\tilde{I} \in [\tilde{I}^{ll}, \tilde{I}^{ul}] = [I_a - 3\sigma_{\tilde{I}}, I_a + 3\sigma_{\tilde{I}}]$$

Limits for the Error

- We now aim to establish upper and lower bounds for the relative error $\hat{E}_{rel}$.
- Let us first consider the upper limit $\hat{E}_{max}$.
- We consider three cases based on the relationship between the the approximate integral based on exact data I_a and the exact value I.

Upper Limit

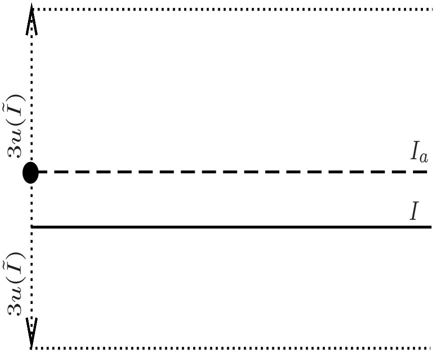
Case 1: $I_a < I$



$$\hat{E}_{max} = \frac{|I - I_a|}{|I|} + \frac{3u(\tilde{I})}{I} = E_{rel} + \frac{3u(\tilde{I})}{I}$$

Upper Limit

Case 2: $I_a > I$



$$\hat{E}_{max} = \frac{3u(\tilde{I})}{I} - \frac{I - I_a}{I} = \frac{|I - I_a|}{|I|} + \frac{3u(\tilde{I})}{I} = E_{rel} + \frac{3u(\tilde{I})}{I}$$

Upper Limit

Case 3: $I_a = I$

- When $I_a = I$ we have

$$\hat{E}_{max} = \frac{|-3u(\tilde{I})|}{|I|} = \frac{|3u(\tilde{I})|}{|I|}.$$

- By definition $u(\tilde{I}) > 0$ and $I > 0$.
- Since $I - I_a = 0$ we have

$$\hat{E}_{max} = \frac{3u(\tilde{I})}{I} = E_{rel} + \frac{3u(\tilde{I})}{I}$$

Lower Limit

- Similarly, we can obtain the lower limit by considering the following three cases:

  Case 1: $|I - I_a| \leq 3u(\tilde{I})$
  Case 2: $|I - I_a| > 3u(\tilde{I})$ and $I_a < I$
  Case 3: $|I - I_a| > 3u(\tilde{I})$ and $I_a > I$

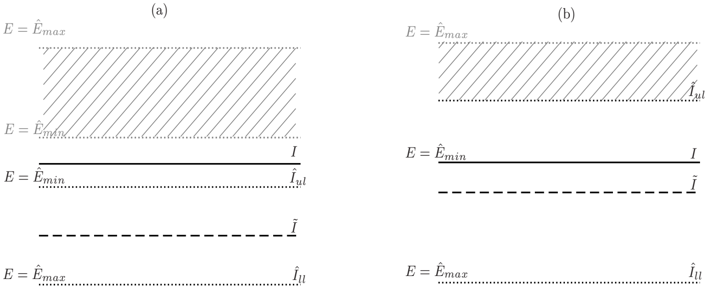
Range for the Error

- We thus arrive at a range for $\hat{E}_{rel}$.
- When $|I - I_a| \leq 3u(\tilde{I})$

$$\hat{E_{rel}} \in \left[0, E_{rel} + \frac{3u(\tilde{I})}{I}\right]$$

- When $|I - I_a| > 3u(\tilde{I})$

$$\hat{E_{rel}} \in \left[E_{rel} - \frac{3u(\tilde{I})}{I}, E_{rel} + \frac{3u(\tilde{I})}{I}\right]$$



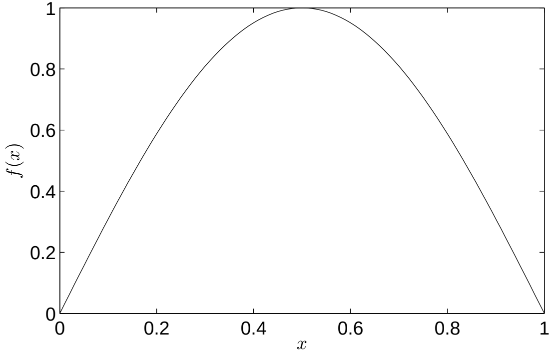
- The given range is not a 99.73% confidence interval for $\hat{E}_{rel}$.
- As the number N of traps increases, $I_a \rightarrow I$ and the probability that $\hat{E}_{rel}$ lies within the given range tends to 99.73%.

Numerical Results

- We now make some numerical computations to verify the theoretical results.
- We then consider some standard test cases to study the impact of noise on the accuracy of a pest abundance estimate.
- The agricultural field is taken as the interval $[a, b]$.
- For each test case we consider regularly spaced traps

$$x_1 = a, \quad x_i = x_{i-1} + h, \text{for } i = 2, \ldots, N-1, \quad x_N = b.$$

$f(x) = sin(\pi x)$



- Noise is introduced to the function value via the following transformation

$$\tilde{f}_i = f_i(1 + \gamma\nu_m/3),$$

- We select $\nu_m = 0.3$ and generate $n_r = 100,000$ sets of data for fixed N.

Verification of the Theory

- We compare the sample mean of the $n_r = 100,000$ realisations of $\tilde{I}$ with the theoretical mean of the random variable $\tilde{I}_F$.

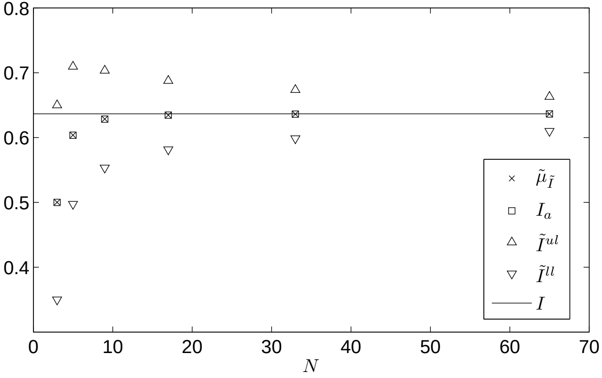
N	$\mu_{\tilde{I}} = I_a$	$\bar{\mu}_{\tilde{I}}$	$\frac{\vert\mu_{\tilde{I}} - \bar{\mu}_{\tilde{I}}\vert}{\vert\mu_{\tilde{I}}\vert}$
3	5.00000e-01	5.00211e-01	4.21928e-04
5	6.03553e-01	6.03672e-01	1.95825e-04
9	6.28417e-01	6.28567e-01	2.38202e-04
17	6.34573e-01	6.34577e-01	6.19468e-06
33	6.36108e-01	6.36039e-01	1.08364e-04
65	6.36492e-01	6.36461e-01	4.88625e-05

Verification of the Theory

- Likewise, we compare the sample standard deviation of the $n_r = 100,000$ realisations of $\tilde{I}$ with the theoretical standard deviation of the random variable $\tilde{I}_F$.

N	$\sigma_{\tilde{I}}$	$s_{\tilde{I}}$	$\frac{\lvert\sigma_{\tilde{I}} - s_{\tilde{I}}\rvert}{\lvert\sigma_{\tilde{I}}\rvert}$
3	5.00000e-02	5.00997e-02	1.99369e-03
5	3.53553e-02	3.53686e-02	3.74599e-04
9	2.50000e-02	2.49087e-02	3.65274e-03
17	1.76777e-02	1.76736e-02	2.28155e-04
33	1.25000e-02	1.25029e-02	2.35095e-04
65	8.83883e-03	8.85956e-03	2.34434e-03

Verification of the Theory



Verification of the Theory

N	$p_{\hat{I}}$	$\frac{\lvert P(3)-p_{\hat{I}}\rvert}{\lvert P(3)\rvert}$
3	9.9730e-01	2.044889e-07
5	9.9753e-01	2.30418e-04
9	9.9735e-01	4.99309e-05
17	9.9723e-01	7.03940e-05
33	9.9716e-01	1.40583e-04
65	9.9713e-01	1.70665e-04

- We are satisfied that the any realisation $\tilde{I}$ lies within the range $[\tilde{I}^{ul}, \tilde{I}^{ll}]$ with probability 99.73%.

Verification of the Theory

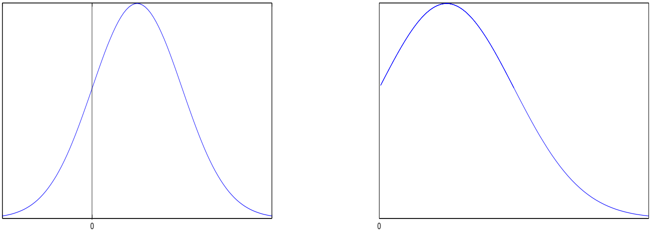
- We now consider the probability that a realisation of the relative error $\hat{E}_{rel}$ lies in the range $[\hat{E}_{min}, \hat{E}_{max}]$.

N	$p_{\hat{E}}$	$\frac{\lvert P(3)-p_{\hat{E}}\rvert}{\lvert P(3)\rvert}$
3	9.9871e-01	1.41361e-03
5	9.9874e-01	1.44369e-03
9	9.9846e-01	1.16294e-03
17	9.9794e-01	6.41528e-04
33	9.9749e-01	1.90310e-04
65	9.9727e-01	3.02857e-05

- We confirm that the probability tends to 99.73% as N gets larger.

Verification of the Theory

- The relative error $\hat{E}_{rel}$ is not a realisation of a normally distributed random variable.
- Instead it is a realisation of a random variable with a *folded normal distribution*[1] .

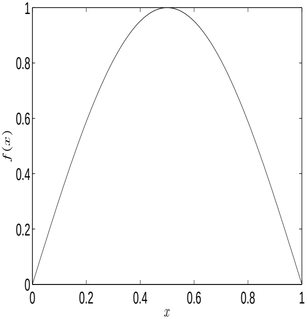


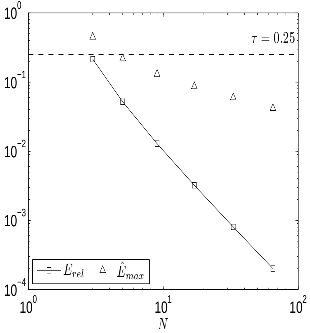
(a) (b)

Figure: (a) Normal distribution (b) Folded normal distribution

[1] *L*eone et al., The Folded Normal Distribution 1961

$f(x) = sin(\pi x)$


1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
$f(x)$
x


10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^0
10^1
10^2
N
$\tau = 0.25$
E_{rel}
$\hat{E}_{max}$

Integration of Noisy Ecological Data

N. Embleton

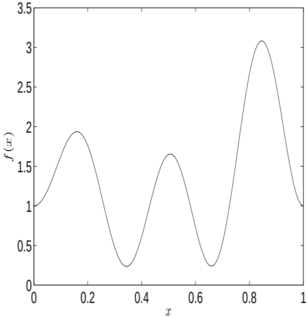
Introduction

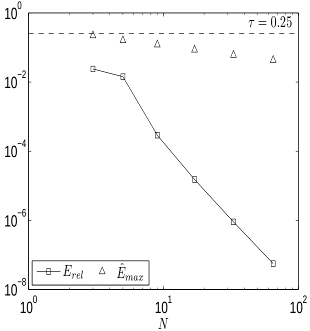
Theory

Test Cases

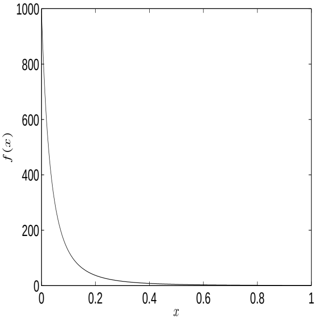
Conclusions

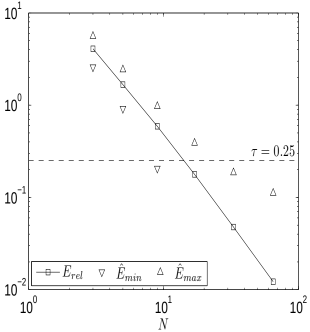
$$f(x) = \exp(x)\sin(3\pi x)^2 + \cos(\pi x)^2$$



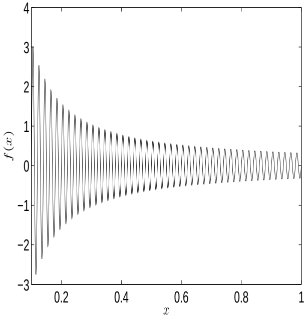


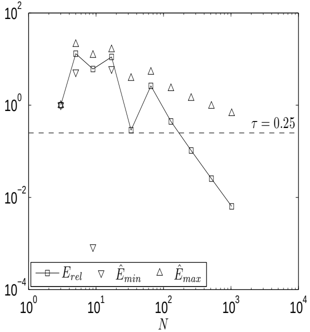
$$f(x) = (x + 0.1)^{-3}$$


1000
800
600
400
200
0
f(x)
0
0.2
0.4
0.6
0.8
1
x

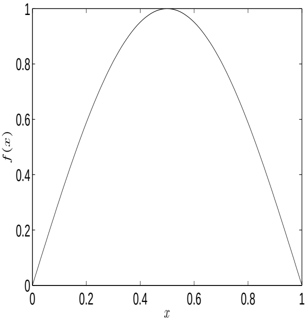

10^1
10^0
10^-1
10^-2
τ = 0.25
E_rel
Ê_min
Ê_max
10^0
10^1
10^2
N

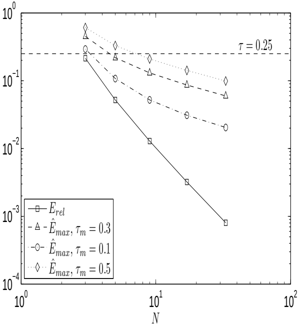
$$f(x) = \frac{sin(100\pi x)}{\pi x}$$





$f(x) = sin(\pi x)$


f(x)
x
0
0.2
0.4
0.6
0.8
1


τ = 0.25
E_rel
Ê_max, τ_m = 0.3
Ê_max, τ_m = 0.1
Ê_max, τ_m = 0.5
N
10^0
10^-1
10^-2
10^-3
10^-4
10^1
10^2

Ecological Test Cases

- We simulate ecologically relevant data using a model.
- We use the spatially explicit form of the Rosenzweig-MacArthur model (e.g. see [1]), which in its dimensionless form is:

$$\frac{\partial u(x,t)}{\partial t} = d\frac{\partial^2 u}{\partial x^2} + u(1-u) - \frac{uv}{u+h}$$

$$\frac{\partial v(x,t)}{\partial t} = d\frac{\partial^2 v}{\partial x^2} + k\frac{uv}{u+h} - mv.$$

- For any fixed time t, $u(x)$ is a spatial distribution of the pest population.
- We have d, h, k and m as the problem parameters, where d is the diffusion coefficient.

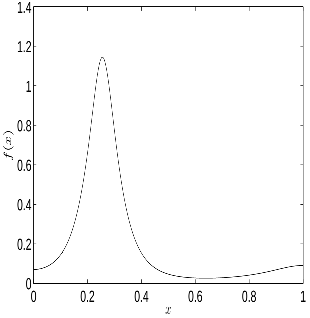
[1] Murray, *Mathematical biology*, 1989

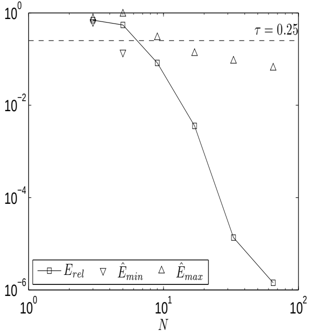
Introducing Noise to Ecological Test Cases

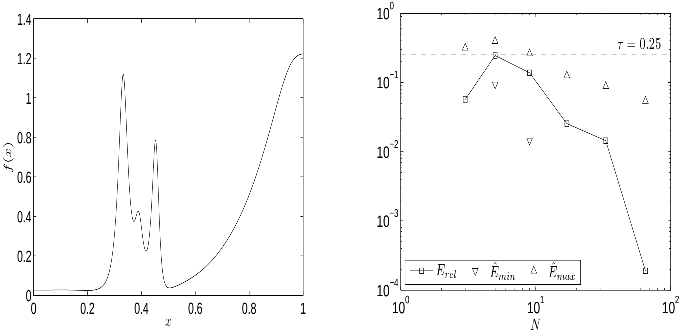
- As before, noise is introduced to the function value via the following transformation

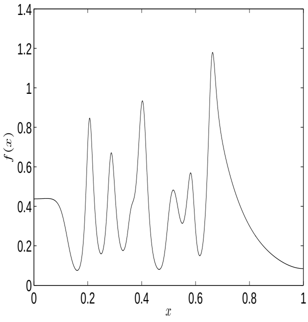
$$\tilde{f}_i = f_i(1 + \gamma\nu_m/3),$$

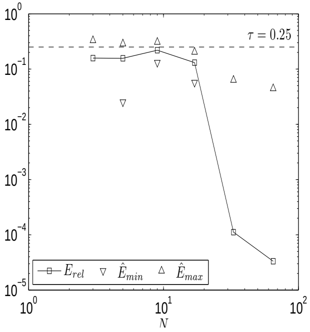
- Since we are now dealing with data with ecological meaning it is important that the perturbed value makes sense in the ecological context.
- The generated measured values must therefore be non-negative.
- A check is performed on each generated $\tilde{f}_i$.
- Any negative values are replaced with zero.


1.4
1.2
1
0.8
0.6
0.4
0.2
0
f(x)
0
0.2
0.4
0.6
0.8
1
x


10^0
10^-2
10^-4
10^-6
τ = 0.25
E_rel
Ê_min
Ê_max
10^0
10^1
10^2
N


1.4
1.2
1
0.8
0.6
0.4
0.2
0
$f(x)$
0
0.2
0.4
0.6
0.8
1
x
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
$\tau = 0.25$
E_{rel}
$\hat{E}_{min}$
$\hat{E}_{max}$
10^0
10^1
10^2
N


1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
$f(x)$
x


10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^0
10^1
10^2
N
$\tau = 0.25$
E_{rel}
$\hat{E}_{min}$
$\hat{E}_{max}$

Conclusions

- Given the measurement tolerance ν_m we can obtain bounds for the relative error of an estimate of pest abundance.
- The probability that error lies within this bound is at least 99.73%.
- On coarse grids the upper limit of $\hat{E}_{rel}$ remains relatively close to E_{rel}.
- The convergence of the upper limit is significantly slower than that of the estimate based on unperturbed data.
- As the number of traps increases the contribution to the error from the noise becomes dominant over the approximation error.
- This may be due to the fact we have considered the input data to be uncorrelated.

Further Work

- Extend the results to the 2D problem.
- Consider correlated measured pest densities.

Thank you for your attention.