Global Sensitivity Analysis for the Interpretation of Machine Learning Algorithms

Sonja Kuhnt and Arkadius Kalka

Abstract Global sensitivity analysis aims to quantify the importance of model input variables for a model response. We highlight the role sensitivity analysis can play in interpretable machine learning and provide a short survey on sensitivity analysis with a focus on global variance-based sensitivity measures like Sobol' indices and Shapley values. We discuss the Monte Carlo estimation of various Sobol' indices as well as their graphical presentation in the so-called FANOVA graphs. Global sensitivity analysis is applied to an analytical example, a Kriging model of a piston simulator and a neural net model of the resistance of yacht hulls.

Keywords Interpretable machine learning · Global sensitivity analysis · Sobol' indices · Shapley values · FANOVA graph · Kriging

1 Introduction

Machine learning is a set of methods that improve automatically through experience, i.e. it is based on data. Popular machine learning methods are, e.g. support vector machines (SVMs [2]), artificial neural networks (ANNs) and random forests (RFs). Machine learning algorithms are increasingly applied in science and business and have achieved impressive performances in diverse tasks, outperforming humans. However, for several machine learning algorithms it is hard to tell what the machine has actually learned from the data. For example, in the case of ANNs, what was learned is hidden in the weights and biases of the neurons involved. If a machine learning model performs well, one might simply trust the model and ignore why it made a certain decision. However, such an attitude goes against human curiosity and thirst for knowledge. This raises the issue of interpretability [24, 25]. The straightforward way to achieve interpretability in statistical learning is to use only

S. Kuhnt (✉) · A. Kalka
University of Applied Sciences and Arts, Dortmund, Germany
e-mail: sonja.kuhnt@fh-dortmund.de; arkadius.kalka@fh-dortmund.de

A. Steland, K.-L. Tsui (eds.), *Artificial Intelligence, Big Data and Data Science in Statistics*, https://doi.org/10.1007/978-3-031-07155-3_6

interpretable models. Interpretable models are, e.g. linear models, generalized linear models [23], generalized additive models [12], decision trees and rules. On the other hand, model-agnostic interpretation methods are more flexible and can be applied to any machine learning algorithm. Graphical model-agnostic methods are, e.g. Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots. We suggest to provide model-agnostic methods to evaluate the influence of different regressors and their interactions by applying methods from the statistical field of global sensitivity analysis (GSA). Sensitivity analysis is the study of how the uncertainty in the output of a mathematical model or function can be divided and allocated to different sources of uncertainty in its inputs [14, 29]. Given any real-valued function on several variables –whether analytical or given by a black-box– one wants to know which input variables affect the variability of the function the most. GSA has proven to be a valuable tool in analysing expensive to evaluate computer models with a surrogate model, e.g. a Kriging model, build first. Cheng et al. [3] use support vector regression as surrogate model within GSA, whereas [37] built a new feature selection approach upon GSA. Like in [4] we suggest to achieve an understanding and interpretability of, e.g. ANNs, SVMs and RFs by combining GSA and visualization.

2 Global Sensitivity Analysis

This section reviews sensitivity analysis with a focus on global variance-based sensitivity measures, but we also discuss derivative-based global sensitivity measures briefly.

2.1 Global Sensitivity Indices

Consider a function $f : \Delta \subseteq \mathbb{R}^d \to \mathbb{R}$ that is square integrable w.r.t. a d-dimensional product measure μ. The functional analysis of variance (FANOVA) decomposition (also called Hoeffding-Sobol' decomposition) of $f \in L^2(\mu)$ is the unique decomposition

$$f(X) = f_0 + \sum_i f_i(X_i) + \sum_{i<j} f_{i,j}(X_i, X_j) + \cdots + f_{1,\ldots,d}(X_1, \ldots, X_d) \qquad (1)$$

such that $E(f_I(X_I) \mid X_J) = 0$ for all $J \subset I \subseteq [d] := \{1, \ldots, d\}$. In particular, we have $E(f_I(X_I)) = 0$ for all $\emptyset \neq I \subseteq [d]$, see e.g. [6]. Furthermore, this implies orthogonality of all summands in the decomposition, i.e. $E\left(f_I(X_I) f_J(X_J)\right) = 0$ for all $I \neq J \subseteq [d]$. The FANOVA decomposition can be computed recursively by

$$f_0 = E(f(X)), \text{ and } f_I(X_I) = E(f(X) \mid X_I) - \sum_{J \subset I} f_J(X_J). \tag{2}$$

Orthogonality allows for an ANOVA like decomposition of the variance of f:

$$D = Var(f(X)) = \sum_{I \subseteq [d]} Var(f_I(X_I)). \tag{3}$$

The variance of each term, $D_I = Var(f_I(X_I))$, gives a sensitivity index of its effect. The standardized indices

$$S_I = D_I / D \tag{4}$$

are known as *Sobol' indices* [31]. Especially, Sobol' indices $S_i = S_{\{i\}}$ for individual variables are referred to as first-order indices and $S_{ij} = S_{\{i,j\}}$ as second-order indices. The same holds for the unstandardized versions D_i and D_{ij}. Sobol' introduced the *closed sensitivity index* to describe the influence of a group of variables:

$$D_I^{cl} = \sum_{J \subseteq I} D_J = Var(E(f(X)|X_I)). \tag{5}$$

The *total sensitivity index* by Homma and Saltelli [13] describes the total contribution of a set of variables including all interactions of any order and is defined by all partial variances containing at least one of the variables, i.e.

$$D_I^T = \sum_{I \cap J \neq \emptyset} D_J, \quad S_I^T = \frac{D_I^T}{D}. \tag{6}$$

For $I = \{i\}$, this total sensitivity index is defined by considering all supersets. An extension [21] of the concept of *superset importance* is given by

$$D_I^{sup} = \sum_{J \supseteq I} D_J. \tag{7}$$

In particular, the unnormalized and normalized *total interaction* indices (TIIs) [9] are given by

$$D_{i,j}^{sup} = \sum_{J \supseteq \{i,j\}} D_J \quad \text{and} \quad S_{i,j}^{sup} = \sum_{J \supseteq \{i,j\}} \frac{D_J}{D}. \tag{8}$$

So, each of these indices characterizes a different aspect of the sensitivity of the model response to individual input variables or interactions between them.

2.2 Shapley Values

A similar problem to the FANOVA decomposition has been studied in game theory and economics, namely the problem of attributing the value created in a team effort to individual team members. Consider the setting where one can measure the value $val(I) \in \mathbb{R}$ created by any subset $I \subseteq [d]$ of the d-member team. In that case the so-called *Shapley values* ϕ_i are the unique choice that satisfy the following four natural criteria [30, 36].

1. (Efficiency) $\sum_i^d \phi_i = val([d])$.
2. (Symmetry) $val(I \cup i) = val(I \cup j) \quad \forall I \subseteq [d] \setminus \{i, j\}$ implies $\phi_i = \phi_j$.
3. (Dummy) $val(I \cup i) = val(I) \quad \forall I \subseteq [d]$ implies $\phi_i = 0$.
4. (Additivity) The game with value $val^{(1)} + val^{(2)}$ has Shapley values $\phi^{(1)} + \phi^{(2)}$ with $\phi^{(1)} = \phi(val^{(1)})$ and $\phi^{(2)} = \phi(val^{(2)})$.

Then the Shapley value of an individual variable is given by

$$\phi_i = \frac{1}{d} \sum_{I \subseteq [d] \setminus \{i\}} \binom{d-1}{|I|}^{-1} (val(I \cup i) - val(I)). \tag{9}$$

Shapley values are connected to the FANOVA decomposition by Owen [27]. In that context, for any subset I of input variables, their combined value $val(I)$ is the "variance explained" in the FANOVA decomposition. More precisely, the choice in [27] is $val(I) = D_I^{cl}$. Then, using the properties (1) − (4), it can be shown that the Shapley value is

$$\phi_i = \sum_{I \subseteq [d], i \in I} \frac{D_I}{|I|} \tag{10}$$

according to Theorem 1 in [27]. The Shapley value does not coincide with any first-order Sobol' index, but it is bracketed between the closed and total sensitivity index [27]:

$$D_i^{cl} \leq \phi_i \leq D_i^T. \tag{11}$$

A normalized Shapley value may be defined as $\phi_i^* = \phi_i / D$. Because these indices are comparatively easy to compute, Sobol' indices provide effectively computable bounds for the Shapley value. An exact computation of the Shapley value is computationally expensive because there are 2^d subsets of $[d]$, representing coalitions of variables. Štrumbelj and Kononenko [34] and Song et al. [33] propose effective algorithms to estimate Shapley values using Monte Carlo sampling.

2.3 Derivative-Based Global Sensitivity Measures

Based on the work of [32] derivative-based global sensitivity measures (DGSM) were introduced by Kucherneko et al. [20] as

$$\nu_i = \int \left(\frac{\partial f}{\partial x_i}(x) \right)^2 d\mu. \tag{12}$$

A normalized DGSM can be defined by $\nu_i^* = \nu_i / \sum_j^d \nu_j$. DGSMs are not associated with a functional decomposition, but they are connected to total sensitivity indices by the inequality $D_i^T \leq C(\mu_i)\nu_i$ if for the measure μ the Poincare inequality

$$\int g(x)^2 \mu \leq C(\mu) \int ||\nabla g(x)||^2 d\mu \tag{13}$$

holds for all centred functions $g \in L^2(\mu)$ with $\int g(x)d\mu = 0$ and $||\nabla g|| \in L^2(\mu)$. Friedman and Popescu [7] introduced crossed DSGMs, in particular, for interactions:

$$\nu_{i,j} = \int \left(\frac{\partial^2 f}{\partial x_i \partial_j}(x) \right)^2 d\mu. \tag{14}$$

Roustant et al. [28] provide an inequality to link crossed DGSMs to superset importance.

2.4 Estimation of Indices

For analytically tractable test functions, the indices above may be calculated by evaluating the integrals involved. In general, the function f is not known analytically and will be treated as black-box function. In *Monte Carlo estimation*, we take a high number of n samples $x^{(1)}, \ldots, x^{(n)}$ from the distribution μ and approximate the integral by

$$\frac{1}{n} \sum_{k=1}^{n} f(x^{(k)}) \quad \overset{n \to \infty}{\longrightarrow} \int f(x) d\mu = E(f(X)). \tag{15}$$

The approximation is unbiased and convergent with probability one according to the law of large numbers. For the estimation, we require a representation of the sensitivity indices that is suitable for Monte Carlo integration. A popular choice is based on the pick-and-freeze formula $D_I^{cl} = E(f(X)f(X_I, Z_{-I})) - f_0^2$ which

gives the pick-and-freeze Monte Carlo estimator

$$\hat{D}_I^{cl} = \frac{1}{n}\sum_{k=1}^{n} f(x^{(k)}) f\left(x_I^{(k)}, z_{-I}^{(k)}\right) - f_0^2. \tag{16}$$

Here Z is an independent copy of the random variable X, and $-I$ denotes the complement set $[d] \setminus I$. Since, the pick-and-freeze estimator gets a large variance if f_0 is large, other formulas have been suggested that avoid the subtraction of f_0^2 [18, 31]. In particular, the total sensitivity index can be computed using the *Jansen formula* $D_I^T = 1/2E((f(X) - f(Z_I, X_{-I}))^2)$.

Computationally cheaper than Monte Carlo estimation, but also slightly biased, are frequency-based estimation methods. The first frequency-based estimation method was the so-called Fourier amplitude sensitivity test (FAST) by Cukier et al. [5]. TIIs can be easily estimated via the relationship with closed sensitivity indices using pick-and-freeze. Of particular interest is a direct method using the formula of [21]:

$$D_{i,j}^{sup} = \frac{1}{4}E((f(X_i, X_j, X_{-\{i,j\}}) - f(X_i, Z_j, X_{-\{i,j\}}) \tag{17}$$

$$-f(Z_i, X_j, X_{-\{i,j\}}) + f(Z_i, Z_j, X_{-\{i,j\}}))^2). \tag{18}$$

The corresponding Liu-Owen Monte Carlo estimator is unbiased, and it is non-negative since it is a sum of squares. This implies that if the true TII is zero, then the estimator is zero as well.

3 Visualizing Interaction Structures by FANOVA Graphs

In this section the FANOVA graph, an intuitive tool to visualize the most valuable information of the FANOVA decomposition, is introduced [8, 10, 26]. Estimation and thresholding of FANOVA graphs is discussed, and we apply GSA to a standard non-linear test function.

3.1 General Idea of FANOVA Graphs

Usually, it is infeasible to look at all $2^d - 1$ terms of the decomposition of a function with d input variables individually, and quite often only main effect Sobol' indices are considered. The primal intention of FANOVA graphs is to overcome this problem and to visualize the interaction structure contained in the FANOVA decomposition by a mathematical graph [26]. The so-called *FANOVA graph* is defined as a graph $G = (V, E)$ where each of the d input variables is identified

by an element of the vertex set $V = \{1, \ldots, d\}$. An edge is included in the edge set $(i, j) \in E$ iff there exists a superset $J \supseteq \{i, j\}$ such that $f_J(X_J) \neq 0$. That is, the pair of input variables (X_i, X_j) has a non-zero two-way interaction or is involved in a higher order non-zero interaction. Equivalently an edge (i, j) is not in G iff all Sobol' indices $S_J = 0$ for $J \supseteq \{i, j\}$. This is exactly captured by a non-zero TII, i.e. $S_{i,j}^{sup} \neq 0$.

A FANOVA graph can be further enhanced by displaying the thicknesses of each edges (i, j) proportional to the strength of the TII of the two involved input variables. In the same way, each vertex i can be displayed by circles with lines proportional in strength to the main effect Sobol' index S_i.

Let us consider the so-called Ishigami function which is frequently used for illustrating sensitivity analysis [16]. The function, given by

$$f(X_1, X_2, X_3) = \sin(X_1) + 7\sin^2(X_2) + 0.1X_3^4 \sin(X_1), \tag{19}$$

depends on three input variables (X_1, X_2, X_3) and obviously contains a non-linear interaction between X_1 and X_3 (see Fig. 1c). For this test function Sobol' indices can be computed analytically. Assuming a uniform distribution on $[-\pi, \pi]$ for each input variable, analytical calculation of the FANOVA decomposition and Sobol' indices gives us the following values

$$D_1 = 4.346,\ D_2 = 6.125,\ D_3 = 0,\ D_{12} = 0,\ D_{13} = 3.374,\ D_{23} = 0,\ D_{123} = 0. \tag{20}$$

This leads to the following first-order Sobol' indices and normalized TIIs

$$S_1 = 0.314,\ S_2 = 0.442,\ S_3 = 0,\ S_{12}^{sup} = 0,\ S_{13}^{sup} = 0.244,\ S_{23}^{sup} = 0. \tag{21}$$

Figure 2 shows a bar plot and the FANOVA graph displaying the Sobol' indices and TIIs for the Ishigami function. Main effect stands for the normalized first-order Sobol' indices and interaction is the difference between the scaled total sensitivity

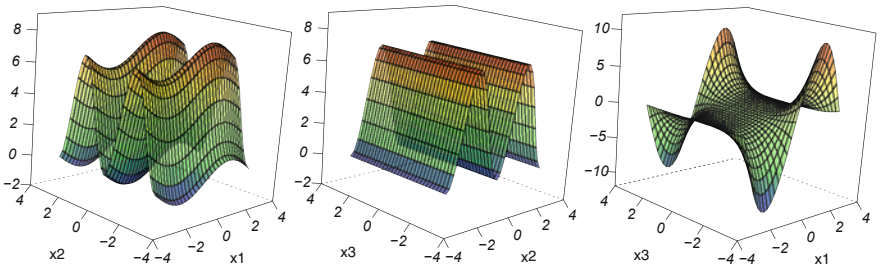

Fig. 1 3d-plots for the Ishigami function. (**a**) $f(X_1, X_2, 0)$. (**b**) $f(0, X_2, X_3)$. (**c**) $f(X_1, 0, X_3)$

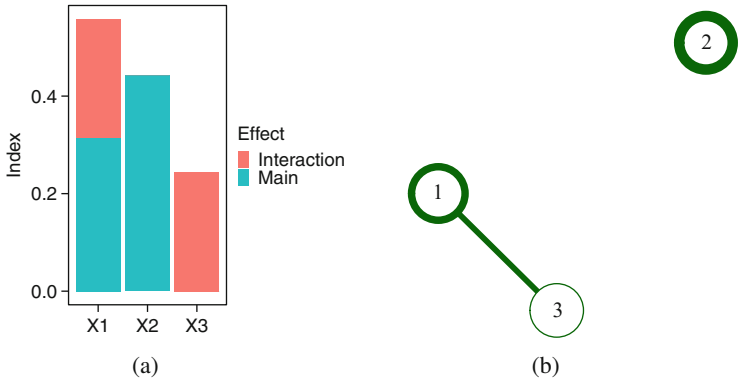

Fig. 2 Bar plot and FANOVA graph displaying Sobol' indices and TIIs for the Ishigami function. (**a**) Bar plot. (**b**) FANOVA graph

index and the Sobol' index. From the FANOVA graph it becomes immediately obvious that the input variable X_2 has the highest impact on the response, followed by X_1. Also, the interaction between X_1 and X_2 is easily detected.

In summary, FANOVA graphs visualize both first- and second-order GSA. First-order analysis in the sense of detecting the inputs X_i for which S_i is very small or even zero. Second-order in the sense of looking at pairs of input variables and detecting influential interactions and their strength, i.e. the pairs $\{i, j\}$ with $S_{i,j}^{sup} > 0$.

3.2 Estimation and Thresholding

In practice, S_i and $S_{i,j}^{sup}$ are usually not analytically available and replaced by estimates $\hat{S}_{i,j}^{sup}$ and $\hat{S}_i$. Moreover, we often even apply GSA not to the actual black-box model but a meta-model or surrogate model of it. Then, estimates are typically not exactly equal to zero even if the "true" or analytically calculated sensitivity index would be. The resulting graph becomes confusing and uninformative. Therefore, edges (i, j) may be included into the graph only if

$$\hat{S}_{i,j}^{sup} > \delta \tag{22}$$

for some small threshold δ, e.g. $\delta = 0.01$ [26]. The computation of the FANOVA graph has been implemented in the R package `fanovaGraph`, providing several estimation methods as well as a thresholding functionality [8, 10].

To exemplify, let us now assume that we cannot analyse the Ishigami function analytically. Based on a random Latin hypercube design with 100 design points, we build a Kriging model of the Ishigami function. Our Kriging model has the usual

Table 1 Sobol' indices (first order) and TIIs for the Kriging model of the Ishigami function

	$\hat{S}_i$	$\hat{\phi}_i^*$	$\hat{S}_i^T$		$\hat{S}_{i,j}^{sup}$
X_1	0.300	0.447	0.603	X_1X_2	0.008
X_2	0.391	0.406	0.420	X_1X_3	0.273
X_3	0.009	0.148	0.285	X_2X_3	0.010

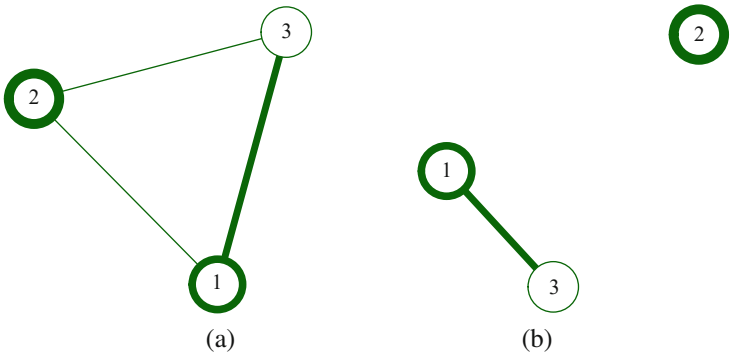

Fig. 3 Fanova graphs without and with thresholding for the Kriging model of the Ishigami function. (**a**) Fanova graph. (**b**) Fanova graph with threshold $\delta = 0.025$

Matern 5/2 covariance structure, no trend and no nugget effect. Table 1 shows the results for the estimators of the first-order Sobol' indices $\hat{S}_i$, normalized Shapley values $\hat{\phi}_i^*$ and TIIs $\hat{S}_{i,j}^{sup}$ of the Kriging model. These estimators have been computed using the R packages `fanovaGraph` [10] and `sensitivity` [15]. Remember that the inequalities

$$S_i^{cl} \leq \phi_i^* \leq S_i^T \tag{23}$$

hold and that $S_i^{cl} = S_i$, which is reflected by the order of the values in Table 1. Comparison also shows that the estimates slightly deviate from the true values given above.

Figure 3 displays on the left hand side the resulting pure FANOVA graph. This is a complete graph as all estimated TIIs are different from zero, even if only slightly. Therefore, we threshold the values by $\delta = 0.025$ and gain the graph on the right hand side, which is the same as for the analytical evaluation of the Ishigami function. The TII's and the FANOVA graph help to discover an underlying block-additive structure of the function f, i.e. we can find a decomposition into cliques of input variables such that variables in different cliques do not interact. As outlined in [26], the detected interaction structure by the FANOVA graph can be a valuable aid in constructing block-additive Kriging models. Therefore, the `fanovaGraph` package also contains methods for block-additive Kriging analysis. The block-additive decomposition provided by the FANOVA graph can also be used in a parallelized global optimization procedure [17].

4 Fields of Applications

The interpretation of a machine learning model by global sensitivity indices and FANOVA graph is in general applicable to any kind of model with a continuous response variable. We show examples of a Kriging model of a piston simulator and an ANN of resistance of sailing yachts.

4.1 Kriging Model of a Piston Simulator

As an example for the application in the field of the design and analysis of computer experiments we are using the piston simulator from the `mistat` package in R [1], first presented in [19]. A piston is moving within a cylinder. The piston's performance is measured by the time it takes to complete one cycle, in seconds. Here, we take the mean of 50 cycles as response, since the cycle time of the piston fluctuates strongly. The following factors can affect the piston's performance. The ranges, in which these factors are varied uniformly in our sensitivity analysis, are given in brackets.

m	The impact pressure determined by the piston mass (30–60) $[kg]$.
S	The piston surface area (0.005–0.020) $[m^2]$.
V_0	The initial volume of the gas inside the piston (0.002-0.010) $[m^3]$.
k	The spring coefficient (1000–5000) $[N/m^3]$.
p_0	The atmospheric pressure $(9 \cdot 10^4 - 11 \cdot 10^4)$ $[N/m^2]$.
T	The surrounding ambient temperature (290–296) $[K]$.
T_0	The filling gas temperature (340–360) $[K]$.

Based on a random Latin hypercube design with 70 design points, we build a Kriging model of the piston simulator. The Kriging model has a Matern 5/2 covariance kernel, no trend and no nugget effect. Table 2 shows the results for the Sobol' indices (first-order and total) and the Shapley values of the piston simulator. The slightly negative value for, e.g. $\hat{\phi}_6^*$ is of course an artefact of the estimation method. We observe that the piston surface $X_2 = S$ and the spring coefficient $X_4 = k$ have the largest effect on the cycle time.

Figure 4 displays the FANOVA graph for the Kriging model of the piston simulator after thresholding by $\delta = 0.005$.

In Fig. 4a both the edges as well as the vertices of the graph are presented by lines proportional to the values of the respective indices. It becomes obvious that $X_2 = S$ and $X_4 = k$ have the highest impact, followed by $X_1 = m$. However, as the values of the TIIs are noticeably smaller than the first-order Sobol' indices, it is not possible to detect which interactions are the largest. Therefore, in the FANOVA

Table 2 Sobol' indices for the Kriging model of the piston simulator

	Sobol' $\hat{S}_i$	Shapley $\hat{\phi}_i^*$	Total Sobol' $\hat{S}_i^T$
X_1	0.109	0.091	0.103
X_2	0.375	0.416	0.423
X_3	0.062	0.069	0.082
X_4	0.353	0.401	0.414
X_5	−0.000	0.014	0.003
X_6	−0.002	−0.009	0.007
X_7	0.026	0.018	0.036

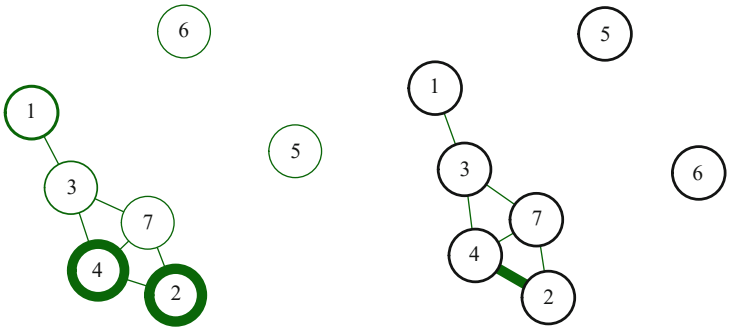

Fig. 4 FANOVA graph with threshold $\delta = 0.005$ for Kriging model of the piston simulator. (**a**) FANOVA graph. (**b**) FANOVA graph (only TIIs displayed)

graph in Fig. 4b we only vary the edges in strength proportional to the values of the TIIs. The largest TII is observed for the interaction $X_2X_4 = Sk$ with $\hat{S}_{2,4}^{sup} = 0.033$.

4.2 Neural Net Model of Resistance of Sailing Yachts

The residuary resistance of a ship is its total resistance minus the viscous resistance. In this section we are studying the residual resistance of sailing yachts in dependence of their hull geometry and the yacht velocity.

The Delft systematic yacht hull series data set [11] comprises $308 = 22 \cdot 14$ experiments with yacht models of scale 6.25 performed at the Delft Ship Hydromechanics Laboratory. In total, 22 different hull forms were tested with 14 different velocities. Based on the Delft series, semi-empirical models were developed [11] which are widely used in the yacht industry [22]. The Delft data set has 6 regressors and one dependent variable, all of which are dimensionless, i.e. their unit is 1 or % or ‰. Let the weight displacement Δ be the weight of water equivalent to the immersed volume of the hull. Then the dependent variable is the

ratio R_r/Δ of the residuary resistance R_r to the weight displacement, given in ‰. The independent variables are as follows.

X_1 The longitudinal centre of buoyancy (LCB) is the longitudinal distance, given in % of some characteristic length, from a point of reference (often midships) to the centre of the displaced volume of water.

X_2 The prismatic coefficient $C_p = \nabla/L_{WL}A_m$ with A_m the cross-sectional area of the underwater slice at midships. C_p displays the ratio of the immersed volume of the hull to a volume of a prism with equal length and cross-sectional area A_m.

X_3 The length-displacement ratio $L_{WL}/\nabla^{1/3}$ where the volume displacement ∇ is the volume of water displaced by the hull.

X_4 The beam-draught ratio B_{WL}/T where the draught T is the maximal distance from the water line to the bottom of the keel.

X_5 The length-beam ratio L_{WL}/B_{WL} is the ratio of length to maximal width at water line.

X_6 The Froude number $Fr = u/\sqrt{gL_{WL}}$. Here u is the flow velocity relative to the yacht, g the gravitational acceleration, and L_{WL} is the length of the hull at water line.

We train a single hidden layer ANN to learn the relationship between input and output variables in the Delft data set. Such ANNs are implemented in the `nnet` package in R [35]. For regression, we choose an ANN with a linear activation function to the output neuron. The data set is divided into training, validation and testing subsets, containing 50%, 25% and 25% of the samples, respectively. The hyperparameter to be tuned is the number n of neurons in the hidden layer. We choose the ANN with lowest validation error, i.e. highest R^2-value for the validation data. That is according to Table 3 an ANN with 8 hidden neurons, i.e. a 6-8-1 net with $6 \cdot 8 + 8 = 56$ weights and $8 + 1 = 9$ biases.

For the chosen ANN as black-box function we perform a GSA. We compute the Sobol' indices as well as the scaled TIIs using the Liu-Owen method with $n = 100,000$ Monte Carlo samples with the help of the fanovaGraph package in R. Table 4 displays these sensitivity indices with the following coding of the regressors $X_1 = LCB$, $X_2 = C_p$, $X_3 = L_{WL}/\nabla^{1/3}$, $X_4 = B_{WL}/T$, $X_5 = L_{WL}/B_{WL}$ and $X_6 = Fr$. The Sobol' indices and scaled TIIs are graphically displayed in the bar plot in Fig. 5a.

Table 3 R^2 values for trainings, validation and test data for ANNs with different number of hidden neurons

m	2	3	4	5	6	7	8	9
R^2_{train}	0.9969	0.9992	0.9991	0.9998	0.9997	0.9998	0.9998	0.9997
R^2_{valid}	0.9972	0.9970	0.9859	0.9989	0.9985	0.9969	0.9990	0.9988
R^2_{test}	0.9962	0.9943	0.9939	0.9971	0.9924	0.9954	0.9957	0.9970

Table 4 Sobol' and Shapley values for the neural net model

	Sobol' $\hat{S}_i$	Shapley $\hat{\phi}_i^*$	Total Sobol' $\hat{S}_i^T$
X_1	0.024	0.020	0.142
X_2	0.006	0.042	0.037
X_3	0.033	0.022	0.071
X_4	0.028	0.163	0.237
X_5	0.021	0.031	0.087
X_6	0.591	0.722	0.688

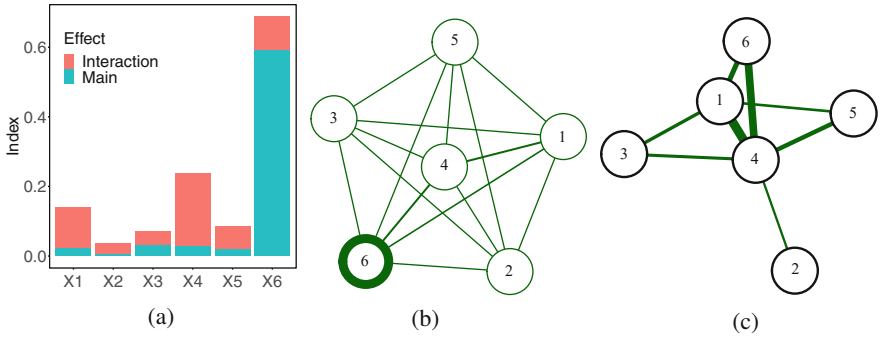

Fig. 5 Bar plot and FANOVA graphs without and with thresholding for the ANN model. (**a**) Bar plot of first-order and total Sobol' indices. (**b**) FANOVA graph. (**c**) FANOVA graph with threshold $\delta = 0.025$ (only TIIs displayed)

Figure 5 displays the FANOVA graph for the ANN model with and without thresholding. The Froude number, a proxy for velocity, has by far the largest impact on the residuary resistance. The largest interactions are X_1X_4, X_4X_6 and X_1X_6.

5 Summary

We have discussed the usefulness of GSA as a tool for interpretable machine learning. Global sensitivity indices based on Sobol' indices, Shapley values as well as derivative-based global sensitivity measures are revisited. FANOVA graphs allow for a very intuitive visualization of interaction structures and the strength of first-order Sobol' indices and TIIs. The approach is exemplified with a Kriging meta-model for a piston simulator and an ANN model for the resistance of yachts.

Acknowledgments The financial support of the Deutsche Forschungsgemeinschaft (SFB 823, project B1) is gratefully acknowledged.

References

1. Amberti, D.: mistat: Data Sets, Functions and Examples from the Book: "Modern Industrial Statistics" by Kenett, Zacks and Amberti (2018). https://CRAN.R-project.org/package=mistat. R package version 1.0-5
2. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Haussler, D. (ed.) Proceedings of the Fifth Annual Workshop on Computational Learning Theory – COLT '92, pp. 144–152. ACM Press, New York (1992). https://doi.org/10.1145/130385.130401
3. Cheng, K., Lu, Z., Zhou, Y., Shi, Y., Wei, Y.: Global sensitivity analysis using support vector regression. Appl. Math. Model. **49**(4), 587–598 (2017). https://doi.org/10.1016/j.apm.2017.05.026
4. Cortez, P., Embrechts, M.J.: Using sensitivity analysis and visualization techniques to open black box data mining models. Inf. Sci. **225**(1), 1–17 (2013). https://doi.org/10.1016/j.ins.2012.10.039
5. Cukier, R., Levine, H., Shuler, K.: Nonlinear sensitivity analysis of multiparameter model systems. J. Comput. Phys. **26**(1), 1–42 (1978). https://doi.org/10.1016/0021-9991(78)90097-9
6. Efron, B., Stein, C.: The jackknife estimate of variance. Ann. Stat. **9**(3) (1981). https://doi.org/10.1214/AOS/1176345462
7. Friedman, J.H., Popescu, B.E.: Predictive learning via rule ensembles. Ann. Appl. Stat. **2**(3), 916–954 (2008). https://doi.org/10.1214/07-AOAS148
8. Fruth, J., Roustant, O., Muehlenstaedt, T.: The fanovaGraph Package: Visualization of Interaction Structures and Construction of Block-additive Kriging Models (2013). https://hal.archives-ouvertes.fr/hal-00795229
9. Fruth, J., Roustant, O., Kuhnt, S.: Total interaction index: a variance-based sensitivity index for second-order interaction screening. J. Stat. Plann. Infer. **147**, 212–223 (2014). https://doi.org/10.1016/j.jspi.2013.11.007
10. Fruth, J., Muehlenstaedt, T., Roustant, O., Jastrow, M., Kuhnt, S.: fanovaGraph: building Kriging Models from FANOVA Graphs (2020). https://CRAN.R-project.org/package=fanovaGraph. R package version 1.5
11. Gerritsma, J., Onnink, R., Versluis, A.: Geometry, resistance and stability of the delft systematic yacht hull series. Int. Shipbuild. Prog. **28**(328), 276–297 (1981). https://doi.org/10.3233/ISP-1981-2832801
12. Hastie, T., Tibshirani, R.J.: Generalised additive models. In: Monographs on Statistics and Applied Probability, vol. 43. Chapman and Hall, London (1990)
13. Homma, T., Saltelli, A.: Importance measures in global sensitivity analysis of nonlinear models. Reliab. Eng. Syst. Saf. **52**(1), 1–17 (1996). https://doi.org/10.1016/0951-8320(96)00002-6
14. Iooss, B., Lemaître, P.: A review on global sensitivity analysis methods. In: Dellino, G., Meloni, C. (eds.) Uncertainty Management in Simulation-Optimization of Complex Systems, Operations Research/Computer Science Interfaces Series, vol. 59, pp. 101–122. Springer US, Boston (2015). https://doi.org/10.1007/978-1-4899-7547-8_5
15. Iooss, B., Veiga, S.D., Janon, A., Pujol, G., with contributions from Baptiste Broto, Boumhaout, K., Delage, T., Amri, R.E., Fruth, J., Gilquin, L., Guillaume, J., Le Gratiet, L., Lemaitre, P., Marrel, A., Meynaoui, A., Nelson, B.L., Monari, F., Oomen, R., Rakovec, O., Ramos, B., Roustant, O., Song, E., Staum, J., Sueur, R., Touati, T., Weber, F.: Sensitivity: global Sensitivity Analysis of Model Outputs (2020). https://CRAN.R-project.org/package=sensitivity. R package version 1.22.1
16. Ishigami, T., Homma, T.: An importance quantification technique in uncertainty analysis for computer models. In: 1990 Proceedings of First International Symposium on Uncertainty Modeling and Analysis, pp. 398–403. IEEE Computer Society Press, Washington (1990). https://doi.org/10.1109/ISUMA.1990.151285

17. Ivanov, M., Kuhnt, S.: A parallel optimization algorithm based on FANOVA decomposition. Qual. Reliab. Eng. Int. **30**(7), 961–974 (2014). https://doi.org/10.1002/qre.1710
18. Jansen, M.J.: Analysis of variance designs for model output. Comput. Phys. Commun. **117**(1–2), 35–43 (1999). https://doi.org/10.1016/S0010-4655(98)00154-4
19. Kenett, R., Zacks, S., Amberti, D.: Modern Industrial Statistics: With Applications in R, MINITAB and JMP, 2nd edn. Statistics in Practice. Wiley, Chichester (2014)
20. Kucherenko, S., Rodriguez-Fernandez, M., Pantelides, C., Shah, N.: Monte Carlo evaluation of derivative-based global sensitivity measures. Reliab. Eng. Syst. Saf. **94**(7), 1135–1148 (2009). https://doi.org/10.1016/j.ress.2008.05.006
21. Liu, R., Owen, A.B.: Estimating mean dimensionality of analysis of variance decompositions. J. Am. Stat. Assoc. **101**(474), 712–721 (2006). https://doi.org/10.1198/016214505000001410
22. Lopez Gonzalez, R.: Neural networks for variational problems in engineering. PhD Thesis. Technical University of Catalonia (2008)
23. McCullagh, P., Nelder, J.A.: Generalized linear models. Monographs on Statistics and Applied Probability, vol. 37, 2nd edn. Chapman and Hall, London (1989)
24. Molnar, C.: Interpretable Machine Learning. lulu.com (2020)
25. Molnar, C., Casalicchio, G., Bischl, B.: Quantifying model complexity via functional decomposition for better post-hoc interpretability. In: Cellier, P., Driessens, K. (eds.) Machine Learning and Knowledge Discovery in Databases. Communications in Computer and Information Science, vol. 1167, pp. 193–204. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-43823-4_17
26. Muehlenstaedt, T., Roustant, O., Carraro, L., Kuhnt, S.: Data-driven kriging models based on Fanova-decomposition. Stat. Comput. **22**(3), 723–738 (2012). https://doi.org/10.1007/s11222-011-9259-7
27. Owen, A.B.: Sobol' indices and Shapley value. SIAM/ASA J. Uncertain. Quantif. **2**(1), 245–251 (2014). https://doi.org/10.1137/130936233
28. Roustant, O., Fruth, J., Iooss, B., Kuhnt, S.: Crossed-derivative based sensitivity measures for interaction screening. Math. Comput. Simul. **105**, 105–118 (2014). https://doi.org/10.1016/j.matcom.2014.05.005
29. Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S.: Global Sensitivity Analysis. The Primer. John Wiley & Sons, Ltd, Chichester (2007). https://doi.org/10.1002/9780470725184
30. Shapley, L.S.: A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (eds.) Contributions to the Theory of Games (AM-28), vol. II, pp. 307–318. Princeton University Press, Princeton (1953). https://doi.org/10.1515/9781400881970-018
31. Sobol, I.M.: Sensitivity analysis for non-linear mathematical models. Math. Modeling Comput. Experiment **1**(4), 407–414 (1993)
32. Sobol, I., Gershman, A.: On an alternative global sensitivity estimators. In: Proceedings of SAMO 1995, Belgirate, pp. 40–42 (1995)
33. Song, E., Nelson, B.L., Staum, J.: Shapley effects for global sensitivity analysis: theory and computation. SIAM/ASA J. Uncertain. Quantif. **4**(1), 1060–1083 (2016). https://doi.org/10.1137/15M1048070
34. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. **41**(3), 647–665 (2014). https://doi.org/10.1007/s10115-013-0679-x
35. Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, 4th edn. Springer, New York (2002). ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4
36. Winter, E.: The Shapley value. In: Handbook of Game Theory with Economic Applications, vol. 3, pp. 2025–2054. Elsevier, Amsterdam (2002). https://doi.org/10.1016/S1574-0005(02)03016-3
37. Zhang, P.: A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model. Appl. Soft Comput. **85**, 105859 (2019). https://doi.org/10.1016/j.asoc.2019.105859