Analysis of complex experiments

Hugo Maruri Aguilar
H.Maruri-Aguilar@qmul.ac.uk

jede6: Pamplona
5-7 de Junio de 2023

Analysis of complex experiments

Computer simulations are widely used as substitute of experiments in situations where physical experimentation is costly or even impossible to do. My talk will describe through examples some of the challenges associated with analyzing data and creating models for simulation experiments.

My talk will discuss through three examples of models for

A infectious disease

B component of an engine and

C motorway traffic data.

. . . not just my work, but joint with L. Bastos, R. Bates, A. Boukouvalas, D. Cornford, P. Curtis, T. van-Effelterre, A. Farid, J.P. Gosling, H. Kurt-Elli and H. Wynn,

What is a computer experiment?

A computer experiment consists of simulation of a computer model which is expected to **mimic** or represent some aspect of reality. The analysis of computer simulations is a relatively recent newcomer in the bag of tools available for the statistics practitioner.

Although simulations do not neccesarily represent reality with accuracy, it is possible to **gain knowledge** about a certain phenomena through the analysis of such simulations, and the role of the statistician is to **design** efficient experiments to explore the parameter region and to **model** with a reasonable degree of accuracy the response. The model (emulator, surrogate model) is used for analysis instead of the simulator.

In each of the examples (disease, engine, motorway), I will go through description of **simulator**, **data collection** (experiment setup) and **analysis** stage.

Example A Rotavirus' and its epidemiological model

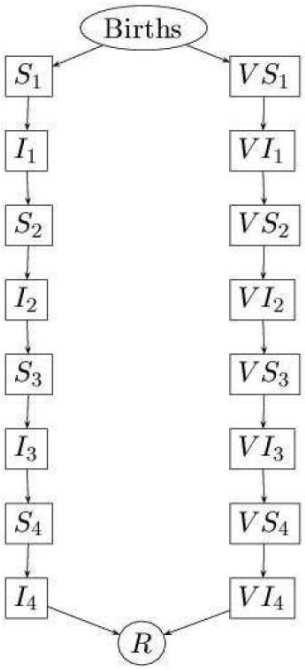
- Rotavirus is a cause of gastroenteritis; one of the major causes of diarrhoea related deaths among young children.
- Almost everyone has the virus at some point in their lives; natural immunity built up through multiple infections.Asymptomatic infections after three or four contractions of the virus.
- Rotavirus gastroenteritis is a **vaccine** preventable disease; the vaccination works by infecting the patient with a form of rotavirus.

The rotavirus model is used to back up claims about the efficacy of the vaccine on a population and to help inform required level of coverage.

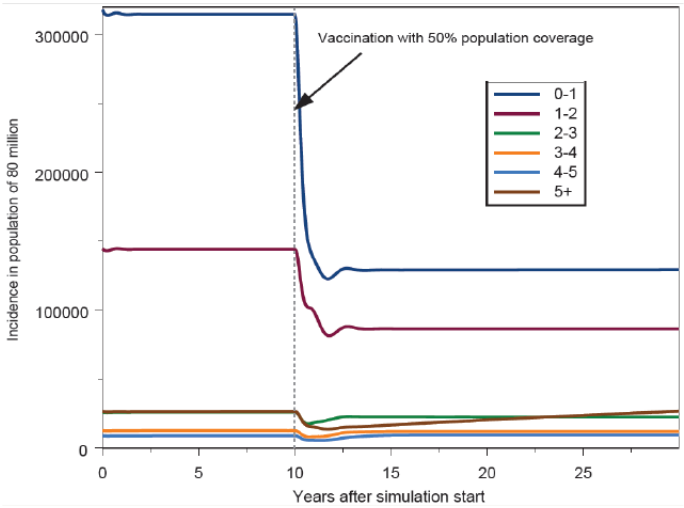
We cannot really perform an experiment of vaccinating a population: time constraints, prohibitive cost, ethical considerations and finally, not enough experimental units.

Example A Rotavirus' epidemiological model 2

This is a deterministic compartmental model with 672 compartments of 16 disease stages (8 vaccinated + 8 (non-)) across 42 age classes.



There are 20 input parameters: 6 transmission parameters, 9 reduction in risk parameters, 4 of other disease properties, and vaccination coverage.



Outputs are time series of incidence levels for each group. At a specific time point we have

$$y_{\text{age group},\text{years after vaccination}},$$

for example, $y_{3-4,10}$.

Four time points and six groups hence 24 highly correlated outputs.

Example B Fan blade assembly

Fan blade assemblies are rotating parts of turbine engines, used in airplanes.

It is known that imbalancing one of the blades triggers a response in the system, however this response under perturbations is not well understood, and the relation between imbalances and responses can be quite complex (8-blade, 24-blade assemblies).

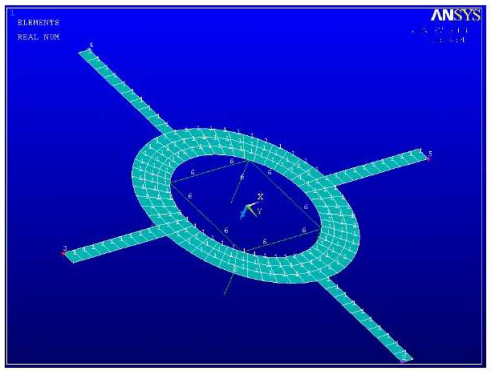
Aim: to study the system response under different imbalance conditions.

Physical experimentation with turbines and parts of them, although not impossible, can be quite costly and it is only reserved for specific cases.

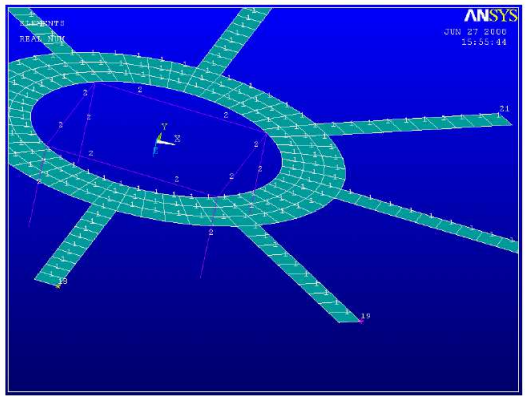
The simulator is coded in ANSYS, relatively inexpensive to run, run time around 1 minute. It is a small scale model of fan assembly, yet reasonably realistic.

Example B Fan blade assembly 2

4- and 8-fan assemblies, built upon basic mass-spring-damper model.

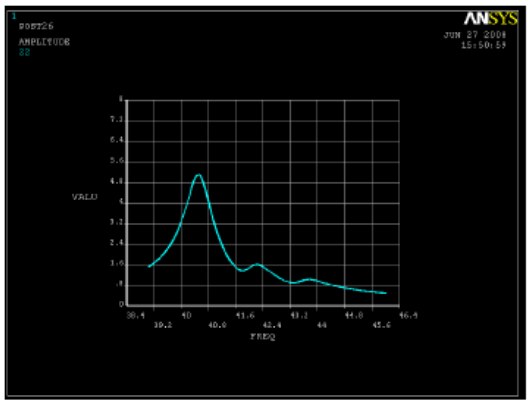


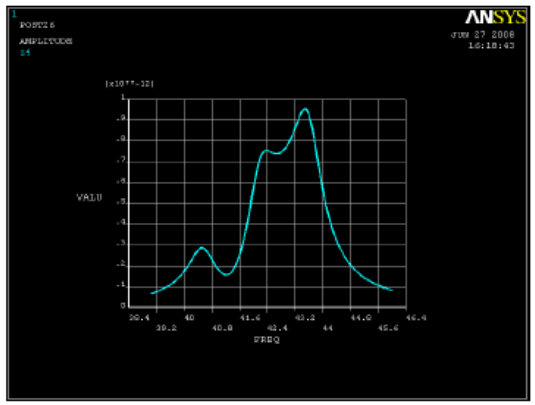
1- Thickness elements
2,3,4,5- Masses
6- Springs



8 blade assembly

In simulation, the assembly is subject to inputs of different frequencies; the response is amplitude-phase, i.e. this is Fourier frequency analysis.

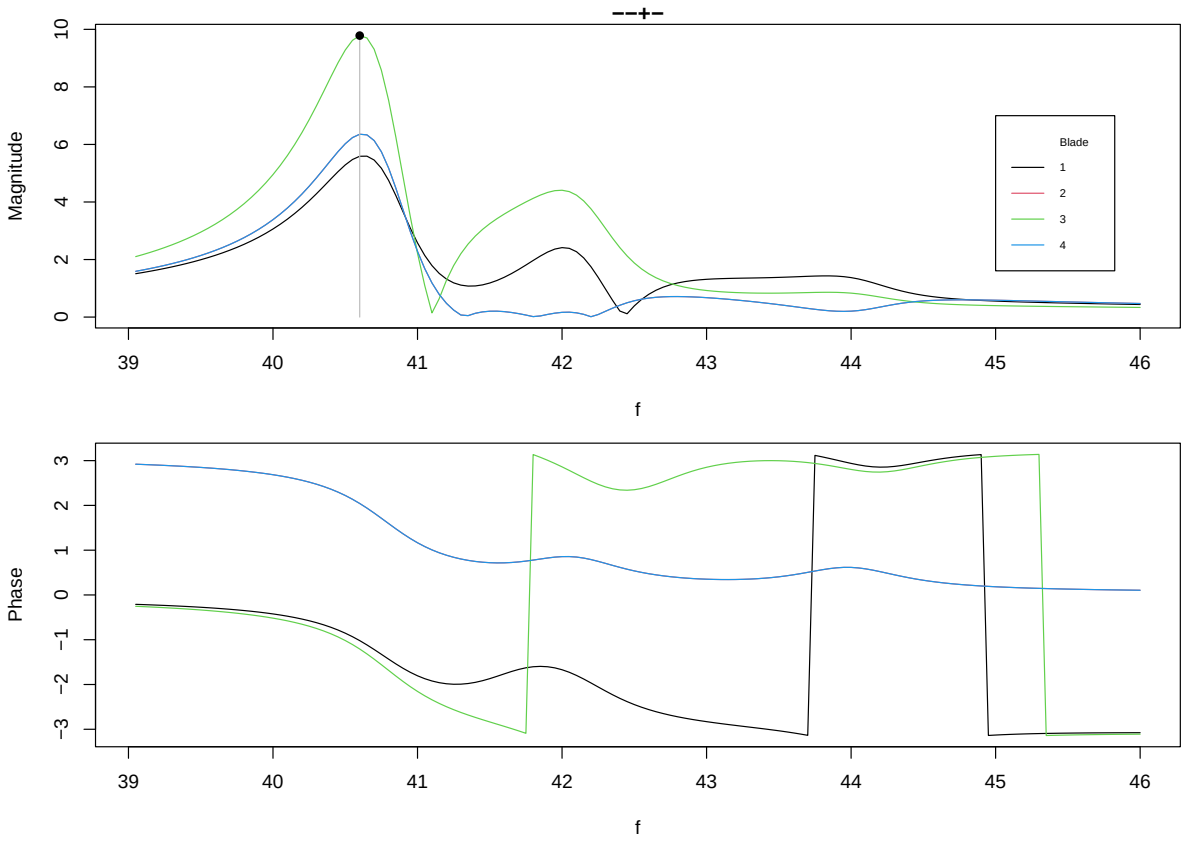




☜ Harmonic effect of high vs. low damping

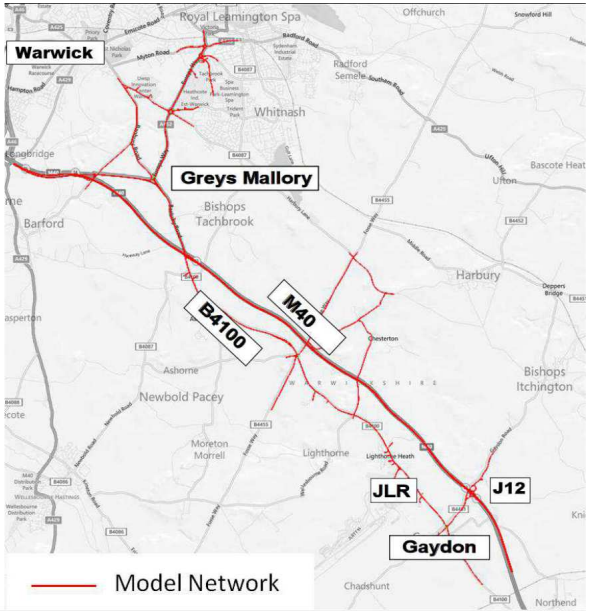
Example B Example of simulator response

Steady state amplitude and phase curves for each blade. We consider the bivariate response (a_i, f_i) (maximum) amplitude of oscilations and associated frequency, hence $8(16)$ outputs for the $4(8)$ blade assembly.



Example C Motorway traffic simulation

Microsimulation was used for a section of the M40 motorway between J12 and J14 in the midlands. The model covers 58 km of road with 44 zones acting as sources and sinks of traffic, with route choice and dynamic routing imparted to some (not all) vehicles in the simulation.



The area is overcongested with flow breakdown on the motorway and significant queuing in two key junctions.

The model has $p = 20$ inputs. Execution time is $\sim 10 - 30$ minutes; manageable but expensive for many runs.

Analysis focused on four timepoints at each of nine locations.

The overall goal is to test options for traffic management and road improvements aiming to reduce chronic congestion.

Experiments for simulations

The traditional principles of design: replication, randomization and local control ("block what you can, randomize what you cannot") do not quite apply to computer simulations.

No replication because replicated simulations give identical results; randomization of runs does not remove unwanted sources of variation; it is unclear the usefulness of blocking.

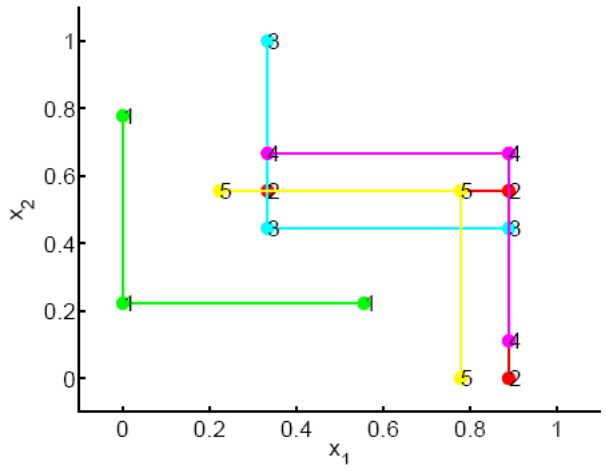
But we still require and do experimental design for simulations:

- we have an input design space that we want to explore,
- a budget for the experiment that puts constraints on the use of resources and
- a modelling objective, usually through a **surrogate** model (emulator).

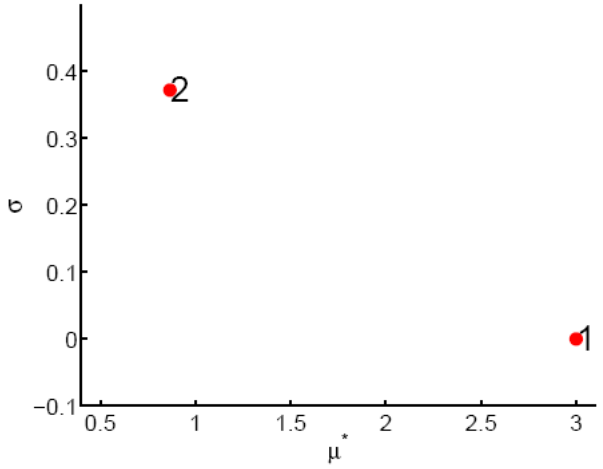
Designs for simulations include space-filling designs (low discrepancy sequences, optimized latin hypercubes), also factorial designs.

Example A epidemiological: screening design

The initial aim is to screen some of the $p = 20$ inputs form the analysis. A sequential screening design [2] following the Morris' method [7] was performed with two trajectories of $21 = p + 1$ inputs each; so in total $42 = 2 \cdot (p + 1)$ runs.



Trajectories

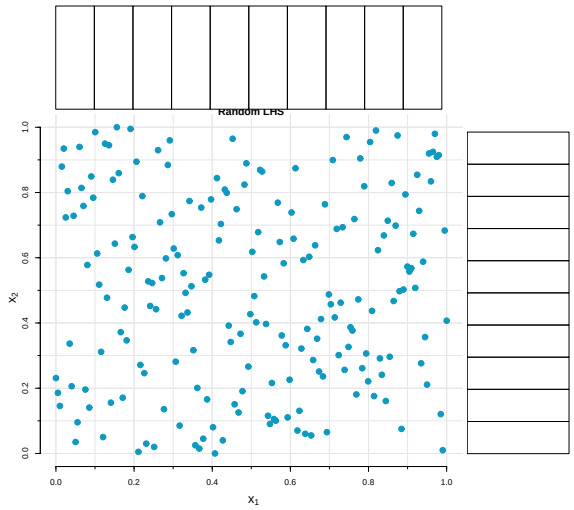


Moments

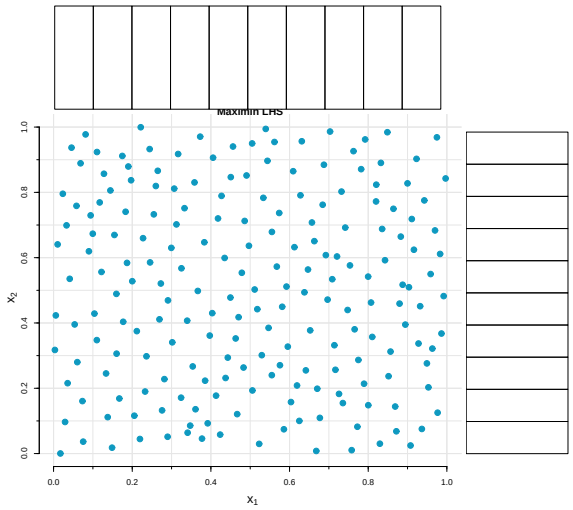
Screening results: every input was important for at least one of the outputs. Thus at this point, no input factors were removed by screening.

Example A epidemiological: data collection and design 2

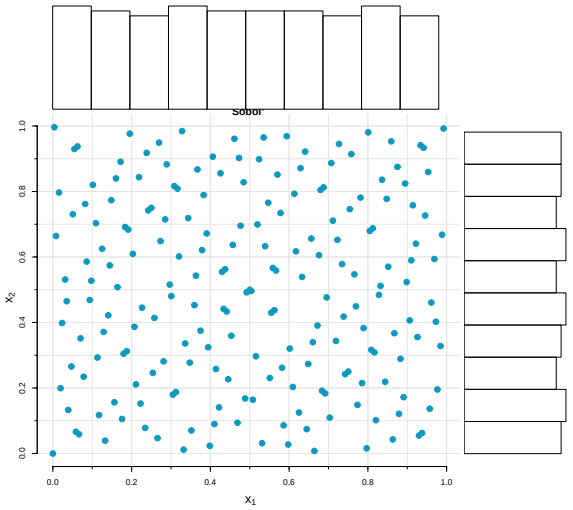
In preparation for emulation, a space filling design was built; it was a maximin latin hypercube (LHS) of $n = 200$ runs in $p = 20$ factors. A maximin LHS for validation of $n = 100$ runs was also built.



Random LHS



Maximin LHS

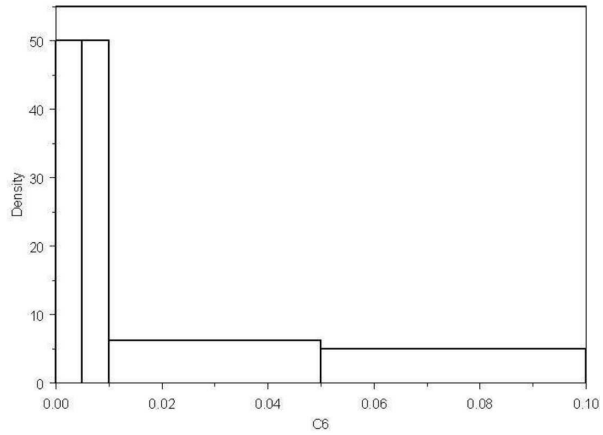


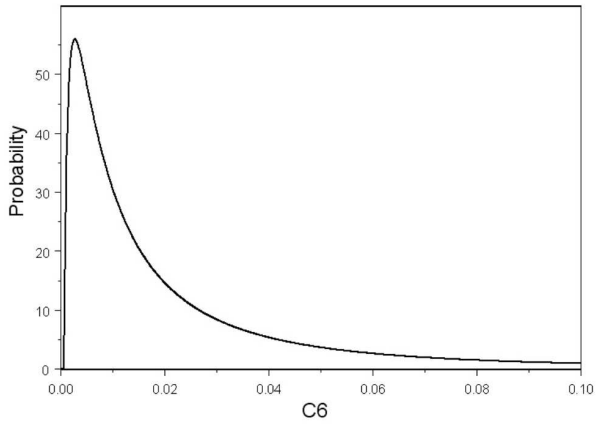
Sobol'

Example A epidemiological: data collection and design 3

Range and parameter elicitation for the twenty inputs required expert opinion and the structured questionnaire of the SHELF approach [4].

Below are elicitation results for a model input ("elicited prior"). On the left, histogram of elicited beliefs and on the right, fitted lognormal. The massive skew of the elicited distribution was typical of our exercise.





For some input parameters, the ranges were inadequate for trial runs of the simulator and had to be adjusted as part of the elicitation exercise. Eliciting beliefs from the expert is not a trivial task!

Example B 4 blade engine: Equivalence on design

The simulation experiment consisted in adding/removing small amounts of **mass** x_i (coded to ± 1) to each of the four blades $i = 1, 2, 3, 4$.

Due to the **symmetry** of the fan assembly, design points become equivalent under rotation, which lowers the cost of the simulation. For example, the run $- + -+ = (x_1 = -1, x_2 = +1, x_3 = -1, x_4 = +1)$ gives the same results as the run $+ - +-$, when observing the response for the adjacent blade.

Representative run	Equivalent 2^4 runs
$- - --$	$- - --$
$- - -+$	$- - -+, - - +-, - + --, + - --$
$- - ++$	$- - ++, - + +-, + + --, + - -+$
$- + -+$	$- + -+, + - +-$
$- + ++$	$- + ++, + - ++, + + -+, + + +-$
$+ + ++$	$+ + ++$

Example B 4 blade engine: Equivalence on design 2

The equivalence in runs allows efficiency on design as we can perform a reduced number of runs which are adapted to reproduce the response of the blade assembly.

There is interest in studying the imbalance in responses in the assembly. We know that an experiment like $--\,--$ or $++\,++$ will give no imbalance, but we'd want to find other configurations that give low imbalance.

Experiment An initial space filling design of $116 = 29 \times 4$ runs in $\mathcal{X} = [-1, 1]^4$ was used to train the surrogate model which consisted of 126 terms up to degree 5. To estimate the error, we used $16 = 2^4$ factorial points at the corners of the region $\mathcal{X}$.

The response was the **imbalance** between amplitudes observed in the blade assembly.

Example C motorway: The waves

This study consisted of a sequential experiment, each iteration is known as a **wave**. The intention is to reduce the region of experimentation to only consider parameter values that are consistent with observed data.

Starting with an **initial** design over a region of interest $\mathcal{X}$, an emulator (surrogate model) is built. Unimportant factors are dropped and so are points which do not match the observed data.

This narrows the study region, and a **second wave** design is created in a smaller region, a model fitted and implausibility criterion used to define regions which have not been ruled out yet (**NROY**) from the study. That is, we update the region $\mathcal{X}_{NROY}$.

After a **few waves** we will (hopefully) end with a good fitting model and a reduced NROY region.

Example C motorway: Implausibility for the waves

- Simulation data y is assumed to consist of emulator $f(x)$ plus a discrepancy term η. The simulation of a physical model is related to observational data z through an error term e:

$$z = \underbrace{f(x) + \eta}_{y} + e = y + e.$$

- Implausibility is a measure of the fit of the modelling scheme to real data z. For an input value x_0, the **implausibility** can be defined as

$$\mathcal{I}(x_0) = (z - E[f(x_0)])^T \left(Var\left(z - E[f(x_0)]\right)\right)^{-1} (z - E[f(x_0)]),$$

  which requires specification of a covariance structure, see [8].
  High values of implausibility indicate that the model does not agree with data. Our search from one wave to another is guided towards areas of low implausibility.

Example C motorway: Implausibility for the waves 2

Low implausibility values are used together with a threshold to define regions which have not been ruled out yet (**NROY**) from the study:

$$\mathcal{X}_{NROY} = \{x \in \mathcal{X} : |\mathcal{I}(x)| \leq a\}.$$

A rule of thumb value for the implausibility threshold is $a = 3$.

When dealing with several output variables, the implausibility is usually computed using a worst-case scenario as the maximum of implausibilities for different outputs.

Up to this point . . .

. . . the experimentation for **A epidemiological** case is directed towards reducing the dimensionality of the input space and quantifying contributions of inputs to the output of interest "incidence levels" turned into PC.

. . . for the **B engine** case, no dimensionality reduction is needed nor quantification of the inputs in the output. We want to build a reasonably useful surrogate model that will be used for finding regions of low imbalance for the blade assembly.

. . . in the **C motorway** experimentation is directed towards matching the simulator with actual data. This is of course not done with the simulator itself, but using the emulator.

Example A epidemiological: Analyses and results

By Principal Components Analysis and using the information of $n = 200$ space filling runs, the response space was reduced from $p = 24$ dimensions to only four components that explain nearly 95% of the total variability. We used in total $42 + 200 + 100 = 342$ runs of the model for our analysis, and although we tried multi-output emulation, that made little difference with respect to independent emulation of the outputs.

What we learnt from emulation and analysis [5]:

Screening: we could not reduce the dimensionality of the input space, neither from the Morris' design nor from sensitivity from emulation. We need the right parameters being investigated which may require freeing hard-coded parameters. As example, the 6×6 transmission matrix between age groups, has a restrictive multiplicative structure. Freeing up from this restriction will increase the number of parameters.

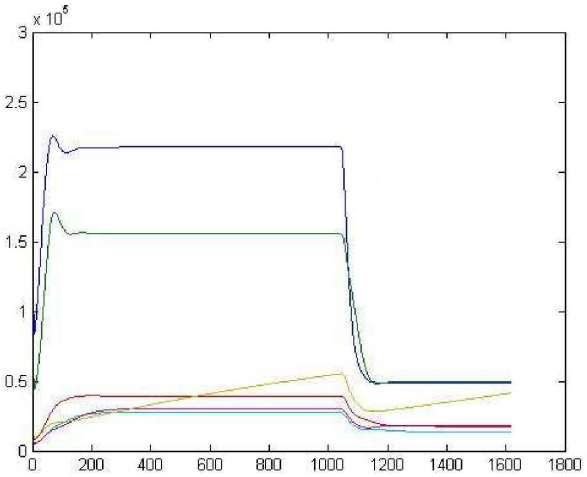
Example A epidemiological: Analyses 2

Elicitation: this is a complicated and time-consuming process that requires great attention to detail. Under increasing number of parameters, elicitation becomes quite difficult, specially under parameter dependence. In some cases, the elicitation relies on numerical shortcuts by the analyst because of the difficulty of relating to parameters.

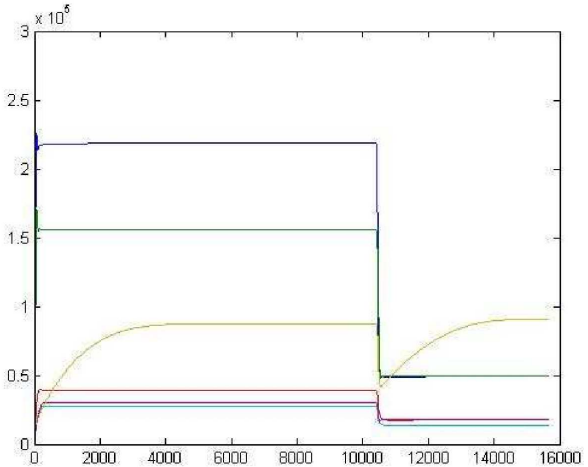
Emulation: We produced and validated emulators for the outputs. With these emulators, we quantified the contribution of inputs to the output, that is, we performed sensitivity analyses taking advantage of the elicited probabilistic beliefs about parameters. This was done at a fraction of the cost of earlier analyses done by the expert using Monte-Carlo directly with the outputs.

Example A epidemiological: Analyses 3

Model: We pointed out cases in which the model did not perform as expected: model not converging; prohibited ranges of the input parameters; vaccination was being administered before the model reaches steady state (see figure). Some of these issues were not known to the analyst and may lead to revised versions of the model.



Short run



Longer run

Example B 4 blade engine: Model and results

We fitted a regression model of the type $y = X\theta + \epsilon$, and the vector of regression coefficients minimizes a penalised criterion

$$(y - X\theta)^T(y - X\theta) + \lambda\theta^T K\theta, \text{ with } K = \int_{\mathcal{X}} \sum_{1 \le i,j \le p} f^{(i,j)^T} f^{(i,j)} dx,$$

where $f^{(i,j)} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ [1]. The scalar λ balances the relative importance of fitting to data versus smoothing. This is a form of regularization, similar to ridge regression (Tikhonov regularization) and

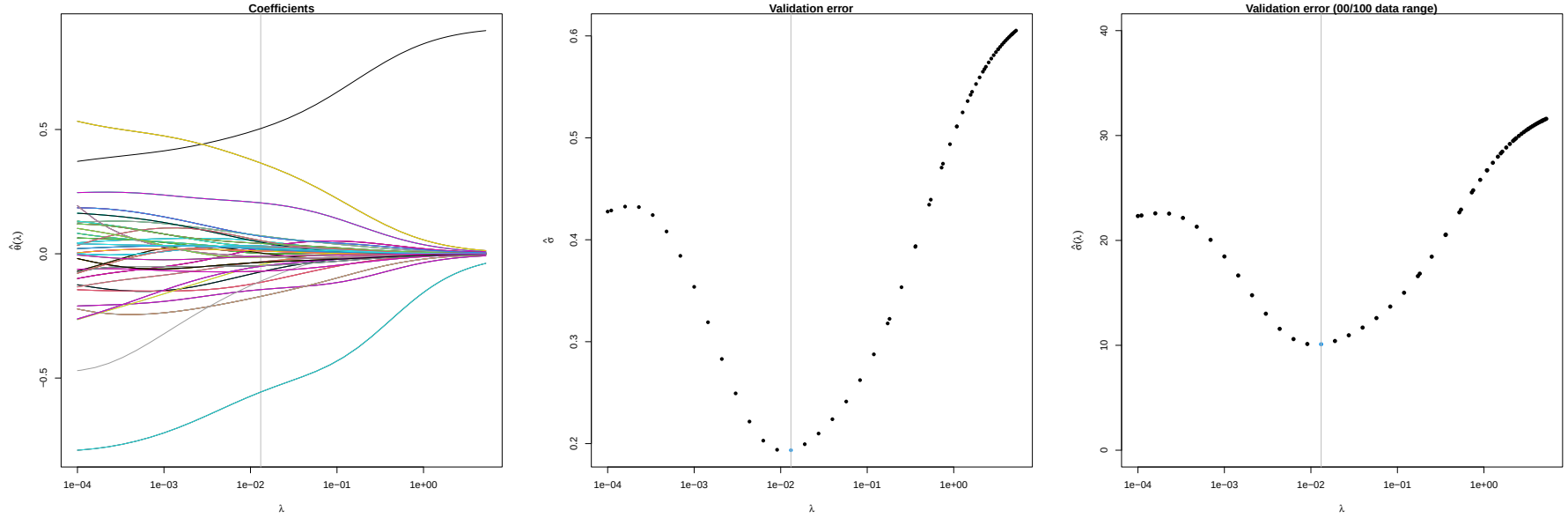
$$\hat{\theta} = (X^T X + \lambda K)^{-1} X^T y.$$

The predictive model $\hat{y}(x) = f(x)^T \hat{\theta}$ no longer interpolates the data, and an approximate prediction interval can be computed as

$$\hat{y}(x) \pm z_{\alpha/2} \sqrt{V(\hat{y}(x))}.$$

Example B 4 blade engine: Model and results 2

For a range of λ values, the smooth path was computed. Using the validation design, a model with $\lambda^* \approx 0.01$ was selected.



We obtained a simple, interpretable, fast smooth polynomial surrogate model. We explored the level set of this surrogate model to establish regions of low imbalance, not linked to the inputs being all equal.

Example C motorway: The waves [3]

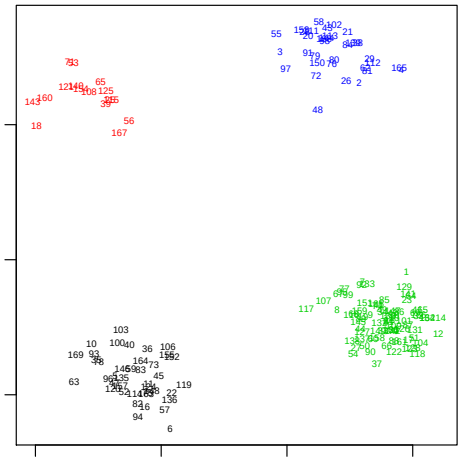
Wave 1 consisted of a maximin LH of 484 points with 5 replicates at each point to build a dual response emulator, with 250 extra points for validation.

The emulators of Wave 1 were evaluated at a maximin LH of 1100 runs in the original design space. The implausibility criterion classified 233 of these as non implausible. Thus **Wave 2** consisted of the latter points replicated 20 times and split in training/validation as $233 = 213 + 20$.

Wave 3 was created from another maximin LH of 1100 points over the whole input space and using emulators of the first two waves. This resulted in 32 more simulator runs added to the area of low implausibility.

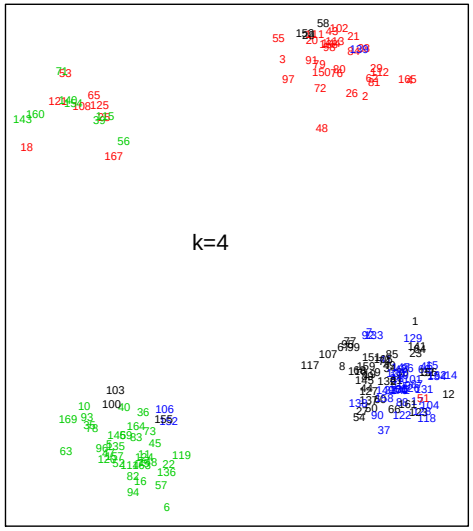
Waves 1 to 3 had 484×5, 233×20 and 32×20 runs, respectively. Without emulation, exploring the parameter space and reaching areas of alignment between data and simulator would be prohibitively expensive.

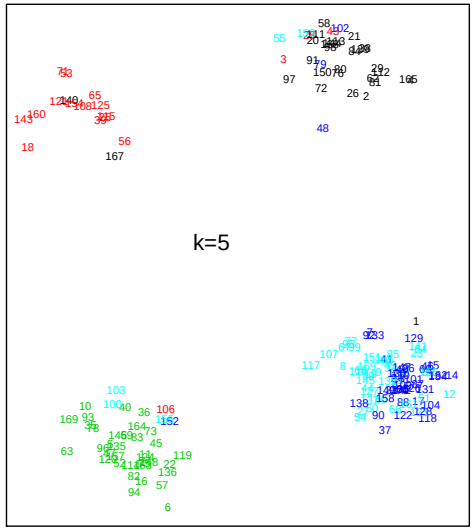
Example C motorway: Describing the low implausibility region



Two binary variables in the data strongly influence spatial clustering of points with low implausibility. The plot shows the first two components of multi-dimensional scaling of the data, colored according to four cases of the binary variables.

We clustered with k-medoids adapted to mixed data [6], suggesting 4 or 5 clusters [3].

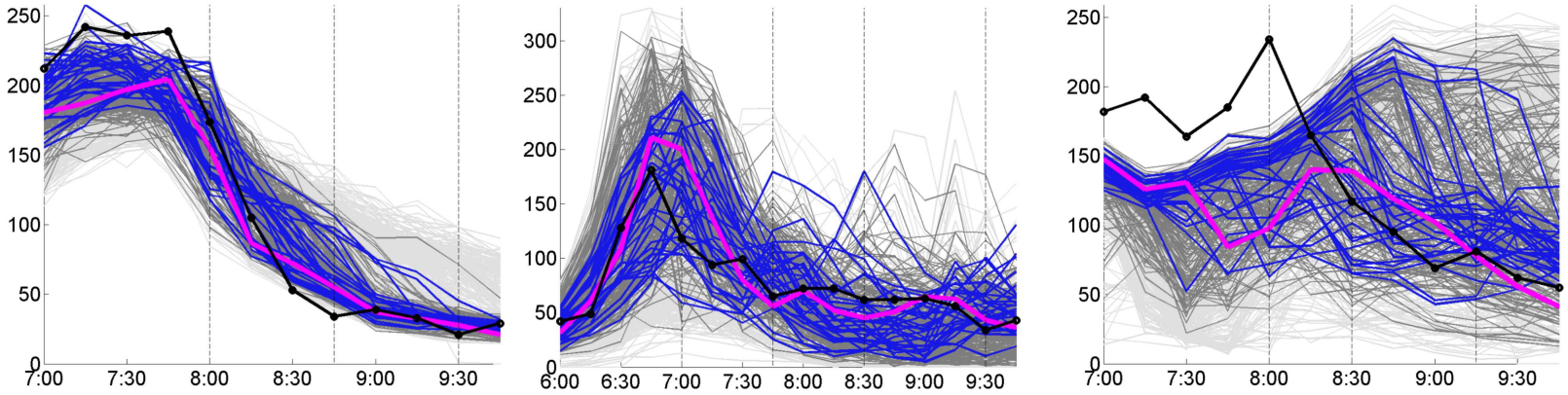




Persistent homology was used to describe the structure of the data.

One of the clusters (the largest, bottom right in figures) was found to have some topological complexity.

Example C motorway: The waves



(a) SE Entrace From B4100 S

(b) Gallows hill from Warwick by-pass to Banbury Road

(c) M40 J12 on-off slip

Runs generated using wave 1 (light grey), wave 2 (dark grey) and wave 3 (blue) parametrisations. The parameter configuration used by Warwickshire Council is shown in magenta and the observations by black circles. The simulator is run repeatedly for each parameter set and the mean of each configuration is plotted. Vertical dashed lines denote the times where the implausibility was examined. There were 4 time points of interest per location [3].

All in all

I have surveyed three examples of analysis of simulations I was involved in. Each project involved a team; this is what I did in each case:

A epidemiology [5] I was involved in the screening design stage, where I employed a sequential Morris design strategy in which I have worked previously [2].

B engine In this project I was fully involved, doing the design, running the simulations in ANSYS and using the data as a testbed for smooth models that I have also helped develop [1].

C motorway [3] I helped describe the region of low implausibility; this was descriptive statistics adapting others' cluster work [6] and persistence homology (former `pHom`) to describe clouds of points.

References

1. Bates et al. (2013) JSCS 84(11), 2453-2464.
2. Boukouvalas et al. (2014) Techno. 56, 422-431.
3. Boukouvalas et al. (2014) IEEE TITS 15(3), 1337-1347.
4. Gosling (2018) Sheffield ellicitation framework. Ellicitation. Springer.
5. Gosling et al. (2023-est) Sensitivity of epidemiological model (Chapter).
6. Hennig, Liao (2013). Appl. Statist. 62(3), 1–25.
7. Morris (1991) Techno. 33(2), 161-174.
8. Williamson, Vernon (2014) `arXiv:1309.3520v1`

Contact details:

`H.Maruri-Aguilar@qmul.ac.uk`