National Physical Laboratory

NPL Report MS 58

GOOD PRACTICE IN SENSITIVITY ANALYSIS FOR METROLOGY

J. VENTON, L. WRIGHT

DECEMBER 2024

Department for
Science, Innovation
& Technology

NPL Report MS 58

Good practice in sensitivity analysis for metrology

J. Venton, L. Wright

Data Analytics & Modelling Group
Data Science Department

NPL Report MS 58

ISSN 1754-2960

https://doi.org/10.47120/npl.MS58

National Physical Laboratory
Hampton Road, Teddington, Middlesex, TW11 0LW

This work was funded by the UK Government's Department for Science, Innovation & Technology through the UK's National Measurement System programmes.

Approved on behalf of NPLML by Peter Harris,

Science Area Leader for Data Analytics and Modelling

GLOSSARY

ANOVA	analysis of variance
DoE	design of experiments method
GSA	Global Sensitivity Analysis
OAT	"one at a time" method
PAWN	Sensitivity analysis method, named after authors Pianosi and Wagner

CONTENTS

TABLES

FIGURES

1 INTRODUCTION

This report is a deliverable of the 2022-2023 National Measurement System project “Tools for Trustworthiness” (NPL project number 127116). It provides an overview of current practice in sensitivity analysis and discusses how metrology can benefit from good practice in sensitivity analysis.

The report structure following the introduction and notation section is:

- Section 2 provides an overview of what sensitivity analysis is and why it is relevant and useful for metrology.
- Section 3 introduces the various classes of method and summarises some of the strengths and weaknesses of each class.
- Section 4 summarises some useful papers that act as introductions or reviews of current practice.
- Section 5 describes a test case used to compare two sensitivity analysis methods on two common benchmark functions and presents the findings.
- Section 6 summarises good practice in sensitivity analysis based on what has gone before.
- Section 7 highlights future topics for investigation that have been raised by the work.

1.1 SCOPE AND NOTATION

This report largely focusses on the sensitivity of model outputs to model inputs. In many cases of interest to metrology, the model in question will be a "measurement model" as defined by the Guide to the Expression of Uncertainty in Measurement (GUM)[1], i.e. a function which calculates a measurement result when given values of its input quantities. However, there are many models in use within metrology that are not thought of by their users as being measurement models. Models used in design of equipment or prediction of the behaviour of systems are not commonly considered as measurement models despite them generating numerical estimates of physical quantities. The language in this report therefore uses "model" rather than "measurement model" in order to include both groups of people.

The following notation conventions are adopted: Suppose we have a model

$$Y = F(\mathbf{X}) \tag{1}$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_N)^\top$ is a vector of input quantities and Y is a single output quantity. The methods discussed in this report can also be applied directly to models that have vectors of output quantities to obtain the sensitivity of individual outputs.

Values of an input or an output quantity are denoted in lower case, so that x_n is a value of X_n the quantity. The use of this notation becomes important when distinguishing between a quantity that is a function of the input quantities $X_i, i = 1, 2, \ldots, N$, and a value that has been calculated using a set of sample values $x_i, i = 1, 2, \ldots, N$ of those quantities.

Throughout, $\mathbb{E}(Y)$ denotes the expectation (mean) of Y and $\mathbb{V}(Y)$ denotes the variance of Y, modelled here as a random variable.

2 WHAT IS SENSITIVITY ANALYSIS?

Sensitivity analysis and uncertainty evaluation are frequently conflated in scientific literature. A meta-analysis of highly cited literature reported in [2] showed that of 280 papers published between 2012 and 2017 that included "sensitivity analysis" in the title, abstract or keywords, 24 were in fact carrying out pure uncertainty analysis. Uncertainty evaluation is a cornerstone of metrology and therefore the distinction between sensitivity and uncertainty is perhaps more straightforward for a metrology audience.

The opening sentence of Sensitivity Analysis by Saltelli et al [3], a widely cited reference, states:

> Sensitivity analysis is the study of how the variation in the output of a model (numerical or otherwise) can be apportioned, qualitatively or quantitatively, to different sources of variation, and of how the given model depends upon the information fed into it.

The aspect of this definition that draws the main distinction between uncertainty evaluation and sensitivity analysis is the retention of the link between the variation of the inputs and the variation of the outputs. This link is present for some forms of uncertainty evaluation, such as the use of the standard approach outlined in the GUM [1] where the use of sensitivity coefficients enables comparison between the effects of different inputs, but it is not present when sampling methods such as the Monte Carlo method are used.

Sensitivity analysis is useful to metrology because understanding the links between the variation of the inputs and the variation of the output can help identify which inputs would reduce the variance of the output most significantly if their variances or uncertainties were reduced. This knowledge is useful for measurement because it enables the metrologist to identify which aspects of their process would bring most benefit if they were controlled more tightly. It also brings benefit for modellers because it identifies the model inputs that could reduce the uncertainty associated with the model output if their associated uncertainties were reduced. In both cases, the knowledge enables the user to focus their efforts in areas where more knowledge or better control will bring the largest benefits.

Sensitivity analysis also benefits metrology when planning comparison of measurements of the same measurand and the same reference object performed using different devices. For the type of intercomparisons that are carried out between National Metrology Institutes the analysis is perhaps less likely to bring benefit as the measurements are well understood, but for applications where metrology is less well established an initial sensitivity analysis can help to inform the trial design. An example is intercomparison of magnetic resonance imaging scanners, where it is difficult to deconvolve scanner-to-scanner variability from patient-to-patient variability, and scanners have many different settable parameters that may affect results. An initial sensitivity analysis can provide guidance on planning a series of measurements to assist with this deconvolution by identifying the key parameters that will affect comparability of measurements made on different devices.

Sensitivity analysis has become more important as predictive models have become more complicated and data processing chains have become more elaborate. Sensitivity analysis can also be a useful "sanity check" as an early step in model validation. If an expert in the system being simulated expects a model output to increase when a given parameter increases, but a sensitivity analysis shows no response of the model to the input then it is likely that the model contains an error. Conversely, being able to explain observed results brings additional confidence that the model is working correctly.

A recent example comes from an EMPIR project, MedalCare [4], which used a multi-step modelling process illustrated in figure 1 to generate features associated with electrocardiographs (ECGs) that a clinician would use to diagnose a patient. The approach used two separate computationally expensive biophysical models to calculate the electrical signal generated at the surface of the body during a heartbeat; a simple linear scaling and stitching to generate a realistic synthetic ECG signal; and a software toolkit in common clinical use to extract the features of interest. The biophysical models had a total of 19 inputs between them that were considered as varying. The models could then be used to generate the 12 ECG signals typically obtained during a 12-lead ECG recording by placing electrodes at various positions on the body. The toolkit could extract many different features from the signal in each lead, such as peak heights or intervals between particular parts of the signal. This problem was therefore a multi-stage process with many inputs and many outputs, and was too complicated for intuition alone to be a reliable guide to which input quantities affected the outputs most. The project team carried out a sensitivity analysis to identify which input quantities most affected key features. Many of the most significant sensitivities could be explained from anatomical considerations, and allowed the model owners to understand which aspects of their models might most benefit from more knowledge of inputs of the underlying processes. Sensitivity to a model input is likely to indicate sensitivity to the biophysical process that the model input describes, and in some cases a change in the way the process is described mathematically may bring more benefit than lowering the uncertainty associated with the input.

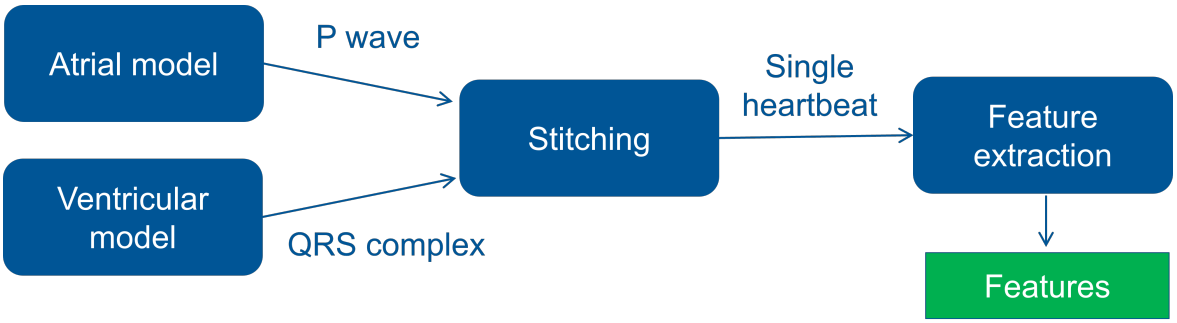


Figure 1: Sketch of MedalCare modelling and processing chain.

It is worth noting that the above definition of sensitivity analysis is not restricted to computer models. It is possible to carry out sensitivity analysis on real systems by setting values of control parameters and analysing the results. This topic has been addressed in a previous NPL report [5]. In general the main differences between sensitivity analysis for computational models and for real systems are:

- the effort required to conduct real world tests often reduces the number of tests that can be run and hence the number of results available for analysis, whereas many (but

not all) computational models can be evaluated repeatedly more quickly and easily than the real world equivalents, and

- repeated instances of a real world test are likely to bring varying results even when control parameters are held constant whereas repeated evaluations of a deterministic computer model will produce identical results.

Both of these differences affect which analysis methods are most suitable for conducting a sensitivity analysis. It is also interesting to consider the possibility of combining information from properly validated simulations with information from tests to make the most of information from both sources.

3 SENSITIVITY ANALYSIS METHODS

Many metrologists are familiar with the idea of sensitivity analysis via the sensitivity coefficients used in the standard GUM uncertainty evaluation methodology [1]. The GUM approach requires definition of a measurement function that calculates the value of the measurand (the output quantity of the model) when given values of a set of input quantities. The square of the combined standard uncertainty associated with the measurand is given by summing the standard uncertainties associated with the input quantities in weighted quadrature. The weights used are the sensitivity coefficients, given by the partial derivative of the measurement function with respect to each input quantity, evaluated at the estimated of all of the input quantities.

This approach can be considered as approximating the measurement function as a linear function, and applying the fact that for a linear function

$$Y = a_1 X_1 + a_2 X_2 + \ldots a_N X_N \tag{2}$$

the variances are linked by

$$\mathbb{V}(Y) = a_1^2 \mathbb{V}(X_1) + a_2^2 \mathbb{V}(X_2) + \ldots a_N^2 \mathbb{V}(X_N) \tag{3}$$

This approach gives one measure of sensitivity: the partial derivative of the function with respect to the input. For a linear function this measure is entirely adequate, because the derivatives are constant for all values of the inputs. The measure is not adequate for nonlinear functions such as that plotted in figure 2, however, where the partial derivative changes magnitude and sign. The partial derivative for a nonlinear function provides useful information about local behaviour, but does not give a quantified indication of the overall effect of the variance of the input on the variance of the output. Such an indication requires a global sensitivity method.

This observation highlights the difference between local sensitivity analysis and global sensitivity analysis. Both approaches are useful, but provide information that can be used in different ways. Local sensitivity analysis provides granularity of information in particular regions of the space spanned by the model inputs, which is more useful when the model is being evaluated at a specific point in that space. Global sensitivity analysis gives an overall evaluation of the effect of each input. Care must be taken when interpreting local sensitivity coefficients on their own. Sensitivity coefficients associated with model inputs that have

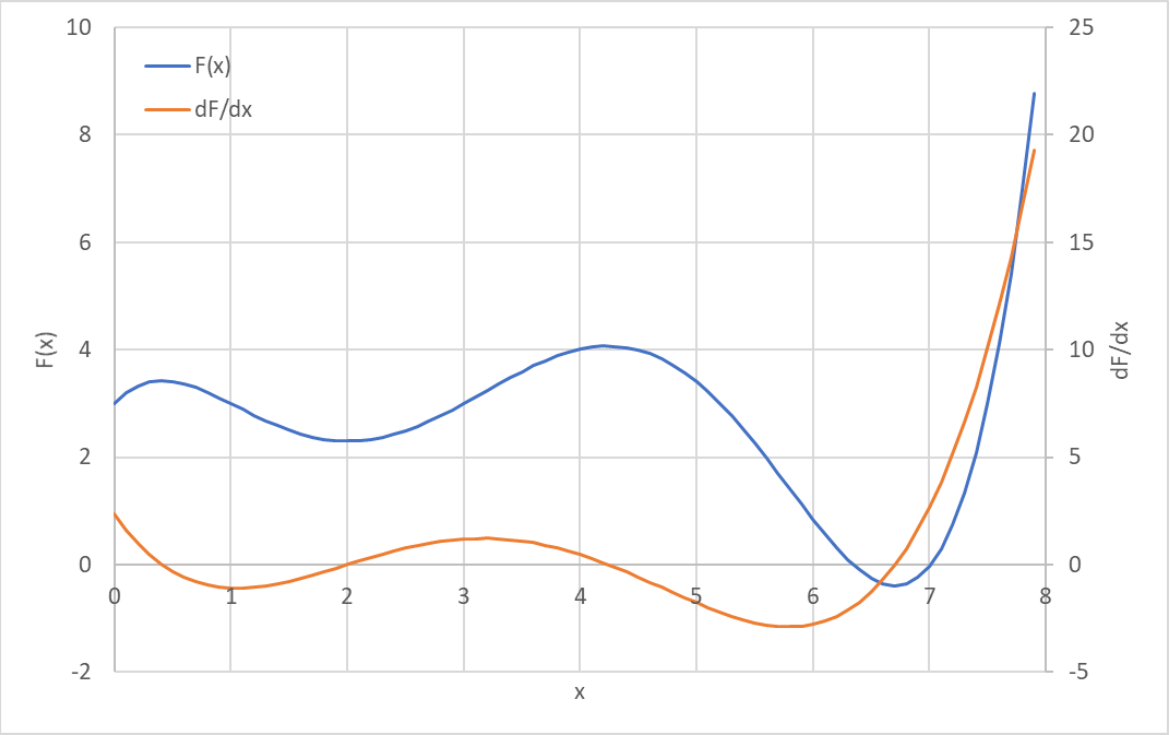


Figure 2: A nonlinear function (blue line, left axis) and its derivative (orange line, right axis).

different units cannot be meaningfully compared, not least because the coefficients will also have different units. The coefficients only really become meaningful when it is known how much the input is varying; a large sensitivity coefficient associated with an input that does not vary significantly will have little effect on the model output. Many metrics for quantitative local sensitivity analysis require definition of a perturbation associated with each input (see section 2 of [6] for several examples of commonly used perturbations) in order to introduce meaning. Whilst this is a practical approach, it means the analyst has to select a set of additional parameters that may affect the result and may be fairly arbitrary and so any such approach must be reported carefully to ensure all parameters chosen are explicitly stated.

3.1 Simple numerical methods

One of the most intuitive forms of sensitivity analysis is the "one at a time" or OAT approach. As the name suggests, this approach varies one input of a model at a time. The user typically defines a nominal or central value of each input quantity (typically the expected value), $\mathbf{x}_C = \{x_1^C, x_2^C, ..., x_N^C\}$ and specifies an upper and lower extreme value for each input,

x_i^U and $x_i^L, i = 1,2,...,N$. The model is then evaluated at a set of $2N$ points, giving

$$\begin{aligned} y_1^U &= F(x_1^U, x_2^C, ..., x_N^C), \\ y_1^L &= F(x_1^L, x_2^C, ..., x_N^C), \\ y_2^U &= F(x_1^C, x_2^U, ..., x_N^C), \\ y_2^L &= F(x_1^C, x_2^L, ..., x_N^C), \\ &\cdots \\ y_N^U &= F(x_1^C, x_2^C, ..., x_N^U), \\ y_N^L &= F(x_1^C, x_2^C, ..., x_N^L), \end{aligned}$$

and a measure of the sensitivity of the output to the value of the input X_i is given by

$$OAT_i = \frac{y_i^U - y_i^L}{x_i^U - x_i^L}. \tag{4}$$

This quantity is fairly clearly analogous to the partial differential coefficients used in the GUM approach and is similarly based on the underlying assumption that the model can be well approximated by a weighted linear combination of the input quantities as in expression 2. This approach neglects the possibility that interactions (i.e. products of inputs or products of functions of inputs) between the different inputs could affect the model output. It also samples within an extremely limited region of the space of input quantities. Despite its limitations, the method is popular, possibly because of its intuitive simplicity. The meta-analysis in [2] showed that 34 % of the papers surveyed used OAT methods.

A more advanced approach is to apply design of experiments (DoE) methods to define a set of values for $\mathbf{X}$ and use appropriate combinations of the resulting model outputs to estimate sensitivity through analysis of variance (ANOVA). Like the OAT method, DoE approaches define a set of values for each input quantity (known as levels) and constructs input values using these values. Unlike OAT, DoE methods vary more than one quantity at once. So for instance if there are three inputs and each input has two levels (U and L say) then the most complete DoE method, a full factorial design, would generate the following values:

$$\begin{aligned} y_1 &= F(\{x_1^U, x_2^U, x_3^U\}), \\ y_2 &= F(\{x_1^L, x_2^U, x_3^U\}), \\ y_3 &= F(\{x_1^U, x_2^L, x_3^U\}), \\ y_4 &= F(\{x_1^L, x_2^L, x_3^U\}), \\ y_5 &= F(\{x_1^U, x_2^U, x_3^L\}), \\ y_6 &= F(\{x_1^L, x_2^U, x_3^L\}), \\ y_7 &= F(\{x_1^U, x_2^L, x_3^L\}), \\ y_8 &= F(\{x_1^L, x_2^L, x_3^L\}), \end{aligned}$$

and would then use expressions (for this specific case) such as

$$\hat{y} = \frac{1}{8}\sum_{i=1}^{8} y_i, \tag{5}$$

$$SST = \sum_{i=1}^{8}(y_i - \hat{y})^2, \tag{6}$$

$$SSF = 4((\frac{y_1+y_3+y_5+y_7}{4} - \hat{y})^2 + (\frac{y_2+y_4+y_6+y_8}{4} - \hat{y})^2), \tag{7}$$

$$\eta_1 = \frac{SSF}{SST}, \tag{8}$$

to calculate the partial sensitivity coefficient η_1, and similar expressions can be obtained for each of the input quantities. There are two grouped terms in the right hand side of expression 7 because the test uses two levels, and the individual terms in each group are evaluations of the function with the same value of X_1. It is also worth noting that for this sample,

$$\begin{aligned} SST &\propto \mathbb{V}(Y), \\ SSF &\propto \mathbb{V}(\mathbb{E}(Y|X_1)), \end{aligned}$$

where $\mathbb{V}(\mathbb{E}(Y|X_1))$ is the variance of the expected value of Y when conditioned on the value of X_1, a quantity relevant to the Sobol coefficients discussed below.

Because this approach varies more than one input at once, similar coefficients can be determined for interactions between inputs. It is clear that as the number of inputs increases, the number of model evaluations required to evaluate every possible interaction increases exponentially (2^N for a full factorial design compared to $2N$ for an OAT approach). Various different partial factorial designs have been developed to reduce the computational burden whilst still giving the most important information. A previous report [5] has described the DoE and ANOVA approach in more detail with a worked example and the interested reader is referred to that document. DoE is also well described in the NIST Engineering Statistics Handbook [7].

A number of methods similar to DoE, known as screening methods, have been developed with a focus on identifying the least significant inputs so that the problem can be simplified by fixing the values of those inputs. Many of these methods seek to minimise the number of times the model must be evaluated, and often require a monotonic model. A slightly more flexible screening method is the method of Morris [8], which evaluates a series of OAT indices at sample points chosen within the input space. The points are chosen in a structured manner with some random aspects. The method is described in section 4 of [3], which covers screening methods in general in some detail.

The DoE approach is an advance on OAT because it can quantify variance caused by interactions of variables. It was noted above that the terms in the analysis are proportional to $\mathbb{V}(Y)$ and $\mathbb{V}(\mathbb{E}(Y|X_i))$ (where the conditioning may be over more than one variable if interaction terms are of interest). For the case with three inputs above, the discrete sums are approximations to the integrals

$$\mathbb{E}(Y|X_1 = x_1) = \int F(x_1, X_2, X_3)\, dX_2 dX_3, \tag{9}$$

$$\mathbb{V}(\mathbb{E}(Y|X_1)) = \int (\mathbb{E}(Y|X_1) - \mathbb{E}(Y))^2\, dX_1. \tag{10}$$

This approach will be discussed in more detail in the section on Sobol indices 3.2 The use of levels effectively attempts to approximate these integrals using a quadrature product rule. Gaussian quadrature rules that use n points give exact values of integrals for polynomials of degree $2n-1$ or less and therefore provide a sensible choice of points and weights to use when approximating such integrals. The link between the ANOVA approach and quadrature is rarely explicitly stated however, and users of the method tend to select endpoints and midpoints of bounding intervals for parameters. This choice of points is likely to affect accuracy for nonlinear models.

A simple example was tested by constructing a series of polynomials of the form

$$Y = \frac{1}{N^2} \sum_{m,n=1}^{N} X_1^m X_1^n,\ X_1, X_2 \in [0,1] \tag{11}$$

for various degrees of N and evaluating η_1 using large scale Monte Carlo simulation (MCS, regarded as ground truth), Gaussian quadrature product rule with two points (GQ2) and with three points (GQ3), and a DoE approach with two levels (DoE2: ends of the interval [0, 1]) and with three levels (DoE3: ends and midpoint). The results are shown in table 1, and they show that Gaussian quadrature outperforms the more simplistic DoE approach for the higher order polynomials. Whilst the level of error is low for the models used here, it is possible that other forms of nonlinear function would cause larger errors.

Table 1: Values of η_1 for polynomials of different order N calculated using Monte Carlo simulation (MCS), two- and three-point Gaussian quadrature (GQ2, GQ3), and Design of Experiments full factorial trials with two (DoE2) and three (DoE3) levels.

N	1	2	3	4	5
MCS	0.49	0.48	0.46	0.45	0.43
GQ2	0.49	0.48	0.46	0.46	0.45
GQ3	0.49	0.48	0.46	0.45	0.43
DoE2	0.47	0.44	0.42	0.41	0.40
DoE3	0.48	0.46	0.43	0.41	0.40

The restrictions that the DoE and ANOVA method imposes may not matter for a large proportion of models. Some references [9] mention the "sparsity of effects principle" which states that most physical systems are dominated by main effects and low-order interactions. Whilst this may be true, for a complicated model there is no simple way to determine whether the principle is valid for the problem of interest, and to assume that it is valid runs the risk of reaching incorrect conclusions.

3.2 VARIANCE-BASED METHODS

The quote from the book "Sensitivity Analysis" above stated that sensitivity analysis was concerned with the variation in the output of a model. It is therefore unsurprising that a class of methods has been developed that focus on the variance of the model output, and how to identify which input quantities contribute the most to that variance.

ANOVA, as described above, uses this approach and effectively imposes a model linear in each input quantity. Other methods in this class impose fewer assumptions. One method that generalised the concept of apportioning the variance of the model outputs to the various inputs via conditional expectation was developed by Sobol [10]. Other similar methods were

developed independently at around the same time (as explained in section 4 of [6]), but the Sobol approach is the most widely used.

It is useful at this point to introduce the notation

$$\int d\mathbf{x}_{\sim i} = \int\cdots\int dX_1\ldots,dX_{i-1},dX_{i+1},\ldots dX_N, \tag{12}$$

so that the subscript $\sim i$ means that the integration takes place over the set of variables excluding X_i and holding the value of X_i fixed. The same notation can be used to exclude more than one variable from the integration, for instance $d\mathbf{x}_{\sim i,j}$.

The Sobol method effectively decomposes the model function F as a sum of functions of increasing dimensionality,

$$F(X_1,X_2,\ldots,X_N) = F_0 + \sum_{i=1}^{N} F_i(X_i) + \sum_{1\le i<j\le N} F_{i,j}(X_i,X_j), + \ldots + F_{1,2,\ldots,N}(X_1,X_2,\ldots,X_N). \tag{13}$$

It can be shown that

$$F_0 = \int\cdots\int F(\mathbf{X})\,d\mathbf{X}, \tag{14}$$

$$F_i(x_i) = -F_0 + \int\cdots\int F(X_1,X_2,\ldots,X_{i-1},x_i,X_{i+1},\ldots,X_N)\,d\mathbf{X}_{\sim i}, \tag{15}$$

and similar expressions can be obtained for higher order terms. The variance of Y is then given by

$$\mathbb{V}(Y) = D = \int F^2(\mathbf{X})d\mathbf{X} - F_0^2. \tag{16}$$

The partial variance due to X_i alone (i.e. not in interaction with any other inputs) is given by

$$D_i = \int F_i^2(X_i)dX_i, \tag{17}$$

and the ratio

$$S_i = \frac{D_i}{D} \tag{18}$$

is a measure of how much of the variability is due to X_i alone, and is called the first-order sensitivity index or main effect index. This quantity can also be written as

$$S_i = \frac{\mathbb{V}_{X_i}[\mathbb{E}(Y|X_i = x_i)]}{\mathbb{V}(Y)} \tag{19}$$

where

$$\mathbb{E}(Y|X_j = x_j) = \int\cdots\int F(X_1,X_2,\ldots,X_{j-1},x_j,X_{j+1},\ldots,X_N)dX_1\ldots,dX_{j-1},dX_{j+1},\ldots dX_N, \tag{20}$$

is a function of X_j only. The notation above leads to the simpler form

$$\mathbb{E}(Y|X_j = x_j) = \int F(X_1,X_2,\ldots,X_{j-1},x_j,X_{j+1},\ldots,X_N)d\mathbf{X}_{\sim j}. \tag{21}$$

Similar definitions follow for indices associated with higher-order terms, and it is useful to note that

$$\sum_{i=1}^{N} S_i + \sum_{1\le i<j\le N} S_{i,j} + \ldots + S_{1,2,\ldots,N} = 1 \tag{22}$$

The total sensitivity index for X_i, S_i^T, is defined as the sum of all indices that involve X_i, and gives a measure of the overall effect of an input on the model output. This index can be defined as

$$S_i^T = \frac{\mathbb{E}_{\mathbf{X}_{\sim \mathbf{i}}}[\mathrm{Var}_{X_i}(Y|X_j = x_j \forall j \neq i)]}{\mathbb{V}(Y)} \tag{23}$$

$$= 1 - \frac{\mathbb{V}_{\mathbf{X}_{\sim i}}[\mathbb{E}_{X_i}(Y|X_j = x_j, \forall j \neq i)]}{\mathbb{V}(Y)} \tag{24}$$

$$= 1 - \frac{D_{\sim i}}{D}. \tag{25}$$

These indices have several useful properties. They are normalised such that values should always lie between 0 and 1, so that indices generated during different analyses can be compared. A value of 0 indicates that the input has no effect on the variance of the output, and a value of 1 indicates that the variance of the output is caused solely by the input (and its interactions if the total index is considered). The approach makes no assumptions about the linearity, continuity, or monotonicity of the underlying model. It only requires that the variance is calculable. The indices can therefore be calculated for any type of model whether the user has knowledge of the equations being solved or not.

The formulae defining these indices are given in terms of integrals of functions. For "black box" models and complicated data processing chains, where the link between inputs and outputs cannot be written as a simple integrable function, these integrals must be evaluated numerically. Numerical integration requires repeated evaluation of the model with varied values of the input quantities, with the accuracy of the estimate thus generated becoming more accurate as the number of model evaluations increases. For computationally expensive models this repeated evaluation is time-consuming and can lead to a limitation on the number of models that can be run, and hence a limitation on the accuracy of the estimate of the integral value.

Work by Saltelli [11] has shown that an intelligent choice of the input quantity values can enable the main effect indices and the total effect indices to be calculated in a computationally efficient manner. There are two aspects to this computational efficiency: definition of the combinations of one sampled value of each input that are used to create the sets of input values that are required to run the model, and generation of a space-filling set of sampled values of each of the inputs.

The combinations of inputs are constructed by creating two M by N sampling matrices, where M is the number of samples and N is the number of variables, A and B. The j^{th} column of A and the j^{th} column of B contain sampled values of the j^{th} variable X_j. A hybrid matrix can then be created by replacing the j^{th} column of A by the j^{th} column of B. This matrix is written as $A_B^{(j)}$, and $B_A^{(j)}$ can be defined similarly by replacing one column of B with a column of A.

Several estimators based on numerical approximation for the integrals $\mathbb{V}_{X_i}[\mathbb{E}(Y|X_i = x_i)]$ and $\mathbb{E}_{\mathbf{X}_{\sim \mathbf{i}}}[\mathbb{V}_{X_i}(Y|X_j = x_j \forall j \neq i)]$ are discussed in [11]. All of the estimators are constructed by evaluating the function of interest at the input values defined by the rows of the matrices A,

B, $A_B^{(i)}$, and $B_A^{(i)}$. The work in [11] shows that the most efficient method to calculate all N of the S_i values and all N of the S_i^T values is by using the approximations

$$\mathbb{V}_{X_i}[\mathbb{E}(Y|X_i = x_i)] \approx \frac{1}{M}\sum_{j=1}^{M} F(B)_j(F(A_B^{(i)})_j - F(A)_j) \tag{26}$$

and

$$\mathbb{E}_{\mathbf{X}_{\sim i}}[\mathbb{V}_{X_i}(Y|X_j = x_j \forall j \neq i) \approx \frac{1}{2M}\sum_{j=1}^{M}(F(A)_j - F(A_B^{(i)})_j)^2 \tag{27}$$

where the notation $F(A)_i$ denotes evaluating the function F using the input values taken from the i^{th} row of the matrix A.

The use of these numerical approximation methods means that the results obtained are dependent on the number of samples used to calculate the indices. It is important to carry out a convergence check, at minimum by evaluating the indices sequentially using an increasing sample size and plotting index value against sample size to look for convergence. An example is shown in figure 3, where values show a reasonable level of convergence for sample sizes above 700.

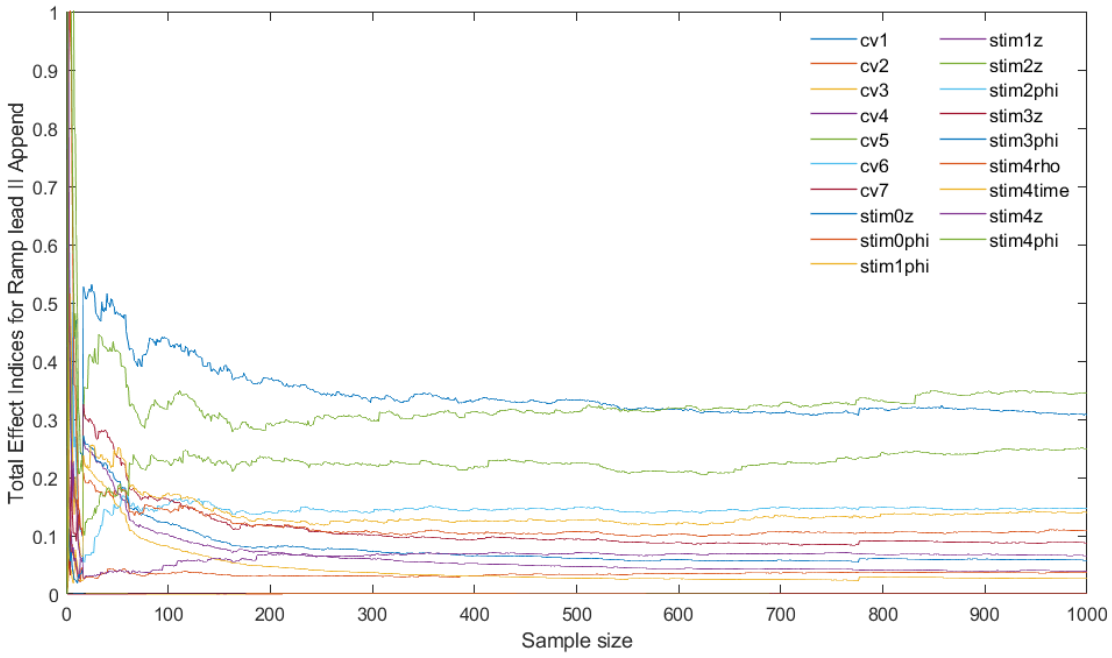


Figure 3: Plot of calculated total index values for different sample sizes, showing convergence. The underlying model, used in MedalCare [4], calculates features of a simulated electrocardiogram based on the values of 14 input quantities.

It should also be noted that the use of numerical techniques may lead to values of indices that are either negative or greater than 1. This possibility is acknowledged in the literature; see for example section 3.1 of [12], which states:

Sobol indices might eventually take on values outside of the range [0,1] due to numerical artefacts created during the computation. This is widely known amongst sensitivity analysts and managed by considering $S_T < 0 \approx 0$ and $S_T > 1 \approx 1$.

This approach is not advisable. "Numerical artefacts" should not be present if the numerical approach uses a sufficiently large sample size for convergence to occur, and it is not possible to predict from a single observation of an unconverged value what the converged value will be. This further highlights the importance of checking for convergence.

3.3 Distribution-based methods

One of the major criticisms of Sobol indices are that the only measure of variation that they consider is variance. For distributions that are not Gaussian, the variance is not sufficient to characterise the shape of the distribution fully. As an example, compare the two probability density functions in figure 4. They have the same mean and standard deviation but are very different distributions.

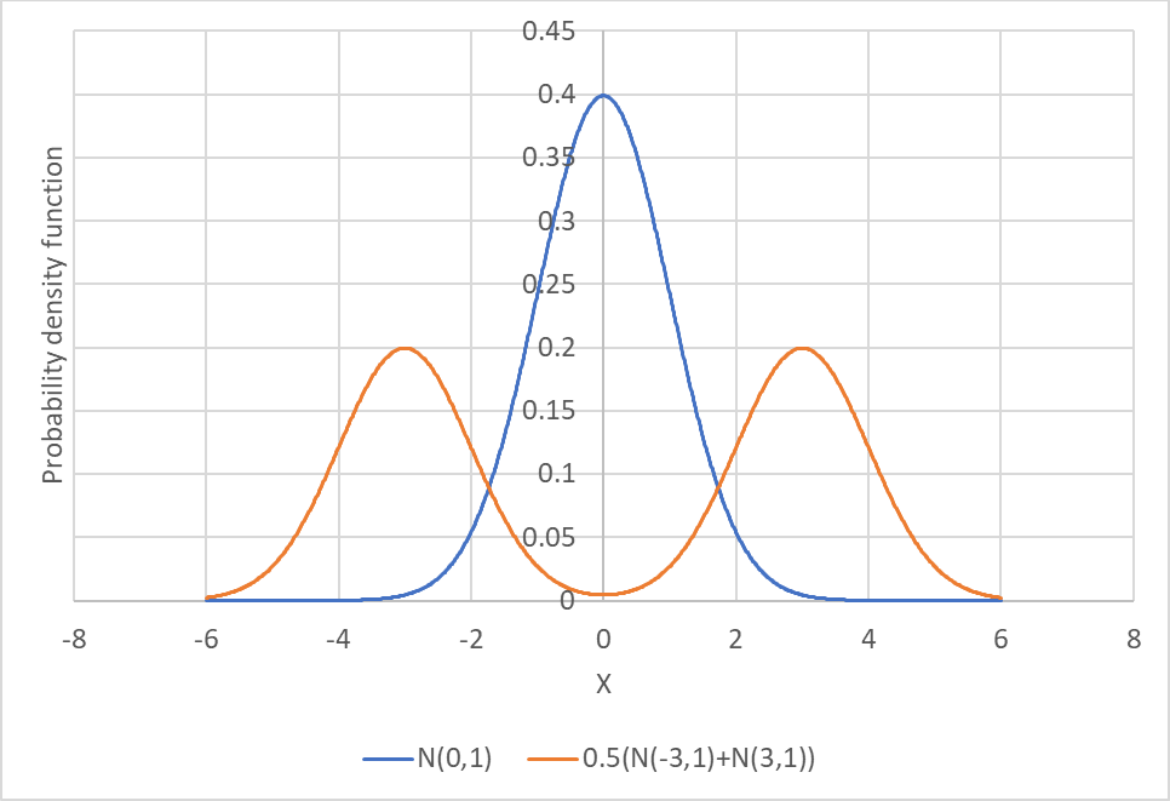


Figure 4: Two distributions with the same mean and standard deviation.

A number of methods that consider the effect of varying one or more input quantities on the overall distribution of the model output have been developed in response to this criticism. This class of methods is an area of ongoing research with new approaches being developed.

One of the first papers on distribution-based approaches [13] starts by specifying four properties that a sensitivity index should have: it should be global, it should be quantitative, it should impose no assumptions about the underlying model, and it should not refer to a particular moment of the distribution of the output quantity. The proposed index, the δ index, is based on consideration of how the probability density function (pdf) of the output Y changes conditional on one (or more) of its inputs. Let $g_Y(y)$ be the pdf of Y, $g_{Y|X_i}(y,x_i)$ be the corresponding density conditional on the value of X_i, and $g_{X_i}(x_i)$ be the (marginal) pdf of the

input quantity X_i. Then the δ index is defined as

$$\delta_i = \frac{1}{2}\int g_{X_i}(x_i)\left[\int |g_Y(y) - g_{Y|X_i}(y,x_i)|dy\right]dx_i. \tag{28}$$

This quantity is half of the expected value of the area between the pdf of Y and the pdf of Y conditional on X_i. The halving of the area ensures that the index has a maximum value of 1 (as would occur if the pdfs were completely disjoint). This index clearly has a value of 0 if changing the value of X_i has no effect on the pdf of Y, so it is a normalised index. The definition above can be extended to look at combinations of inputs by using joint marginal distributions instead of $g_{X_i}(x_i)$ and conditioning Y on multiple variables. It can be shown that if the definition is extended to the full set of parameters, then $\delta_{123\ldots N} = 1$ for any model. The paper provides several examples, including comparison of the δ and Sobol indices for a commonly used test problem. This comparison highlights the fact that the input that most influences the variance of an output is not necessarily the same as the input that most influences the pdf of an output. The indices are not in competition or contradiction with one another, but they tell you different things about the relationship between the inputs and outputs of the model.

An earlier paper [14] uses a similar approach by focussing on conditional pdfs, but fixes the conditioning variable at its mean rather than taking the expectation over all possible values, and uses the Kullback-Liebler entropy to quantify the difference between the pdfs. This approach is useful for consideration of local sensitivity by restricting the limits of integration to a sub-region of the full input space, but suffers from a lack of normalisation and the use of a single value of X_i (although this could be avoided by evaluating the index repeatedly for different values of X_i).

One of the main challenges associated with practical use of the δ index is its reliance on the pdf. For the majority of models, particularly in metrology applications, it is not possible to define an analytical expression for $g_Y(y)$, let alone the conditional distributions required to calculate 28. The distribution of Y is typically constructed numerically from repeated evaluations of the model $F(\mathbf{X})$. Accurate construction of a pdf generally requires a large number of model evaluations, and so the double integral in expression 28 will potentially require a very large number of evaluations. Methods for efficient calculation [15] have been proposed, but other authors have altered the computational expense by shifting their attention to the use of cumulative distribution functions (cdfs) instead of pdfs.

One such index is the PAWN index [16]. The authors start from a list of desirable properties for a sensitivity index. In addition to those listed above (as defined in [13]), they state that an index should not be conditional on an assumed value of any model input, should be easy to interpret, should be easy to compute, and should be stable or at least its stability should be possible to assess. Based on these requirements, the authors propose an index of the form

$$T_i = \mathrm{stat}_{x_i}(\max_y |G_Y(y) - G_{Y|X_i}(y,x_i)|), \tag{29}$$

where stat here denotes any statistic such as the mean, maximum or median, over all values of x_i, and G_Y is the cdf of Y. The term $\max_y |G_Y(y) - G_{Y|X_i}(y,x_i)|$ is the Kolmogorov-Smirnov or K-S statistic and is the largest difference in (cumulative) probability between the conditional and unconditional cumulative distribution functions.

As with the δ index, the PAWN index is equal to zero when Y is independent of X_i and always takes a value less than or equal to 1. The variety of possible statistics the index can

use means that care must be taken when reporting values and when comparing indices to literature values. The authors provide a schematic diagram illustrating the simple calculation approach that can be used if the index is calculated by new evaluations of the model (rather than use of historical data), illustrating their claim of ease of calculation. They also highlight the use of the method to identify non-influential parameters by application of the two-sample Kolmogorov-Smirnov test to test whether the conditional and unconditional distributions are the same. The index is computed for several examples including a standard test function, but comparisons with other indices are not offered. A subsequent paper [17] by the same authors suggests a method of calculating the PAWN index using historical data, by grouping data. This paper also carries out an exploration of how the total number of model evaluations and the number of groups they are put into affects the index result.

At least one paper [12] (by authors who have published regularly on Sobol indices) has criticised the use of PAWN and has used examples to show that the choice of the parameters used to estimate the index can lead to biased results, which they state does not occur for Sobol indices. The developers of PAWN have provided an online response [18] which showed that the low quality results were associated with poor choices of parameter combinations that a user who understood the method would be unlikely to use.

Recent authors [19] have offered a method of combining variance-based and distribution-based approaches. The authors suggest that this combined approach allows for separation of main, interaction and total effects, which allows for a more detailed understanding of the underlying model, whilst considering the effects on the whole of the distribution rather than focussing on variance. The method as described can determine a combined measure from model evaluations from a general sampling of the input space rather than requiring a specific sampling approach. It combines use of the Sobol main effects index and evaluation of the δ index (requiring fitting of a kernel density function), which is taken to be representative of the interaction terms not captured by the Sobol main effects index. The approach is tested on several commonly examined functions, and repeatability and convergence with increasing sample size are both examined.

3.4 METAMODELS

Many of the methods described above require repeated evaluation of the model $F(\mathbf{X})$ for sampled values of $\mathbf{X}$, and in particular reliable evaluation of the results of the methods requires a sample size that produces a converged and stable value of the result. Many models and data processing chains used in metrology are computationally expensive and may take minutes or hours to run. The requirement for a converged result can make some of the methods discussed too computationally expensive for practical usage. Many developers of methods have identified the most efficient numerical approaches for evaluating their indices of interest [11].

An alternative approach is to investigate the use of metamodels. A metamodel is a model that approximates the behaviour of the full model but takes a fraction of the time to evaluate. Metamodels can potentially replace computationally expensive models in any application, including within routines for Monte Carlo sampling, optimisation, or data assimilation. High fidelity models that are used across all areas of physics, and particularly those that use numerical approximation techniques to solve partial differential equations, are very often computationally expensive. The use of such models in robust design, where optimisation and uncertainty evaluation are important, and digital twins, where data assimilation and uncertainty evaluation are key, makes metamodels a very active area of research.

This report will not give a complete overview of the state of the art in metamodelling (for a short introduction see section 5 of [20]), but two papers that use metamodelling are of interest.

The first [21] takes a Bayesian approach to probabilistic sensitivity analysis. The authors discuss the analogy between regression analysis and the functional decomposition given in expression 13, and consider variance-based sensitivity analysis as equivalent to estimation of the expected square error when carrying out regression when one or more of the input variables is perfectly known. This framework is then used to develop a Bayesian approach to evaluation of the main effects and interaction terms. The paper defines a Gaussian process metamodel as the prior and obtains an analytical expression for the posterior distribution of the metamodel conditional on the data. The posterior distribution is then used to infer means and variances of the various conditional expectations used in the determination of Sobol indices. It should be noted that this approach appears to determine the sensitivity of the metamodel to its inputs, and so the quality of the results will depend on the goodness of fit of the metamodel, but the authors supply references for further papers that describe how to develop a suitable metamodel.

Several different approaches to constructing non-parametric metamodels for sensitivity analysis are described and compared in a later paper [22]. Classes of metamodel considered include multivariate adaptive regression splines (MARS), random forests, gradient boosting machines, and adaptive component selection and smoothing operators (ACOSSO) as well as Gaussian process models. Random forests and gradient boosting machines are both machine learning models initially developed for classification problems. The methods are applied to various standard test functions using different sample sizes to construct the metamodels. One of the most useful remarks in the paper is that the construction of a metamodel for a complex model with many inputs requires some form of variable selection (which is itself one of the possible aims of a sensitivity analysis). Methods such as MARS which involve variable selection in the metamodel development procedure result in a more computationally efficient metamodel, as well as identifying insignificant variables directly.

3.5 Does sensitivity analysis need to consider input quantity distributions?

It is interesting to consider at this point whether sensitivity analysis should consider the distributions associated with the input quantities, or should sensitivity be regarded as a property of a model alone. The GUM [1] definition of sensitivity coefficients does not require any knowledge of the distributions associated with the model inputs, but as has been noted above this approach assumes that linearisation of the underlying model is valid and only considers local measures of sensitivity. However, also as previously noted, it can be more useful to consider the product of the absolute value of the coefficient with the standard uncertainty of the input quantity, which does consider the input distribution. An uncertainty budget table will often report such values.

It is also worth noting that if indices of other sensitivity measures are computed numerically by sampling randomly within a finite region of the space of input quantities, then the imposed bounds on the finite region are in some ways equivalent to assuming that the inputs are independently uniformly distributed within these bounds, so a distribution is implicitly being assigned to the inputs even if it is not being explicitly acknowledged.

The importance of the distributions associated with the inputs depends in part on what question the sensitivity analysis is intended to answer. One aspect of this dependence relates to whether the question is considering a specific instance of the model or a more general assessment of model behaviour. Many models are developed such that a wide range of values of their input parameters are valid, and there may be benefits in considering the full range of values that could be input into the model by imposing limits on the inputs and a uniform distribution. If a specific case is of interest then it is more likely to be useful to consider the distributions associated with the inputs in more detail.

As an example, consider a model of a tensile test that outputs the required force to produce a specified stress in the sample under test given the Young's modulus of the sample and a set of other experimental parameters. The model may accept values that are valid for any material from nylon (approximately 3 GPa) to tungsten carbide (approximately 600 GPa). If the user wishes to understand how close to linear their model is for any material that they may want to simulate, the user is essentially accepting all materials, and hence all modulus values within that range, as equally likely and hence is implicitly assuming a uniform distribution. On the other hand, if the user knows that they will be testing samples of a batch of steel whose Young's modulus has a well characterised distribution and wants to know which of the experimental parameters should be most closely controlled during the test, using the distribution associated with the Young's modulus will give more useful information about the response of the system during the experiment.

4 USEFUL LITERATURE

Saltelli et al [2] provides a useful introduction to the uses of sensitivity analysis, including a clear statement of terminology and an explanation of the limited usefulness of one at a time designs. The paper also highlights six recommendations for good practice, which (summarised and reordered) are:

- Sensitivity analysis and uncertainty analysis should be focussed on a question.
- With some exceptions, it is advisable to perform both uncertainty and sensitivity analysis.
- Both uncertainty and sensitivity analysis should be based on a global exploration of the space of input factors.
- When sensitivity analysis is performed, it should allow the relative importance of input factors and combinations of factors to be assessed either qualitatively or quantitatively.
- Sensitivity and uncertainty analysis are themselves uncertain and it should be clear how the uncertainties associated with the input factors and the parameter values (e.g. sample size) used in any analysis method were chosen.
- Even an apparently perfect uncertainty and sensitivity analysis is no assurance against error.

All of these recommendations are relevant to metrology. The first two points are of most importance when planning a sensitivity or uncertainty analysis as they affect what information will be needed for the analysis and potentially how the model to be analysed is constructed. The second two points both affect the practical implementation of the analysis, and the third two points affect the presentation and interpretation of the results of the analysis.

Sensitivity Analysis by Saltelli et al [3] is a good general introduction to the field of sensitivity analysis. It contains sections on most of the most commonly used methods and includes an extensive range of application examples. Since the version referred to was written in 2000, it does not include information about density-based methods such as PAWN [16] which were developed after that date.

Sensitivity analysis: a review of recent advances by Borgonovo and Plischke [6] provides a good overview of the subject that does include more recent developments, and in particular density-based methods. It uses a simple analytical model with four parameters to illustrate and compare methods throughout. There are several points raised in the paper that are useful to discuss in more detail.

The first is that the decomposition of variance on which several of the variance-based methods depend becomes non-unique if the input quantities are correlated. The consequence of this non-uniqueness is that variance decomposition is not a suitable tool for sensitivity analysis when input quantities are correlated. Whether this limits the applicability of the technique partly depends on whether you consider sensitivity analysis to require consideration of the distributions of the inputs, or whether you consider sensitivity to be a property of a function output (perhaps limited to a given region of the space of input quantities).

The second is that section 6 of the paper offers several classes of question that the sensitivity analysis may be aiming to answer, which will affect which method should be used in the analysis. The classes of question mentioned are:

- Factor prioritization
  - Identify which inputs are most important
  - Use global sensitivity methods
- Factor fixing
  - Identify which factors matter least and fix their values
  - Use screening methods
- Model structure
  - Determine whether interactions between factors are significant or whether the model is essentially linear
  - Use high order variance-based methods
- Direction of change
  - Determine whether a given input will increase or decrease the output
  - Use signs of local derivatives or first order terms in ANOVA
- Stability
  - Identify a region of the input space within which the model output does not change much (useful for robust design)
  - Use specialised methods not discussed in this report

The third point of interest is that section 5 of [6] provides an approach to place all global sensitivity measures as the expected value of a statistic. Let $\mathbb{P}$ be a probability distribution and ζ be an operator on probability distributions such that $\zeta(\mathbb{P},\mathbb{P}) = 0$. Then ξ_i, a global sensitivity measure for the input quantity X_i, can be written as

$$\gamma_i(x_i) = \zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_i=x_i}), \tag{30}$$

$$\xi_i = \mathbb{E}[\gamma_i(X_i)], \tag{31}$$

where $\gamma_i(x_i)$ is an inner statistic that gives regional sensitivity information and the outer expectation gives global behaviour of that statistic. For Sobol indices, for instance, the inner statistic is $\mathbb{V}(Y)^{-1}(\mathbb{E}(Y|X_i) - \mathbb{E}(Y))^2$.

This approach leads to a way to derive the properties of a sensitivity measure (such as transform invariance) from its inner statistic, and reframes global sensitivity analysis as quantifying of measures of statistical dependence. This reframing may lead to new methods for sensitivity analysis based on innovations in the field of statistical dependence.

Most papers that discuss specific methods use examples to illustrate these methods. There are several standard functions that are used in many of these papers, which are sometimes

complemented by a more complicated model with an application of particular interest to the author. Two of the most commonly used functions are the Ishigami-Homma function:

$$Y = \sin(X_1) + a\sin^2(X_2) + bX_3^4\sin(X_1) \tag{32}$$

where $x_i \in [-\pi, \pi]$, $a = 2$, $b = 1$, and Sobol's G function

$$Y = \prod_{i=1}^{10} \frac{|4X_i - 2| + a_i}{1 + a_i} \tag{33}$$

where $X_i \in [0,1]$ and $\mathbf{a} = [0,1,2,4,8,99,99,99,99,99]$. Some authors also add a random zero-mean error to the output of these functions.

Another interesting approach is outlined in [9] where a set of nine simple univariate basis functions with different properties (e.g. linear, periodic, discontinuity, etc.) are combined in products and weighted sums to give a framework from which multivariate test functions can be created. These functions can be designed to have particularly challenging features if a particular analysis method may not perform well for functions with a particular feature, or it can be used to generate a suite of random test functions to explore the performance of a method more broadly.

5 ILLUSTRATIVE EXAMPLES

In this section we use two commonly used standard functions to compare the effect of increasing sample size on method convergence for a variance based method (Sobol) and a density based method (PAWN). All sampling, function and sensitivity analysis methods were implemented in Python 3.9 using the sensitivity analysis library package SALib, version 1.4.7 [23, 24]. All random number generator options were fixed to ensure reproducibility. Sections 5.1 to 5.3 describe the methods used and Section 5.4 presents the results and findings.

5.1 SAMPLING

5.1.1 Latin hypercube sampling

Latin hypercube sampling (LHS) generates quasi-random samples of multiple variables. For full details the reader is referred to [25, 26].

5.1.2 Sobol' sequence sampling

Saltelli's extension of the Sobol' sequence is a popular quasi-random sequence used to generate uniform samples of multiple variables. Saltelli's scheme extends the Sobol' sequence in a way to reduce error rates in the resulting sensitivity index calculations. For full details the reader is referred to [11, 27, 28].

5.2 FUNCTIONS

Two commonly used standard functions described in Section 4 are used, and are restated here for convenience. These are the Ishigami-Homma function:

$$Y = \sin(X_1) + a\sin^2(X_2) + bX_3^4\sin(X_1)$$

where $x_i \in [-\pi, \pi]$, $a = 2$, $b = 1$, and Sobol's eight parameter G function

$$Y = \prod_{i=1}^{8} \frac{|4X_i - 2| + a_i}{1 + a_i}$$

where $X_i \in [0,1]$ and $\mathbf{a} = [0, 1, 4.5, 9, 99, 99, 99, 99]$.

5.3 SENSITIVITY ANALYSIS

5.3.1 Sobol sensitivity indices

The Sobol indices were calculated 100 times to generate statistics on variability. The plots in section 5.4.1 show confidence intervals for a level of confidence of 0.95. For full details on the Sobol indices, see Section 3.2.

5.3.2 PAWN sensitivity indices

PAWN sensitivity analysis is a moment-independent approach to Global Sensitivity Analysis (GSA). For full details See Section 3.3. The approach used for index calculation is that described in [17], where calculations use sub-intervals of the input(s) of interest for conditioning rather than that in [16] which uses single values of the input(s) and is more restrictive regarding the use of historical data.

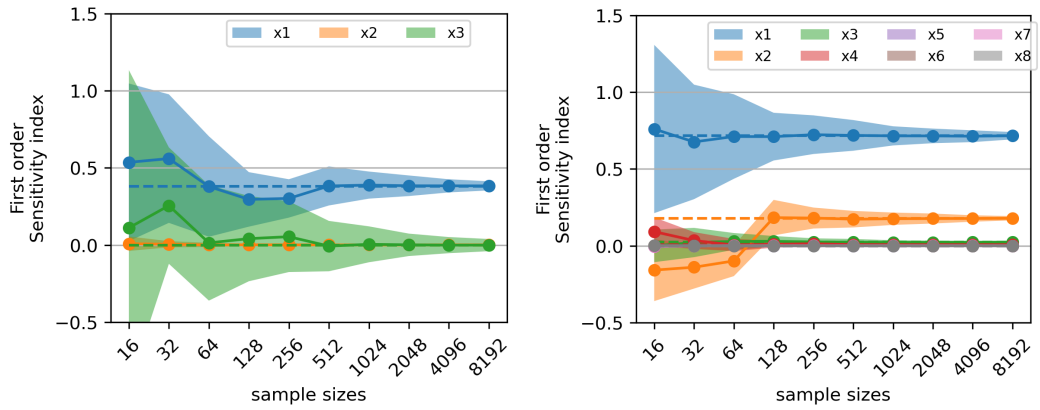
5.4 Results and interpretation

The results will be presented in two parts: Section 5.4.1 presents Sobol sensitivity analysis on the standard functions, and Section 5.4.2 presents PAWN sensitivity analysis on the same functions. In all cases, the dashed lines on the chart represent the true values of the indices given in the literature.

5.4.1 Sobol sensitivity analysis

Convergence of the first order Sobol indices for both the Ishigami-Homma function and eight parameter Sobol G-function for increasing sample sizes can be seen in Figure 5.

As expected, these both converge to the known analytical solutions for the Sobol sensitivity indices for these functions.



(a): Ishigami-Homma function. **(b):** Sobol G-function.

Figure 5: Convergence of first order Sobol sensitivity indices for increasing sample sizes with 95% confidence interval. Shown for 5(a) the Ishigami-Homma function input parameters x_{1-3} and 5(b) the Sobol G-function input parameters x_{1-8}, analytical solutions indicated by dashed line.

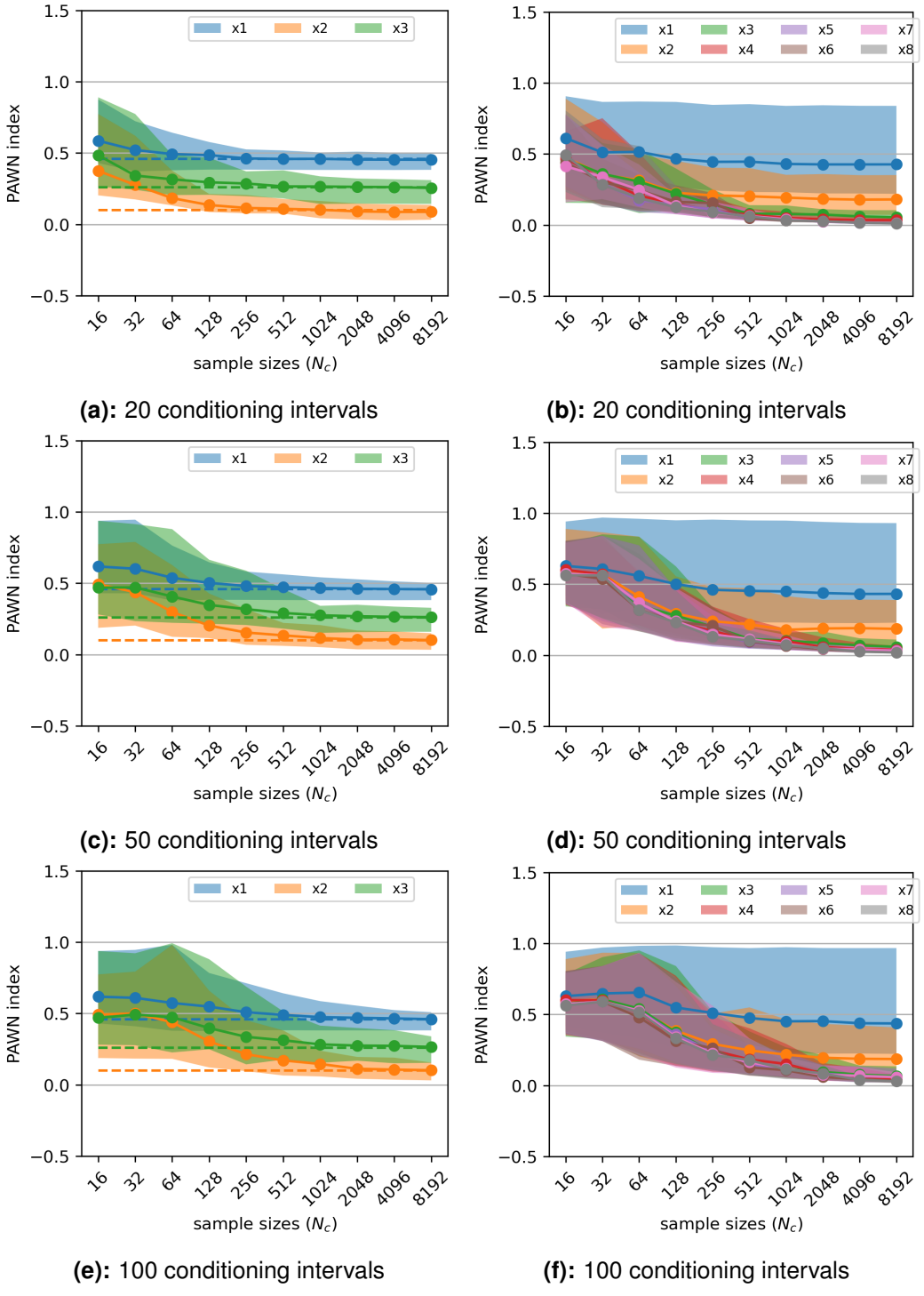
5.4.2 PAWN sensitivity analysis

Convergence of the median PAWN indices for both the Ishigami-Homma function and eight parameter Sobol G-function for increasing sample sizes and number of conditioning intervals can be seen in Figures 6 and 7.

The bands around the median values show the index calculated using a maximum statistic and the index calculated using a minimum statistic rather than the maximum and minimum value of the median across repeated samples, and so the width of the bands for a large sample size is not indicative of a lack of convergence.

As was described in section 3.3, the PAWN index is based on the Kolmogorov-Smirnoff (K-S) measure. This measure is the largest difference between the cumulative probabilities for a given value of Y, the model output, and for two cumulative distribution functions (cdfs) of Y. A sketch of this measure is shown in figure 8. For the PAWN index, the two distributions used are the total distribution, and a distribution conditioned on a specific interval of one or more of the variables.

The PAWN index when implemented:



(a): 20 conditioning intervals

(b): 20 conditioning intervals

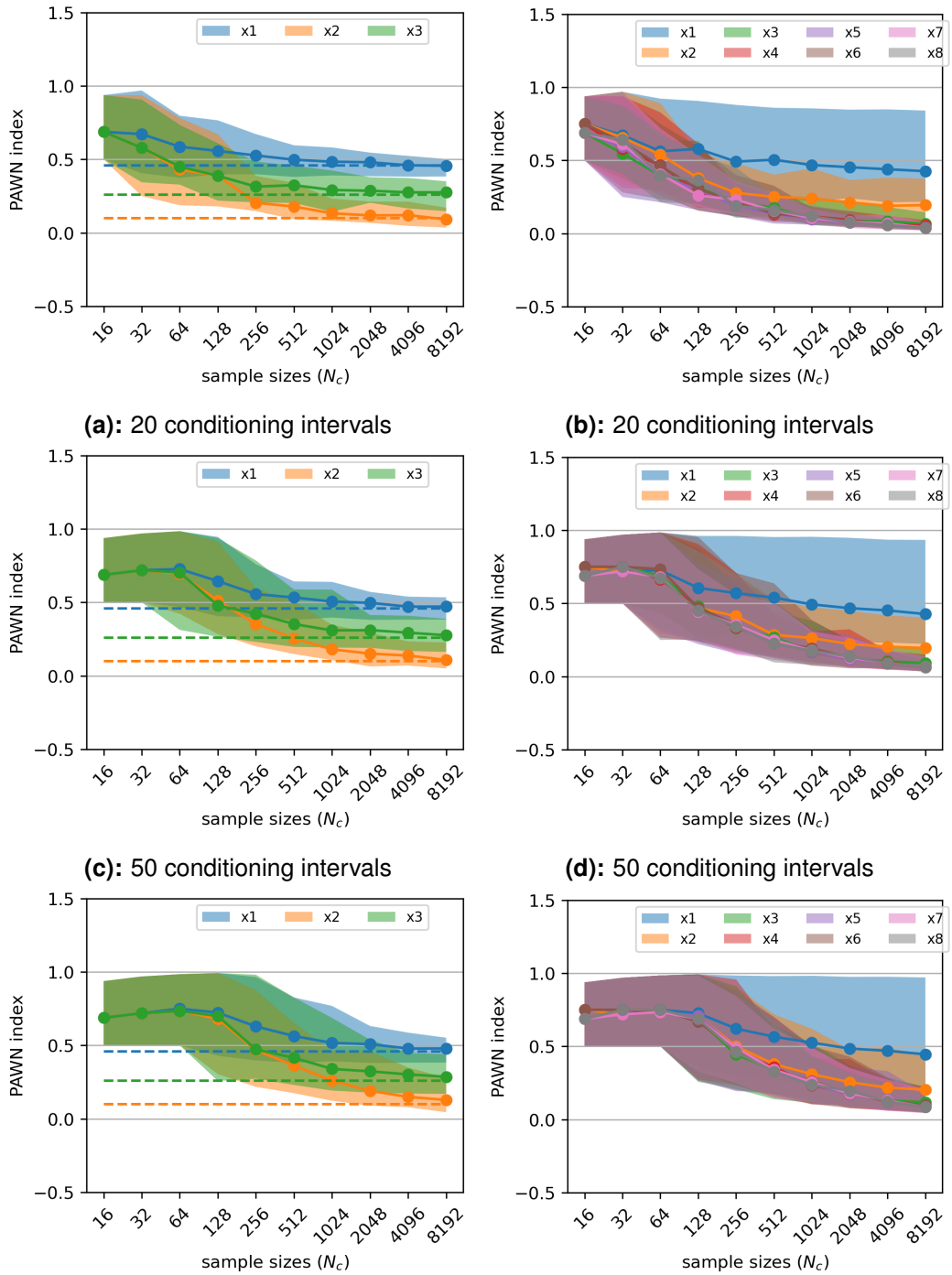
(c): 50 conditioning intervals

(d): 50 conditioning intervals

(e): 100 conditioning intervals

(f): 100 conditioning intervals

Figure 6: Convergence of median PAWN indices and for increasing sample sizes, shaded band indicates minimum and maximum values. Shown for (left) the Ishigami-Homma function and (right) the Sobol G function with all input samples generated using Sobol sampling, analytical solutions indicated by dashed line. Number of PAWN conditioning intervals indicated in subcaptions.



(a): 20 conditioning intervals

(b): 20 conditioning intervals

(c): 50 conditioning intervals

(d): 50 conditioning intervals

(e): 100 conditioning intervals

(f): 100 conditioning intervals

Figure 7: Convergence of median PAWN indices and for increasing sample sizes, shaded band indicates minimum and maximum values over the number of conditioning intervals. Shown for (left) the Ishigami-Homma function and (right) the Sobol G function with all input samples generated using Latin hypercube sampling. Solution generated using a custom PAWN Matlab code indicated by dashed lines where available. Number of PAWN conditioning intervals indicated in subcaptions.

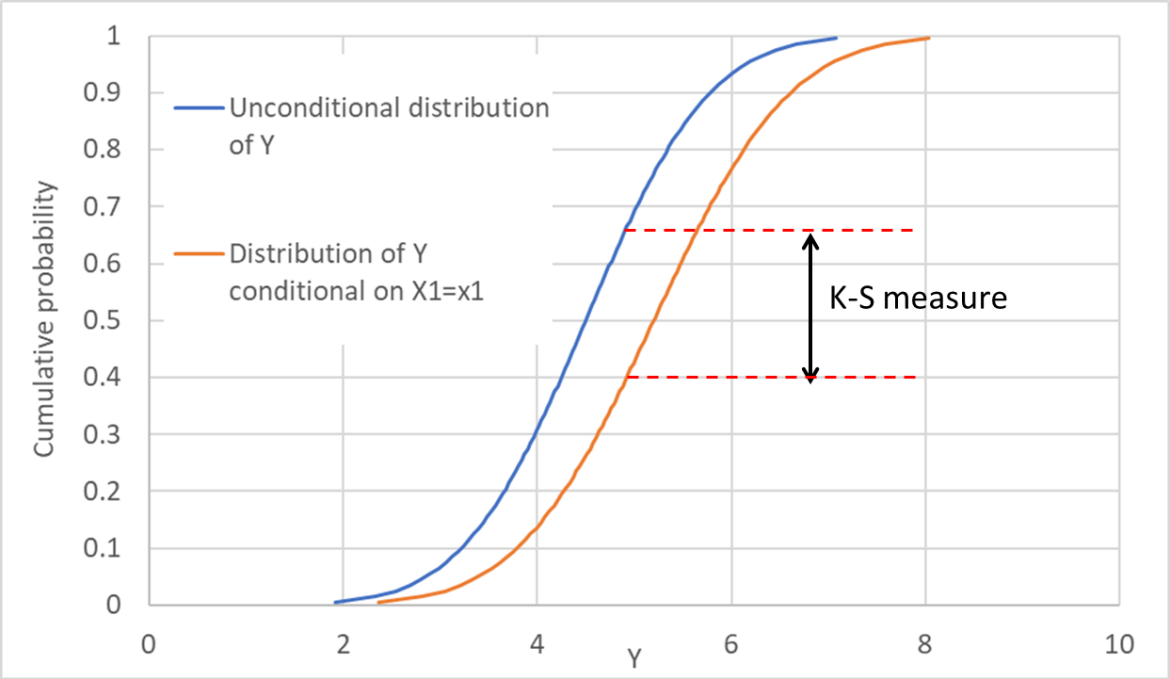


Figure 8: Sketch of the Kolmogorov-Smirnoff measure. The value shown is the largest difference in cumulative probability for any value of Y.

- Uses discrete samples of size N_c to estimate the cdf.
- Repeats the calculation of a KS value for n different conditioning intervals of the variable.
- Allows the user to select what statistic should be calculated from the n different KS values.

Commonly used statistics are the maximum, minimum, mean, median and coefficient of variation of the K-S values. Both N_c and n will have an effect on the value of the index. As N_c increases the shapes of the cdfs will converge and so the values of the K-S measure, and hence all of the statistics, will converge. As n increases, there will be more values to be a maximum or a minimum and more values in the calculation of a mean and a coefficient of variation. The maximum will increase as n increases but its rate of increase will decrease as n increases. Similarly the minimum will decrease and its rate of decrease will decrease as n increases.

This qualitative behaviour is shown in figures 6 and 7. It is also interesting to note that comparisons between these two figures suggest that for the same function and the same number of conditioning intervals, the maximum and minimum index values converge more quickly using Latin hypercube sampling than they do with Sobol sampling.

It is important to note that the maximum and minimum do not provide a confidence interval for the PAWN index: they are themselves different versions of the PAWN index.

6 STEPS IN SENSITIVITY ANALYSIS AND A SUMMARY OF GOOD PRACTICE

The descriptions of sensitivity analysis methods and the literature summary above have given a broad description of how to carry out a sensitivity analysis and identified some points of good practice. This section will present that information in a more structured format.

It is assumed in this section that the inputs and outputs are well defined and that a model linking the two already exists. It is further assumed that the model is sufficiently complicated that analytical approaches for evaluation of expectations and variances are not usable.

It was noted by [2] that a sensitivity analysis should start from a question. The nature of the question will affect which method is most suitable for the analysis. In particular, focussing on tasks common in metrology:

- If the aim is to identify which inputs are most important (and hence may benefit from having their associated uncertainties reduced), a global sensitivity method should be used.
- If the aim is to identify which inputs are least important (and hence fix their value and omit them from any subsequent uncertainty evaluation), a screening method should be used.
- If the aim is to learn more about the model structure and determine whether the model is approximately linear, high order variance-based methods should be used (for instance comparison of Sobol main effects indices and total effects indices for each variable).

It is worth noting that more than one question may be asked sequentially: an initial screening method may be used to reduce the problem size before carrying out a global sensitivity analysis on the remaining inputs.

The link between the question being asked and the distributions associated with the inputs was discussed in section 3.5 with the emphasis on how the question imposes requirements on the distributions. It is also the case that the information available about the distributions associated with the inputs affects the question that can be answered and therefore the choice of method. This dependency does not necessarily mean that a method should not be used, but it does mean that the results of the analysis should be interpreted bearing the imposed distributions in mind.

This point is one facet of a wider aspect of good practice: most methods for sensitivity analysis make assumptions about either the model or the input quantities, and the results of the sensitivity analysis should be interpreted bearing those assumptions in mind. If assumptions are made about the model, these should be checked. For instance if a DoE and ANOVA approach is used with two levels then the results are more likely to be accurate if the model is close to linear in each input quantity. This requirement could be checked by fitting a linear model to the DoE data, evaluating the model at an additional point using the mid-point of the ranges of each variable, and comparing the value of the linear fit at the new point and the value output by the full model.

Many models in use in metrology have complicated multi-step algorithms linking their inputs and their outputs, and are therefore effectively "black boxes". In many cases the models can

take minutes to evaluate. The "black box" nature of the models means it is necessary to use numerical approximation techniques to evaluate many of the sensitivity indices discussed above. The computational expense of the model may mean that the number of model evaluations that can be carried out practically within a sensitivity analysis may be limited. In such cases, it is better to use a method that imposes assumptions on the model (for instance by using a metamodel) and to interpret the results in the light of the assumptions rather than using too few model evaluations to obtain a converged value of (for instance) a Sobol index.

If the sensitivity method being used requires numerical approximation methods to evaluate an integral or to approximate a distribution function, it is essential to check for convergence. This check can be made visually by plotting the calculated index value against the number of evaluations used in its calculation. Some methods (for instance the general PAWN method described in [17]) may have a dependence on other parameters, and the robustness of the calculated value to these parameters should also be examined.

Another way to gain an understanding of the behaviour and reliability of an index is to include a dummy input. A dummy input is an additional input variable that does not affect the output of the model, so that the sensitivity of the model result to the input should be equal to zero. Including such a parameter in an analysis can give an indication of what the likely error in calculated values is.

Finally, as with all data analysis procedures, the results should be critically examined rather than accepted without further consideration. It is rare for the user of a model to have absolutely no idea of how the input values affect the model results, and checking that the sensitivity analysis results are consistent with the user's expert knowledge is a sensible and simple step.

7 FUTURE WORK

In this report we examined the impact of sample size on the convergence of two sensitivity analysis methods, Sobol and PAWN. These were applied to two well known test functions with well defined sensitivity indices that can be analytically derived. Ultimately, sensitivity analysis is applied to a range of different functions and models, and the impact of factors other than sample size on the sensitivity indices for these models is important.

Areas for future work include:

- How does varying input parameter ranges affect sensitivity indices?
- When constraints are applied to input parameters (as is sometimes the case in real world models), how are sensitivity indices impacted and can density-based methods provide an advantage over variance-based methods where there is an assumption of no correlation between the variables?
- How does the number of input parameters affect the analysis?
- How does altering the ordering of input parameters affect Sobol sensitivity analysis and can screening methods be used to mitigate for this?

ACKNOWLEDGEMENTS

This work was supported by the UK Government's Department for Business, Energy and Industrial Strategy (BEIS).

REFERENCES

[1] BIPM et al. *Evaluation of measurement data — Guide to the expression of uncertainty in measurement*. Joint Committee for Guides in Metrology, JCGM 100:2008. URL: https://www.bipm.org/documents/20126/2071204/JCGM%5C_100%5C_2008%5C_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6.

[2] A. Saltelli et al. "Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices". In: *Environmental Modelling and Software* 114 (2019), pp. 29–39.

[3] A. Saltelli, K. Chan, and E. M. Scott. *Sensitivity Analysis*. John Wiley & Sons, 2001.

[4] MedalCare. *MedalCare project website*. 2019. URL: https://www.ptb.de/empir2019/medalcare/home/.

[5] I. X. Partarrieu, N. Smith, and P. Harris. *What Am I Measuring? Using Experimental Design to Understand and Improve Measurement Pipelines, with an Example Application to Radiomics Measurements.* NPL Report. MS 35. 2022.

[6] E. Borgonovo and E. Plischke. "Sensitivity analysis: A review of recent advances". In: *European Journal of Operational Research* 248(3) (2016), pp. 869–887.

[7] NIST. *NIST Engineering Statistics Handbook*. 2012. URL: https://www.itl.nist.gov/div898/handbook/pri/pri.htm.

[8] M. D. Morris. "Factorial sampling plans for preliminary computational experiments". In: *Technometrics* 33(2) (1991), pp. 161–174.

[9] W. Becker. "Metafunctions for benchmarking in sensitivity analysis". In: *Reliability Engineering and System Safety* 204 (2020), p. 107189.

[10] I. M. Sobol. "Sensitivity analysis for non linear mathematical models." In: *Mathematical Modeling & Computational Experiment* 1 (1993), pp. 407–414.

[11] A. Saltelli et al. "Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index". In: *Computer Physics Communications* 181.2 (2010), pp. 259–270. ISSN: 00104655. URL: http://dx.doi.org/10.1016/j.cpc.2009.09.018.

[12] A. Puy, S. Lo Piano, and A. Saltelli. "A sensitivity analysis of the PAWN sensitivity index". In: *Environmental Modelling and Software* 127 (2020), p. 104679.

[13] E. Borgonovo. "A new uncertainty importance measure". In: *Reliability Engineering and System Safety* 92(6) (2007), pp. 771–784.

[14] H. Liu, W. Chen, and A. Sudjianto. "Relative Entropy Based Method for Probabilistic Sensitivity Analysis in Engineering Design". In: *Journal of Mechanical Design* 128(2) (2006), pp. 326–336.

[15] P. Wei, L. Zhenzxhou, and X. Yuan. "Monte Carlo simulation for moment-independent sensitivity analysis". In: *Reliability Engineering and System Safety* 110 (2013).

[16] Francesca Pianosi and Thorsten Wagener. "A simple and efficient method for global sensitivity analysis based on cumulative distribution functions". In: *Environmental Modelling and Software* 67 (2015), pp. 1–11. ISSN: 13648152.

[17] Francesca Pianosi and Thorsten Wagener. “Distribution-based sensitivity analysis from a generic input-output sample”. In: *Environmental Modelling and Software* 108.August 2018 (2018), pp. 197–207. ISSN: 13648152.

[18] SAFE Toolbox. *Review of manuscript A sensitivity analysis of the PAWN sensitivity index*. 2022. URL: https://www.safetoolbox.info/wp-content/uploads/2022/02/Review-of-manuscript.pdf.

[19] G. Baroni and T. Francke. “An effective strategy for combining variance- and distribution-based global sensitivity analysis”. In: *Environmental Modelling and Software* 134 (2020), p. 104851.

[20] Rasmussen K. et al. *Best practice guide to uncertainty evaluation for computationally expensive models*. 2015. URL: https://www.ptb.de/emrp/fileadmin/documents/nmasatue/NEW04/Papers/BPG-WP2-NEW04.pdf.

[21] J. Oakley and A. OHagan. “Probabilistic sensitivity analysis of complex models: A Bayesian approach”. In: *Journal of the Royal Statistical Society B* 66(3) (2004), pp. 751–769.

[22] C. Storlie et al. “Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models”. In: *Reliability Engineering and System Safety* 94(11) (2009), pp. 1735–1763.

[23] Jon Herman and Will Usher. “SALib: An open-source Python library for Sensitivity Analysis”. In: *The Journal of Open Source Software* 2.9 (Jan. 2017). URL: https://doi.org/10.21105/joss.00097.

[24] Takuya Iwanaga, William Usher, and Jonathan Herman. “Toward SALib 2.0: Advancing the accessibility and interpretability of global sensitivity analyses”. In: *Socio-Environmental Systems Modelling* 4 (May 2022), p. 18155. URL: https://sesmo.org/article/view/18155.

[25] M. D. McKay, R. J. Beckman, and W. J. Conover. “A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code”. In: *Technometrics* 21 (2 May 1979), p. 239. ISSN: 00401706.

[26] Ronald L. Iman, Jon C. Helton, and James E. Campbell. “An Approach to Sensitivity Analysis of Computer Models: Part IIntroduction, Input Variable Selection and Preliminary Variable Assessment”. In: *https://doi.org/10.1080/00224065.1981.11978748* 13 (3 July 2018), pp. 174–183. ISSN: 0022-4065. URL: https://www.tandfonline.com/doi/abs/10.1080/00224065.1981.11978748.

[27] Andrea Saltelli. “Making best use of model evaluations to compute sensitivity indices”. In: *Computer Physics Communications* 145 (2 May 2002), pp. 280–297. ISSN: 0010-4655.

[28] I. M. Sobol. “Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates”. In: *Mathematics and Computers in Simulation* 55 (1-3 Feb. 2001), pp. 271–280. ISSN: 0378-4754.