DEALING WITH SPARSE AND NOISY DATA IN PROBLEMS OF SPATIAL ECOLOGY

Natalia Petrovskaya

School of Mathematics, n.b.petrovskaya bham.ac.uk

Introduction

Spatial ecology

Predictive modelling:

– better understanding of *spatio-temporal population dynamics* in various ecological systems

– development, study, and (numerical) solution of relevant mathematical models

Spatial ecology

Predictive modelling:

– better understanding of *spatio-temporal population dynamics* in various ecological systems

– development, study, and (numerical) solution of relevant mathematical models

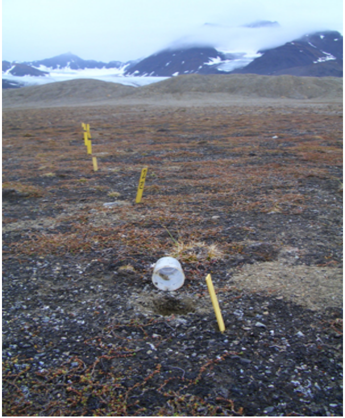
Processing *spatial* ecological data:

– making well-informed conclusions about the ecological system of interest

- (accurate) spatial pattern reconstruction
- (accurate) evaluation of ecological indices (functionals)

– validation of ecologically relevant mathematical models

Examples of spatial data collection



A line of pitfall traps, Kongsfjord.
©S.J. Coulson

Bat detectors, Fruska Gora National Park.
©The Rufford Foundation

Baseline problem: evaluation of the pest abundance

- The information about pest population size is obtained through trapping
- Once the samples (trap counts) are collected, the total number of the pest animals in the field is evaluated

The need in reliable methods to estimate the pest population size in order to avoid unjustified pesticides application and yet to prevent pest outbreaks.

Evaluation of functionals from sparse data

Evaluation of the pest abundance from discrete data

- *A functional of interest is the total pest population size (pest abundance).*

- If the density $u(x, y)$ is known at any point (x, y) of the domain D, the pest abundance I is given by

$$I = \iint_D u(x, y) dx dy.$$

- Values $u_i \equiv u(x_i, y_i)$, $i = 1, \ldots, N$ are given at the locations (x_i, y_i) (*grid points*) only and the pest population size I is reduced to computation of a weighted sum of the values u_i

$$I \approx I_a(N) = \sum_{i=1}^{N} \omega_i u_i.$$

- For *evaluation of the integral* $I \approx I_a(N) = \sum_{i=1}^{N} \omega_i u_i$ ecologists use

$$\bar{u} \approx \frac{1}{N} \sum_{i=1}^{N} u_i \ , \quad \text{so that} \ \ I \ \approx \ I_a(N) = A\bar{u} = \frac{A}{N} \sum_{i=1}^{N} u_i,$$

  where A is the area of the domain and $\omega_i = A/N$

- The *evaluation error* (integration error, approximation error) is

$$e(N) = \frac{|I - I_a(N)|}{I}$$

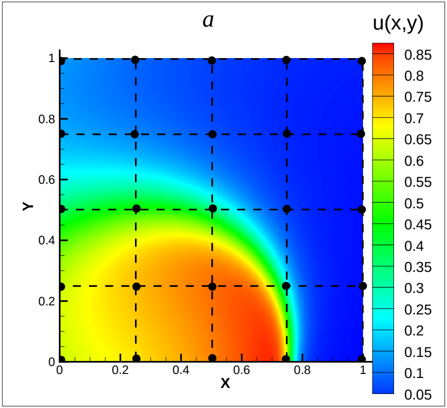
- The *accuracy* requirement is $e \leq \tau$, where τ is the specified tolerance

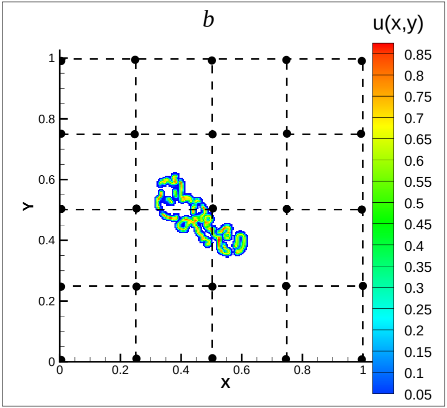
Accuracy of monitoring

How reliable is our evaluation when the data are sparse?

(a) $e(N) \sim 10^{-2}$ ($\tau \in [0.3, 0.5]$)

(b) $e(N) \sim 1.0$

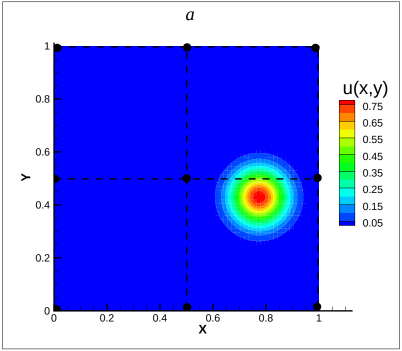


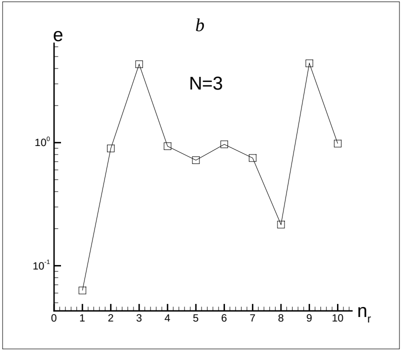


Accuracy of monitoring on a coarse sampling grid

The evaluation error depends on the location of the peak with respect to the nearest grid node on a sampling grid.

- *How big is our chance to get accurate answer e $\leq \tau$?*

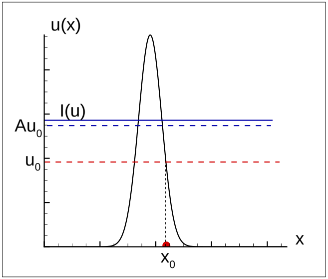




Accuracy of monitoring on a coarse sampling grid

Inaccurate measurement of the density function may occasionally give the accurate value of the functional.

$$I \;\approx\; I_a(N) = A\bar{u} = \frac{A}{N}\sum_{i=1}^{N} u_i = Au_0$$

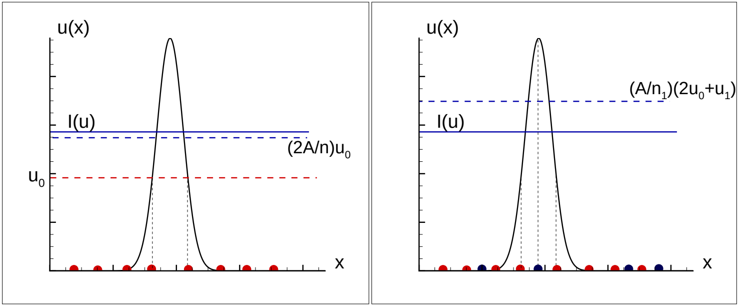


Accuracy of monitoring on a coarse sampling grid

The theory states that increasing the number of points results in better accuracy.

$$I \approx I_a(N) = \frac{A}{N} \sum_{i=1}^{N} u_i$$

- *What is the number N^* of grid points to provide the accuracy required in ecological applications?*



Key questions revisited

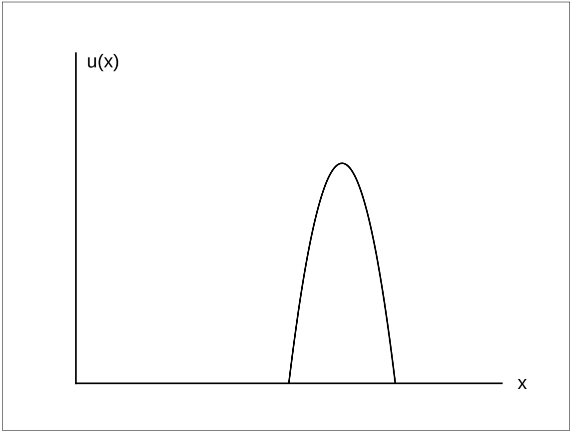
Q1: How big is our chance to get accurate answer $e \leq \tau$?
Q2: What is the number N^* of grid points on a sampling grid to provide the accuracy required in ecological applications?

The previous analysis shows that the evaluation error is a random variable on coarse sampling grids. Questions Q1 and Q2 are re-formulated as

Q1: What is the probability p of the event $e \leq \tau$?
Q2: What is the number N^* of grid points required to provide the probability $p(e \leq \tau) = 1$?

Theoretical study: quadratic function

$u(x) = -A(x - x^*)^2 + B, \quad h = \alpha\delta,$
$h = 1/(N-1)$ is the grid step size, δ is the peak width
$p(e \leq \tau) = ?$



Theoretical study: quadratic function

$$u(x) = -A(x - x^*)^2 + B, \quad h = \alpha\delta, \quad p(e \leq \tau) = ?$$

$\alpha \in [1/2, \alpha^*]$, $p(e \leq \tau) = 1$,

$\alpha \in [\alpha^*, \alpha_1]$

$$p(e \leq \tau) = 2\gamma_{II}(\alpha) + 1 - 2\gamma_{III}(\alpha) < 1, \quad p(e \leq \tau) \sim \frac{1}{h}$$

$\alpha \in [\alpha_1, \alpha_2]$

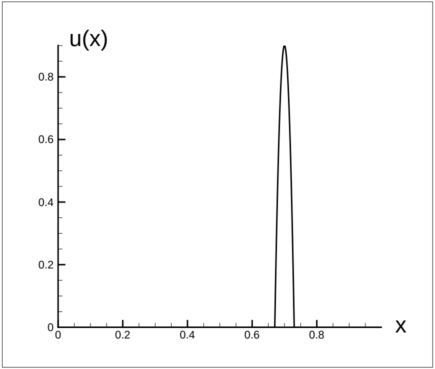
$$p(e \leq \tau) = 2(\gamma_{II}(\alpha) - \gamma_I(\alpha)) + 1 - 2\gamma_{III}(\alpha) < 1, \quad p(e \leq \tau) \sim \frac{1}{h}$$

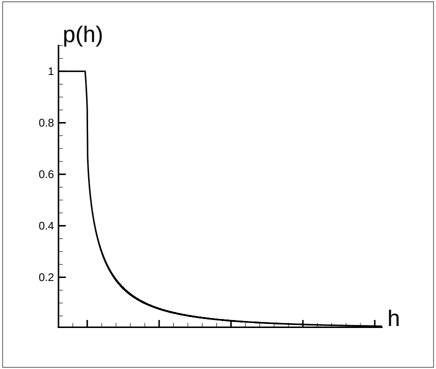
$\alpha > \alpha_2$

$$p(e \leq \tau) = 2(\gamma_{II}(\alpha) - \gamma_I(\alpha)) < 1, \quad p(e \leq \tau) \sim \frac{1}{h}$$

Probability curve for the quadratic function

$$u(x) = -A(x - x^*)^2 + B,$$
$$\tau = 0.25, \; \delta = 0.06$$

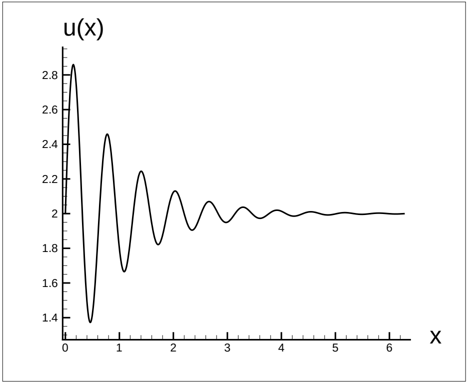


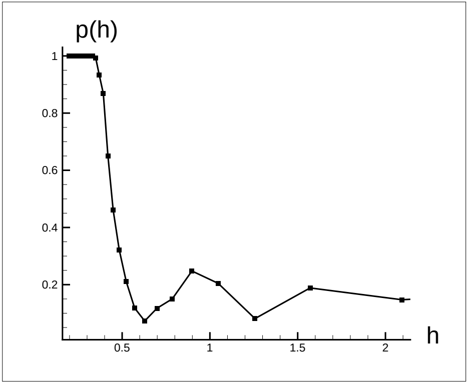


Further examples: probability for an oscillating function

$$u(x) = \exp(-x)sin(Ax) + B$$

$$A = 10.0,\ B = 2.0,\ \tau = 0.01$$





Evaluation of functionals in the presence of noise

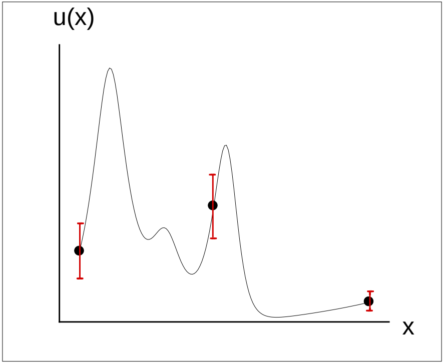
Sparse data vs. noisy data

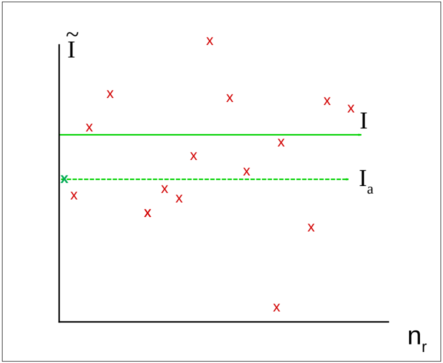
- Measurement errors are always present in the monitoring problem.
- Most of the sampling protocols currently used for the pest control imply that the evaluation error e is much smaller that the measurement error.
- The theory states the approximation error is fully controllable and therefore the measurement error is of the utmost importance.

Is this statement always true?

What is the relationship between the measurement error and the approximation error?

Example: using noisy data on a grid of 3 traps





The lower and upper error bounds

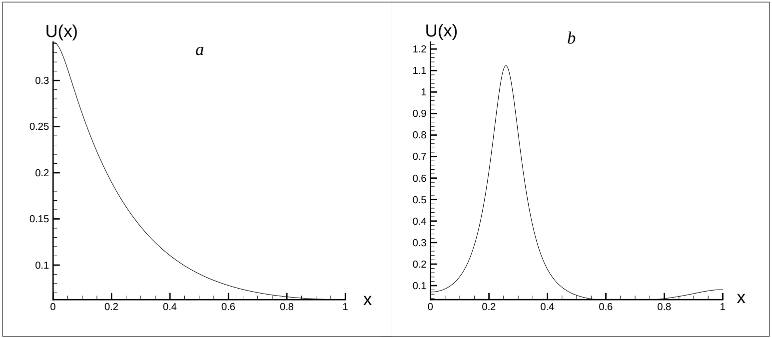
The credible interval $[\tilde{e}_{min}(N), \tilde{e}_{max}(N)]$ to which $\tilde{e}(N)$ belongs with probability $P(z)$.

$$\tilde{e}_{min}(N) = \begin{cases} 0, & \text{for } |I - I_a| \le z\sigma_{\tilde{I}}, \\ e - \dfrac{z\sigma_{\tilde{I}}}{I}, & \text{for } |I - I_a| > z\sigma_{\tilde{I}}, \end{cases}$$

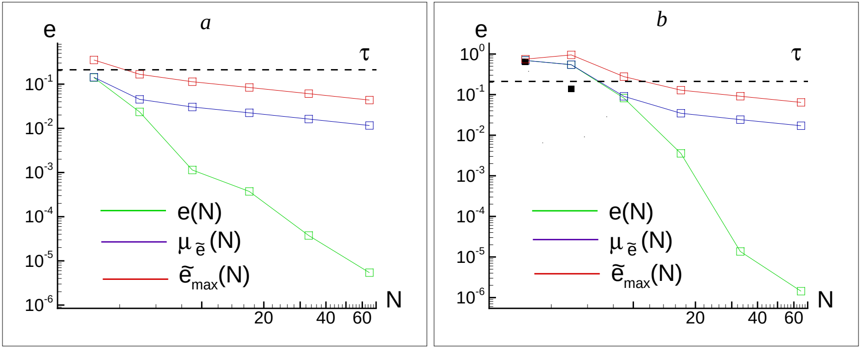
$$\tilde{e}_{max}(N) = \begin{cases} \Phi^{-1}\left[2\Phi(z) - \Phi\left(z + 2\dfrac{|\mu_{\tilde{e}}|}{\sigma_{\tilde{e}}}\right)\right], \quad \text{for } |I - I_a| \le z\sigma_{\tilde{I}}, \\ \\ \Phi^{-1}\left[\Phi(z) - \Phi\left(z - \dfrac{2|\mu_{\tilde{e}}|}{\sigma_{\tilde{e}}}\right) - \Phi\left(z + \dfrac{2|\mu_{\tilde{e}}|}{\sigma_{\tilde{e}}}\right) + 1\right], \\ \quad \text{for } |I - I_a| > z\sigma_{\tilde{I}}. \end{cases}$$

Examples of the pest population density

The R-M model: for an intermediate value of the diffusion coefficient *d*, the pattern can consist of just one or a few peaks only.

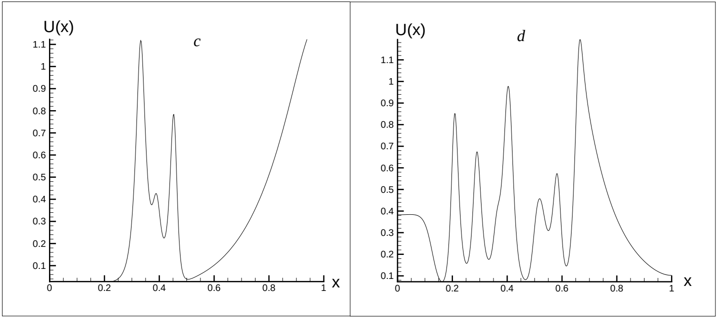


The mean error and the error bounds



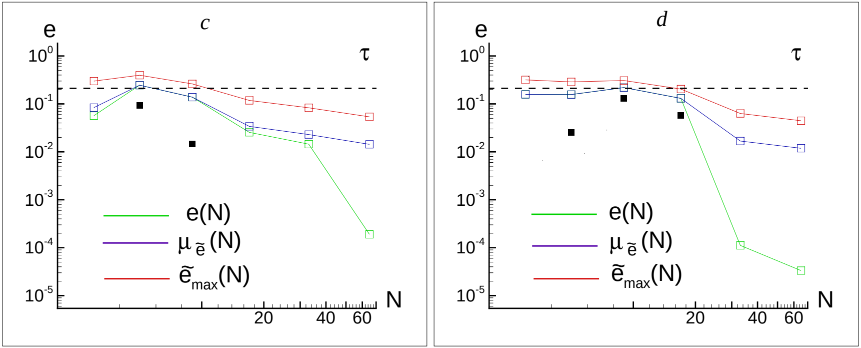
Examples of the pest population density

The number of 'humps' increases for smaller values of d resulting in oscillations.



The accuracy of evaluation depends on a spatial pattern of $u(x)$!

The mean error and the error bounds



Conclusions

- Standard methods of evaluation are not reliable on coarse sampling grids. The results of evaluation should be explained from a probabilistic viewpoint when data used for evaluation are sparse.
- The threshold number N^* of grid points for which we have $p(N < N^*) < 1$ and $p(N \geq N^*) = 1$ depends on the shape of a function (spatial pattern) monitored in the problem.
- The error in evaluation of the pest abundance from randomly perturbed data mostly depends on the evaluation error obtained from exact data.
- The knowledge of spatial pattern is crucial for accurate monitoring and any information about it must be used to its fullest extent.

THANK YOU!