Sensitivity analysis and model calibration by method of Design of Experiment

The purpose of the design of experiments is to identify the effects of factors (i.e. parameters of interest) on the model outcome. This file reports step-by-step process of how we use this approach to conduct sensitivity analysis and model calibration.

Step 1: Determine number of runs and noise standard deviation

In the model simulation, the standard deviation of model output becomes lower when we run it multiple times (i.e. use different seeds). A larger number of runs increases the precision of result, however, consumes more time to compute. Therefore, we need first find a suitable number of runs for each simulation, in order to avoid unnecessary time complexity and get a more efficient calculation for later analysis. 50 times of runs using different random seeds are conducted with default parameter values. Experiment outputs are saved as CSV files and the calculation of standard deviation can be found in supplementary ZIP file called "S1 Result. std calculation".

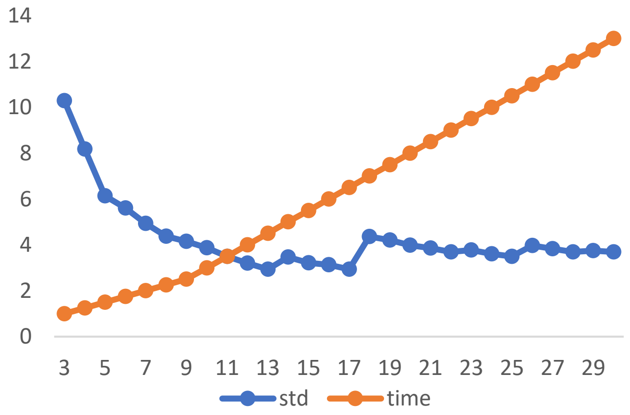

Fig.1: Plot of time consumed and corresponding standard deviation. X-axis represents number of runs. Y-axis represent the time measured in hours. Time refers to the total amount of time spent to run. Std is the standard deviation of model output, and it is rescaled.

Fig.1 shows that the standard deviation of results decreases with the increase in the number of runs. The intersection of time and standard deviation, which is 12 times of run, is expected to be adequate for the following simulation. It's corresponding standard deviation is 0.0157.

Step 2: Construct metamodels for model calibration and sensitivity analysis

Classical model calibration tests all possible values of parameter to find an optimal one that generates closest result to data. However, this approach takes too much time and appears to be infeasible in this study. Therefore, we do model calibration the other way round: A quadratic function is chosen as the metamodel to find the optimal value of sandfly infection factor. The key idea is that we assume the distance between output and data to be a function (i.e. metamodel) of the fitting parameter. As such, we just need a small number of trials (minimum 3 trials) and we can find this function by regression analysis. Nevertheless, one shortcoming has to be point out here, that we can only find the locally optimal value, rather than the global optimum. The manuscript has explained this method and results in the section 3.1. The regression analysis is saved in supplementary ZIP files "S2 Result. Model calibration".

In terms of sensitivity analysis, another metamodel of polynomial equation is constructed to represent the association between parameters and model output. Section 3.2 in the manuscript reported this analysis in detail. The function is also obtained by regression analysis and related results are saved in supplementary files "S3 Result. Sensitivity analysis".

Step 3: Add noise standard deviation

In sensitivity analysis, the effect of parameter is quantified by the coefficient of polynomial equation obtained from regression. There are two options to find the range for this coefficient, either by running each parameter set for 12 times, or by using statistical method to add a noise standard deviation. Here, we choose the latter option and following gives the prove of estimated coefficient interval:

Prove for estimated coefficient interval in the sensitivity analysis

We already know that, given a set of regression data$\{(x_i, y_i): \mathrm{i} = 1,2, \dots, \mathrm{n}\}$ and a fitted model $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$, the fitted coefficient $\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$.

Where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$; $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

In sensitivity analysis, β_1 represents the sensitivity from samples of model results $\{y_i : \mathrm{i} = 1,2,\dots,\mathrm{n}\}$. The parameter sets to be input are denoted as $\{x_i : \mathrm{i} = 1,2,\dots,\mathrm{n}\}$. By adding stochasticity to model output, we plus and minus a noise standard deviation (std) to samples and get ranges $\{y_{i_{range}} : \mathrm{i} = 1,2,\dots,\mathrm{n}\}$ and $y_{i_{range}} \in [y_i - std, y_i + std]$. Now we want to find ${\beta_1}^{min}$ and ${\beta_1}^{max}$ for sample model results $y_{i_{range}}$, that is, to find ${S_{xy}}^{min}$ and ${S_{xy}}^{max}$.

$${S_{xy}}^{range} = (x_1 - \bar{x})(y_{1_{range}} - \bar{y}) + \dots + (x_{\frac{n+1}{2}} - \bar{x})(y_{\frac{n+1}{2}_{range}} - \bar{y}) + \dots + (x_n - \bar{x})(y_{n_{range}} - \bar{y})$$

We have $(x_1 - \bar{x}) < 0$, $\left(x_{\frac{n+1}{2}} - \bar{x}\right) = 0$, $(x_n - \bar{x}) > 0$

Take the absolute value of x part:

$${S_{xy}}^{range} = -|x_1 - \bar{x}|(y_{1_{range}} - \bar{y}) + \dots + \left|x_{\frac{n+1}{2}} - \bar{x}\right|(y_{\frac{n+1}{2}_{range}} - \bar{y}) + \dots + |x_n - \bar{x}|(y_{n_{range}} - \bar{y})$$

To find ${S_{xy}}^{range}$ minimum:

$${S_{xy}}^{min} = -|x_1 - \bar{x}|(y_1 + std - \bar{y}) + \cdots + \left|x_{\frac{n+1}{2}} - \bar{x}\right|\left(y_{\frac{n+1}{2}} - std - \bar{y}\right) + \cdots + |x_n - \bar{x}|(y_n - std - \bar{y})$$

$$= -|x_1 - \bar{x}|(y_1 - \bar{y}) - \mathrm{std}|x_1 - \bar{x}| + \cdots + \left|x_{\frac{n+1}{2}} - \bar{x}\right|\left(y_{\frac{n+1}{2}} - \bar{y}\right) - std\left|x_{\frac{n+1}{2}} - \bar{x}\right| + \cdots + |x_n - \bar{x}|(y_n - \bar{y}) - std|x_n - \bar{x}|$$

$$= -|x_1 - \bar{x}|(y_1 - \bar{y}) + \cdots + \left|x_{\frac{n+1}{2}} - \bar{x}\right|\left(y_{\frac{n+1}{2}} - \bar{y}\right) + \cdots + |x_n - \bar{x}|(y_n - \bar{y}) + std\left(-|x_1 - \bar{x}| - \left|x_{\frac{n+1}{2}} - \bar{x}\right| - |x_n - \bar{x}|\right)$$

$$= S_{xy} - std \sum_{i=1}^{n} |x_i - \bar{x}|$$

$$\beta_1{}^{min} = \frac{S_{xy}{}^{min}}{S_{xx}} = \frac{S_{xy}}{S_{xx}} - std \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{\sum_{i=1}^{n} |x_i - \bar{x}|^2} = \widehat{\beta_1} - std \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{\sum_{i=1}^{n} |x_i - \bar{x}|^2}$$

To find $S_{xy}{}^{range}$ maximum:

$$S_{xy}{}^{max} = -|x_1 - \bar{x}|(y_1 - std - \bar{y}) + \cdots + \left|x_{\frac{n+1}{2}} - \bar{x}\right| \left(y_{\frac{n+1}{2}} + std - \bar{y}\right) + \cdots +$$

$$|x_n - \bar{x}|(y_n + std - \bar{y})$$

$$= -|x_1 - \bar{x}|(y_1 - \bar{y}) + \mathrm{std}|x_1 - \bar{x}| + \cdots + \left|x_{\frac{n+1}{2}} - \bar{x}\right| \left(y_{\frac{n+1}{2}} - \bar{y}\right) + std \left|x_{\frac{n+1}{2}} - \bar{x}\right| + \cdots$$

$$+ |x_n - \bar{x}|(y_n - \bar{y}) + std|x_n - \bar{x}|$$

$$= -|x_1 - \bar{x}|(y_1 - \bar{y}) + \cdots + \left|x_{\frac{n+1}{2}} - \bar{x}\right| \left(y_{\frac{n+1}{2}} - \bar{y}\right) + \cdots + |x_n - \bar{x}|(y_n - \bar{y})$$

$$+ std \left(|x_1 - \bar{x}| + \left|x_{\frac{n+1}{2}} - \bar{x}\right| + |x_n - \bar{x}|\right)$$

$$= S_{xy} + std \sum_{i=1}^{n} |x_i - \bar{x}|$$

$$\beta_1{}^{max} = \frac{S_{xy}{}^{max}}{S_{xx}} = \frac{S_{xy}}{S_{xx}} + std \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{\sum_{i=1}^{n} |x_i - \bar{x}|^2} = \widehat{\beta_1} + std \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{\sum_{i=1}^{n} |x_i - \bar{x}|^2}$$

Therefore, the estimated range is $\beta_1{}^{\widehat{range}} \in [\widehat{\beta_1} - std \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{\sum_{i=1}^{n} |x_i - \bar{x}|^2}, \widehat{\beta_1} + std \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{\sum_{i=1}^{n} |x_i - \bar{x}|^2}]$