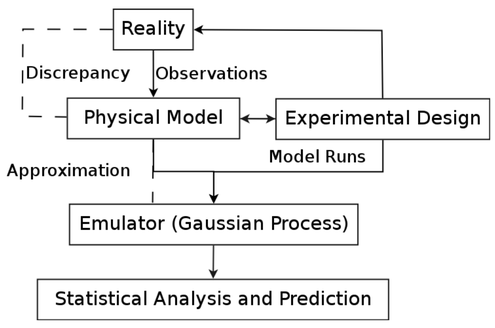
Gaussian process emulation of stochastic models: developments and application to rabies modelling

A. Boukouvalas, D. Cornford, D. Maniyar and A. Singer

Aston University, Aston Triangle, Birmingham, B4 7ET Contact: boukouva@aston.ac.uk

Introduction To Emulation

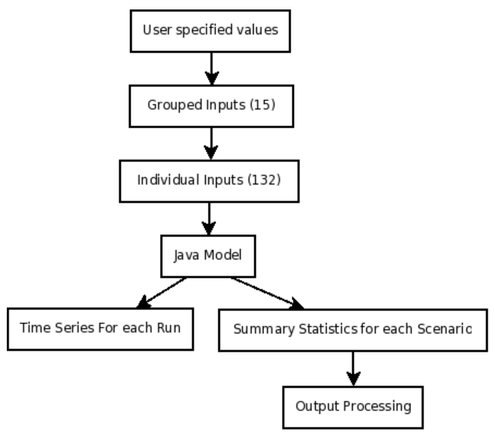
The usage of computer code models to analyse and predict the behaviour of complex systems is becoming increasingly common. Statistical surrogates, termed emulators, are commonly used for the analysis of simulator models whose computational complexity makes direct analysis infeasible. Constructing such "emulators" is a well-understood process for deterministic models. We are investigating novel methods of building such approximations for stochastic models.



Emulation methodology

Rabies Disease Modelling

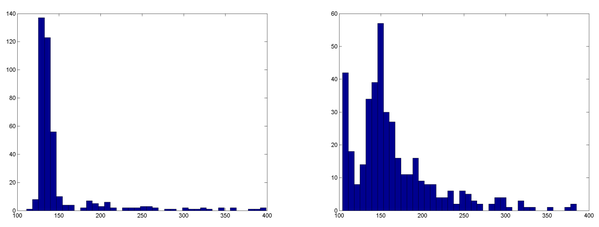
- Rabies disease propagation model with two vector species: raccoon dogs and foxes.
- Two types of output: time series and summary statistics for each run.
- Stochastic model. Output is stochastic but not normally distributed.



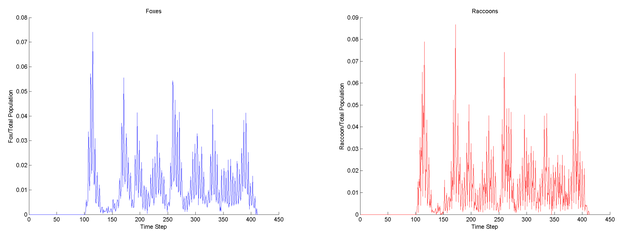
Overview of rabies model.

For policy makers, important outputs are:

- Time steps for rabies to become extinct in the raccoon population.
- Probability that rabies becomes extinct in both species after N years.

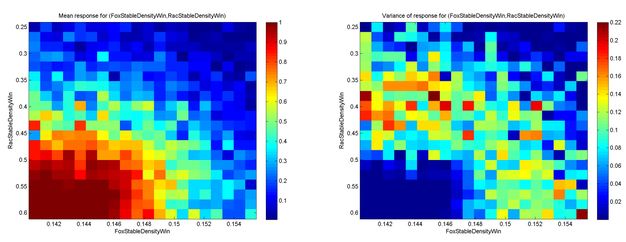


Histogram of time to extinction for two distinct inputs.



Foxes Raccoons

Time series of population sizes for a single model run.



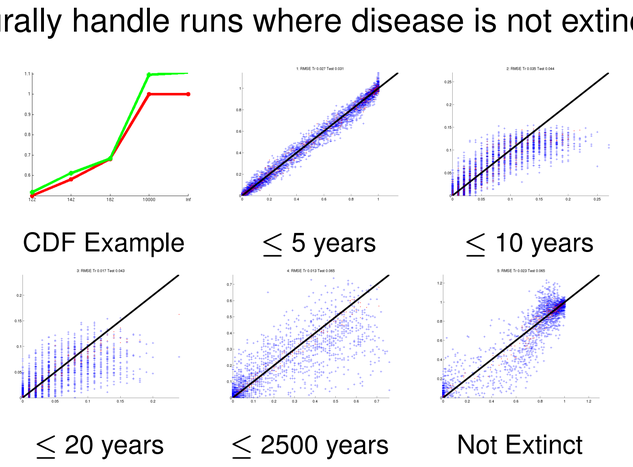
Mean Response Variance of Response

Probability of disease extiction within N=5 years for fox and racoon dog population factors.

Indicator Kriging

Describe the full output distribution.

- Estimate the cumulative distribution function (CDF) from training data.
- Discretize CDF into k slices.
- Train k Gaussian Process (GPs) emulators to estimate each slice.
- Naturally handle runs where disease is not extinct.



CDF Example $\leq$ 5 years $\leq$ 10 years

$\leq$ 20 years $\leq$ 2500 years Not Extinct

Learning the discretized distribution

Root Mean Square Error

MODEL	5 $\leq$	10 $\leq$	20 $\leq$	2500 $\leq$	INF
LINEAR	0.143	0.049	0.031	0.125	0.179
MLP(30)	0.042	0.019	0.016	0.048	0.041
GP	0.036	0.018	0.017	0.051	0.049

- Also of interest is sensitivity analysis.
- Correlation length scales for all indicator outputs agree on 4 most relevant inputs.
  - Raccoon Dog Winter Density
  - Raccoon Dog Death rate
  - Raccoon Dog Birth Rate
  - Winter Hunting proportion
- In line with expert opinion and results of expensive Monte Carlo sensitivity analysis.

Flexible Heteroscedastic Modelling

- Often first two moments are sufficient to describe model distribution.
- For Rabies model however, variance is not homoscedastic.
- In the most likely heteroscedastic framework of (Kersting et al, 2007), a coupled system of two Gaussian Processes is used to predict the mean model response and the input dependent variance.
- We modify this framework by making more efficient use of repeated observations and correcting for finite sample size errors.

Assumed Observation model:

$$t = y(x, w) + \epsilon(x)$$

- If $\epsilon(x) = \epsilon$ we have a homoscedastic noise model since it does not depend on the inputs.

The noisy predictive equations for the mean GP are:

$$\mu_* = K^*(K + R)^{-1}t$$
$$\Sigma_* = K^{**} + R^* - K^{*T}(K + R)^{-1}K^*$$

The definition of the R matrix depends on the system we define:

- *Homoscedastic Single Realisation*. $R = \sigma_n^2 I_n$.
- *Heteroscedastic Single Realisation*. $R = diag(r(x_1) \ldots r(x_n))$.
- *Heteroscedastic Multiple Realisations*. Our approach.

Heteroscedastic Emulation

The algorithm is:

1 Given a dataset D, we estimate a standard homoscedastic GP: G_1 by maximum likelihood.
  - R matrix is :

$$R = \begin{pmatrix} R_s & 0 \\ 0 & R_r \end{pmatrix} = \begin{pmatrix} \sigma_n^2 I_s & 0 \\ 0 & diag(\sigma^2_{\mu_1} \ldots \sigma^2_{\mu_r}) \end{pmatrix}$$

  where $diag(\sigma^2_{\mu_i}) = diag(\frac{\sigma_i^2}{n_i})$ the estimate of the variance of the sample mean.

Heteroscedastic Emulation Continued

2 If no replicate observations are available at x_i, sample from G_1 to estimate noise levels for the data, i.e. $v = var[t_i, G_1(x_i, D)]$. Otherwise directly compute the empirical variance for the training data. To correct for finite sample effects:

$$r = log(v) + (d + d\log(2) - \Psi(d/2))^{-1}$$
$$\sigma_v = \Psi_1(d/2)$$

r is the true log variance, d number of sample points - 1, Ψ and Ψ_1 the digamma and trigamma functions respectively.

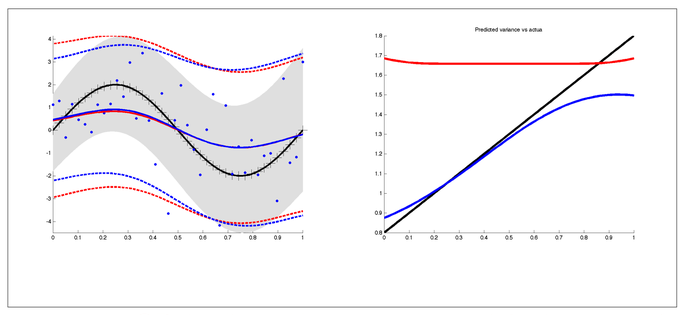
3 Using the training data D and corrected log variances estimate the z-process GP G_2.

4 Estimate the t-process G_3 using G_2 to predict the noise levels r.

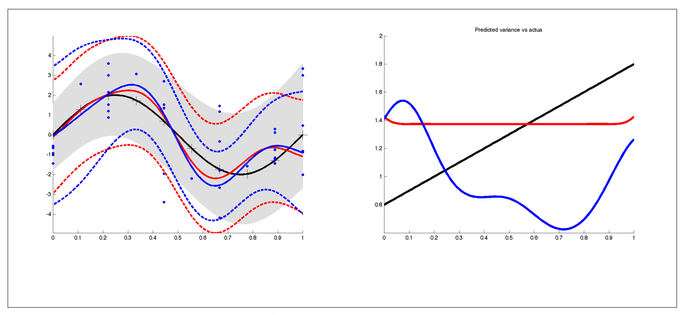
5 If not converged, set $G_1 = G_3$ and repeat from step 2.

Heteroscedastic example

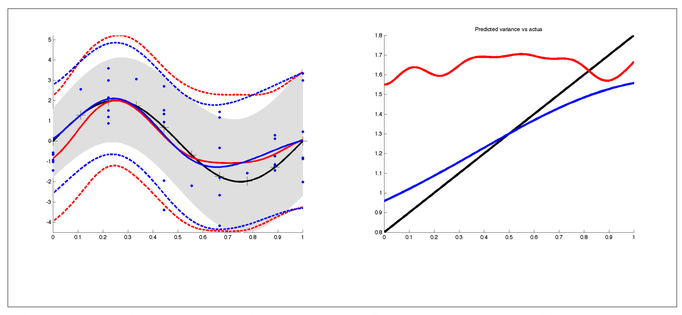
- We apply our method to the modified Goldberg synthetic dataset. The generating function is $y = 2sin(2\pi x) + N(0, x + \frac{2}{5})$.
- We examine three cases, with and without replicate observations and comparing our framework with the standard Kersting method.



Homoscedastic GP (red) vs Heteroscedastic (blue) vs true function (black). Design = 40 points uniformly distributed. Kersting method. Right plot shows predicted variance.



Design = 4 points + 6 points X 6 replications. Kersting.



Design = 4 points + 6 points X 6 replications. Our method.

Conclusions and Future Directions

- Applying heteroscedastic emulation to rabies model shows promising results.
- Indicator Kriging:
  - Build a multi-output emulator to model cross output correlations.
  - Model based framework which enforces valid probabilities in outputs (e.g. log contrasts).
  - More flexible than Heteroscedastic emulation.
- Screening.
  - Common inputs identified by indicator and heteroscedastic emulation approaches.
  - Results may be used for efficient sequential design.
- Experimental Design.
  - In both approaches, experimental design critical.
  - Direction for further research.

Acknowledgements

This research is funded as part of the Managing Uncertainty in Complex Models project by EPSRC grant D048893/1.