Random Forest based Qantile Oriented Sensitivity Analysis indices estimation

Kévin Elie-Dit-Cosaque, Véronique Maume-Deschamps

Computational Statistics

Random forest based quantile-oriented sensitivity analysis indices estimation

--Manuscript Draft--

Manuscript Number:	COST-D-23-00288R1
Full Title:	Random forest based quantile-oriented sensitivity analysis indices estimation
Article Type:	Original Paper
Keywords:	Quantile-oriented sensitivity analysis; Random forest; Cross validation; Out-ofbag samples
Manuscript Classifications:	1.03120: Nonparametric Techniques; 1.04050: Sensitivity Analysis; 1.04560: Statistical Learning

Corresponding Author:	Véronique Maume-Deschamps Universite Claude Bernard Lyon 1 FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universite Claude Bernard Lyon 1
Corresponding Author's Secondary Institution:	
First Author:	Kevin Elie-dit-Cosaque
First Author Secondary Information:	
Order of Authors:	Kevin Elie-dit-Cosaque
	Véronique Maume-Deschamps
Order of Authors Secondary Information:	

Funding Information:	SCOR SE	Pr Véronique Maume-Deschamps

Abstract:	We propose a random forest based estimation procedure for Quantile-Oriented Sensitivity Analysis - QOSA. In order to be efficient, a cross-validation step on the leaf size of trees is required. Our full estimation procedure is tested on both simulated data and a real dataset. Our estimators use either the bootstrap samples or the original sample in the estimation. Also, they are either based on a quantile plug-in procedure (the R-estimators) or on a direct minimization (the Q-estimators). This leads to 8 different estimators which are compared on simulations. From these simulations, it seems that the estimation method based on a direct minimization is better than the one plugging the quantile. This is a significant result because the method with direct minimization requires only one sample and could therefore be preferred.

Kévin Elie-Dit-Cosaque, Véronique Maume-Deschamps
Institut Camille Jordan
Lyon

November 19, 2023

Associate editor and reviewers
Computational Statistics

Dear Editor and Reviewers,

We would like to thank you for your very helpful comments and reports. Please find below our answers to the points raised by the reviewers.

Reviewer 1

- It seems that the random forest estimator can be replaced by any non-parametric estimator. Thus, it is better to describe why the random forest algorithm is employed in this paper.
  Random forest is known to be an efficient and robust non parametric estimation method. Also, its use in GSA develops (see e.g. [Antoniadis et al., 2021]) and this is why we focus on it but other non parametric regression methods could be used. We have added this sentence in the introduction.
- On page 3 (line 54), we can find the setting of this paper, such as $Y = f(\mathbf{X})$. However, in the statistical problems, this assumption is quite unnatural, and the proposed estimators can be applied even for the model $Y = f(\mathbf{X}) + \varepsilon$. Although the authors describe some reasons on page 9 (lines 23-25), this reason is just for the convenience of numerical experiments and not essential for the estimation procedure. In addition, the theoretical true values can be numerically calculated by the Monte Carlo method. Thus, it is better to consider the more realistic model $Y = f(\mathbf{X}) + \varepsilon$ at least in Sections 2-4.
  The setting $Y = f(\mathbf{X})$ is the general Global Sensitivity Analysis framework, see e.g. [Da Veiga et al., 2021]. But this is true that we could also consider noisy models: $Y = f(\mathbf{X}) + \varepsilon$ with ε a centered noise, independent of $\mathbf{X}$, which is closer to statistical frameworks. We have added this sentence at the beginning of Section 2. Also, we have added one simulation with a noisy model (see Table 2 in Section 6).
- Related to the above comments, all current numerical experiments are performed under the setting $Y = f(\mathbf{X})$. However, it is important to know whether we can get similar results even under the more realistic model $Y = f(\mathbf{X}) + \varepsilon$. I understand that redoing numerical experiments with different settings requires a significant amount of effort. Thus, it would be nice if the authors provided a few additional experiments with the model $Y = f(\mathbf{X}) + \varepsilon$ to show that the current experiments are enough to understand the behaviors of all estimators. Especially, I'm curious about the results in Table 1 under the model $Y = f(\mathbf{X}) + \varepsilon$.
  We did one simulation for the model $Y = X_1 - X_2 + \varepsilon$ with the X_i's following an exponential distribution with parameter 1 and ε following a centered normal distribution with standard deviation 0.5. We have computed the theoretical QOSA indices with a Monte Carlo method on a large sample (size 10^7). We have estimated the QOSA indices with $Q^{1,o}$ and $Q^{2,o}$ which have the best compromise *time cost* vs *accuracy*, and the kernel estimator $\widetilde{S}$. The order of magnitude of the RMSE are the same for the noisy and un-noisy model, this can be seen in Table 2, Section 6.
- If possible, it is better to provide some insights into the results of numerical experiments. For example, in Figure 5, it seems that the S_1^{α} with $Q_1^{2,b}$, is depending on α. It would be nice if the authors provided insights into this kind of unexpected behavior.`

This is indeed an unexpected behavior, we do not have a real explanation, so that we did not add comments on that point.

Reviewer 2

- The abstract should provide a more detailed overview of the paper's content. It should highlight the introduction of quantile-based estimators, minimum-based estimators, and the use of the original dataset and bootstrap samples. Additionally, the abstract should incorporate key simulation results and conclusions to better represent the paper's contributions.
  We have added these precisions in the abstract.
- Please ensure proper indentation at the beginning of paragraphs following line breaks, such as line 25 on page 2.
  We are not sure to understand properly the rules on this indentation. We feel that this point would be treated with the editorial team, if the paper is accepted.
- On page 4, lines 45-46: Clarify whether Θ_i, $(i = 1, 2)$ used in $\Theta = (\Theta_1, \Theta_2)$ and (Θ_ℓ), $\ell = 1, ..., k$ defined in the previous sentence refer to the same variables or different ones. The use of these symbols may lead to confusion.
  We changed to $\Theta = (\Theta^1, \Theta^2)$.
- On page 8, lines 18-20: It appears that ... appears to be a scalar. Additionally, please correct the index l in the outer summation to ℓ for consistency.
  Done.
- On page 9, lines 19-20: Provide an explicit explanation of what $\mathcal{E}(1)$ refers to. While on page 18, it is mentioned that $\mathcal{E}(\lambda_i)$ is an exponential distribution, it should be clarified when first introduced in the paper.
  This has been clarified on page 9.
- In the second line of Algorithm 1: It is unclear what $\mathbb{N}^\star$ represents. Please provide a clear definition or explanation to avoid confusion.
  $\mathbb{N}^\star$ has been replaced by $\mathbb{N} \setminus \{0\}$.
- In Figure 5 on page 16: It is observed that $Q_i^{1,o}$ and $Q_i^{2,o}$ exhibit better accuracy in estimation compared to $Q_i^{1,b}$ and $Q_i^{2,b}$. It would be valuable to discuss the reasons behind this difference. Explore why results based on the original sample outperform those based on the bootstrap method in this particular case.
  It seems that the Q-estimators have more bias when the bootstrap sample is reused in the estimation (instead of the original sample). This could be an over-fitting effect which is also observed with less amplitude for the bootstrap R-estimator. We added this sentence in Section 6.

We hope our modifications properly address the issues raised in the reports. We are looking forward to hearing from you.

All the best,

Kévin Elie-Dit-Cosaque, Véronique Maume-Deschamps

References

[Antoniadis et al., 2021] Antoniadis, A., Lambert-Lacroix, S., and Poggi, J.-M. (2021). Random forests for global sensitivity analysis: A selective review. Reliability Engineering & System Safety, 206.

[Da Veiga et al., 2021] Da Veiga, S., Gamboa, F., Iooss, B., and Prieur, C. (2021). Basics and trends in sensitivity analysis: Theory and practice in R.

Random forest based quantile-oriented sensitivity analysis indices estimation

Kévin Elie-Dit-Cosaque *[1, 2] and Véronique Maume-Deschamps †[1]

[1]Universite Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, F-69622 Villeurbanne, France
[2]Model Analysis, SCOR, Paris, FRANCE

November 19, 2023

Abstract

We propose a random forest based estimation procedure for Quantile-Oriented Sensitivity Analysis - QOSA. In order to be efficient, a cross-validation step on the leaf size of trees is required. Our full estimation procedure is tested on both simulated data and a real dataset. Our estimators use either the bootstrap samples or the original sample in the estimation. Also, they are either based on a quantile plug-in procedure (the R-estimators) or on a direct minimization (the Q-estimators). This leads to 8 different estimators which are compared on simulations. From these simulations, it seems that the estimation method based on a direct minimization is better than the one plugging the quantile. This is a significant result because the method with direct minimization requires only one sample and could therefore be preferred.

Keywords: Quantile-oriented sensitivity analysis, Random forest, Cross validation, Out-of-bag samples.

1 Introduction

Numerical models are ubiquitous in various fields, such as aerospace, economy, environment or insurance, and allow to approximate the behavior of physical phenomenon. Their main advantage is that they replace expensive, or even unachievable, real-life experiments and thus provide knowledge about the natural system. The extremely faithful representation of reality, favored by easier use of large datasets nowadays thanks to the increase in computing power, also explains this widespread use. However, this accuracy is often synonymous of complexity, ultimately leading to a difficult interpretation of models. Besides, model inputs are usually uncertain due to a lack of information or the random nature of factors, which means that the resulting output can be regarded as random. It is then important to assess the impact of this uncertainty on the model output. Global Sensitivity Analysis (GSA) methods solve these issues by *studying how the uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model inputs* (Saltelli et al., 2004). Hence, GSA allows to investigate input-ouput relationships by identifying the inputs that strongly influence or not the model response.

*edckev@gmail.com
†veronique.maume@univ-lyon1.fr

Variance-based approaches are well-established and widely used for GSA. Among them, the sensitivity indices developed by Sobol (1993) are very popular. This last method stands on the assumption that the inputs are independent. Under this hypothesis, the overall variance of a scalar output can be split down into different partial variances using the so-called Hoeffding (1948) decomposition. Then, the first-order Sobol' index quantifies the individual contribution of an input to the output variance while the total Sobol' index (Jansen et al., 1994; Homma and Saltelli, 1996) measures the marginal and interaction effects. However, even if they are extremely popular and informative measures, variance-based approaches suffer from some limitation. Indeed, by definition, they only study the impact of the inputs on the expectation of the output as they compare the conditional expectations with the unconditional expectation of the output distribution, by using the variance as a distance measure.

A new class of sensitivity indices, generalizing the first-order Sobol' index to other quantities of interest than the expectation, has been introduced in Fort et al. (2016). These indices called Goal-Oriented Sensitivity Analysis (GOSA) compare the minimum of a specific contrast function to its conditional counterpart when one of the inputs is fixed. The unconditional minimum is reached by the quantity of interest (for instance a quantile).

In this paper, we focus on Quantile-Oriented Sensitivity Analysis (QOSA) measuring the impact of the inputs on the α-quantile of the output distribution. Browne et al. (2017); Maume-Deschamps and Niang (2018) introduced a statistical estimator of the first-order QOSA index based on a kernel approach. Kala (2019) defined the second and higher order QOSA indices as well as a variance-like decomposition for quantiles in the case of independent inputs. Elie-Dit-Cosaque and Maume-Deschamps (2022a) thoroughly studied theoretical properties of QOSA indices and illustrated some of their limitations on various toy models in independent and dependent contexts. That led them to define new generic indices based on the Shapley values named Goal-Oriented Shapley effects (GOSE), in particular, Quantile-Oriented Shapley effects (QOSE) when considering the quantile of the output as feature of interest.
Despite these recents works, the question of the effective estimation of the first-order QOSA index remains open. Indeed, it turns out to be difficult to compute in practice because it requires an accurate estimate of either the conditional quantile of the output given an input, or the minimum of a conditional expectation of the output given an input. Kala (2019) handles this feature with a brute force Monte-Carlo approach. As a matter of fact, for each value of an input variable, realizations of the other inputs are generated conditionally to this fixed value. Therefore, this leads to a computational cost that is too heavy to consider its use in an industrial context when dealing with costly models. Besides, when dealing with dependent inputs, this approach requires the knowledge of the dependency structure of inputs in order to sample the conditional distributions, which is not always the case. Browne et al. (2017); Maume-Deschamps and Niang (2018) developed kernel-based estimators to avoid this double-loop issue. But, when using a small dataset, their performance is highly dependent of the bandwidth parameter. Browne et al. (2017) proposed a cumbersome algorithm for setting an efficient bandwidth that is not straighforward to implement in practice. As for the estimator of Maume-Deschamps and Niang (2018), a large dataset is needed in order to have a low estimation error, as no algorithm of bandwidth parameter selection is established.

To overcome these issues, we explore the random forest algorithm introduced by Breiman (2001) in order to estimate the conditional distribution of the output given an input. Random forest is known to be an efficient and robust non parametric estimation method. Also, its use in GSA develops (see e.g. Antoniadis et al. (2021)) and this is why we focus on it but other non parametric regression methods could be used. The main contribution of this paper is to provide different estimation strategies of the first-order QOSA index based on this method. Some of the

estimators developed provide a better estimation of QOSA indices than the aforementioned ones while requiring less data. The random forest methodology uses bootstrap samples in order to construct several trees. Our estimators use either the bootstrap samples or the original sample in the estimation. Also, they are either based on a quantile plug-in procedure (the R-estimators) or on a direct minimization (the Q-estimators). This leads to 8 different estimators which are compared on simulations. The R-estimators need 2 independent samples while the Q ones require only one sample. The performance of the Q-estimators appears slightly better on our simulations. Also the Q-estimators require only one sample, so that we would advise to use them rather than the R ones.

The paper is organized as follows. We recall in Section 2 the definition of the first-order QOSA index and initiate the estimation process. Section 3 presents the random forest algorithm and several estimators of the first-order QOSA index based on this method are described in Section 4. The entire process is summarized in Section 5. Then, the performance of the estimators is investigated in Section 6 on simulated data and the relevance of this index is highlighted on a real dataset in Section 7. Finally, a conclusion is given in Section 8.

2 Estimation of the QOSA index

Let us consider the input-output system where $\mathbf{X} = (X_1, \dots, X_d) \in \mathbb{R}^d$ is a random vector of d inputs and $Y = f(\mathbf{X})$ is the output random variable of a measurable deterministic function $f : \mathbb{R}^d \to \mathbb{R}$ which can be a mathematical function or a computational code. This is the general GSA framework, see e.g. Da Veiga et al. (2021). We could also consider noisy models: $Y = f(\mathbf{X}) + \varepsilon$ with ε a centered noise, independent of $\mathbf{X}$. This noisy setting is closer to statistical frameworks. Then, given a level $\alpha \in]0,1[$, Fort et al. (2016) introduced the first-order Quantile-Oriented Sensitivity Analysis (QOSA) index, related to the input X_i, as

$$S_i^\alpha = \frac{\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right] - \mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\middle| X_i\right]\right]}{\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right]} ,$$

with the contrast function $\psi_\alpha : (y,\theta) \mapsto (y-\theta)\left(\alpha - \mathbb{1}_{\{y\leqslant\theta\}}\right)$. This function, also called *pinball loss* or *check function* in the literature is the cornerstone of the quantile regression (Koenker and Hallock, 2001). Quantile and conditional quantile are related to this loss function as follows

$$q^\alpha(Y) = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right] \quad \text{and} \quad q^\alpha(Y|X_i) = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\middle| X_i\right] ,$$

where $q^\alpha(Y)$ is the α-quantile of Y and $q^\alpha(Y|X_i)$, the α-quantile of Y given X_i. Thus, the index S_i^α can be rewritten in the following way,

$$S_i^\alpha = 1 - \frac{\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\middle| X_i\right]\right]}{\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right]} = 1 - \frac{\mathbb{E}\left[\psi_\alpha(Y, q^\alpha(Y|X_i))\right]}{\mathbb{E}\left[\psi_\alpha(Y, q^\alpha(Y))\right]} = 1 - \frac{O}{P} ,$$

where O refers to $\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\middle| X_i\right]\right] = \mathbb{E}\left[\psi_\alpha(Y, q^\alpha(Y|X_i))\right]$ and P, to $\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right] = \mathbb{E}\left[\psi_\alpha(Y, q^\alpha(Y))\right]$.

Hence, as stated in Browne et al. (2017), the index S_i^α compares the mean distance between Y and its conditional quantile to the mean distance between Y and its quantile, where the pinball loss function ψ_α is the considered distance. This index has some basic properties requested for

a reasonable sensitivity index such as $0 \leqslant S_i^\alpha \leqslant 1$, $S_i^\alpha = 0$ if Y is independent of X_i and $S_i^\alpha = 1$ if Y is X_i measurable.

It should be mentioned that Kucherenko et al. (2019) proposed new indices K^α to assess the impact of inputs on the α-quantile of the output distribution. They directly quantify the mean distance between quantiles $q^\alpha(Y)$ and $q^\alpha(Y|X_i)$ rather than the mean distance between average contrast functions like in the first-order QOSA index. Different estimation strategies are investigated in their paper (brute force Monte Carlo and double-loop reordering approach). But a major limitation is that a large sample size is required to get an accurate computation of the index (samples of size 2^{18} are used in their paper). Also, as discussed in the theoretical review of quantile-oriented indices carried out in Elie-Dit-Cosaque and Maume-Deschamps (2022a), the practical interpretation of the K^α indices is questionable while QOSA indices give a relevant interpretation of the impact of the inputs on the α-quantile of the output.

Let us now initiate the estimation procedure for the first-order QOSA index S_i^α, associated to a specific input X_i and a level α.
We consider an i.i.d n-sample $\mathcal{D}_n^\diamond = \left(\mathbf{X}^{\diamond j}, Y^{\diamond j}\right)_{j=1,\dots,n}$ such that $Y^{\diamond j} = f\left(\mathbf{X}^{\diamond j}\right), j = 1,\dots,n$. Then, a first natural estimator of the P term of the QOSA index based on the quantity $\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha(Y)\right)\right]$ is proposed

$$\widehat{P}_1 = \frac{1}{n}\sum_{j=1}^{n} \psi_\alpha\left(Y^{\diamond j}, \widehat{q}^\alpha(Y)\right) , \tag{2.1}$$

with $\widehat{q}^\alpha(Y)$, the classical empirical estimator for $q^\alpha(Y)$ obtained from $\mathcal{D}_n^\diamond$.
The P term can be alternatively estimated as follows by using the quantity $\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right]$,

$$\widehat{P}_2 = \min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{j=1}^{n} \psi_\alpha\left(Y^{\diamond j}, \theta\right) ,$$

where the minimum is reached for one of the elements of $\left(Y^{\diamond j}\right)_{j=1,\dots,n}$. As the function to minimize is decreasing then increasing, this estimator therefore requires to compute $\frac{1}{n}\sum_{j=1}^{n} \psi_\alpha\left(Y^{\diamond j}, Y^{\diamond(k)}\right)$, $k = 1,\dots,n$, until it increases, with $Y^{\diamond(k)}$ the order statistics of $\left(Y^{\diamond 1},\dots,Y^{\diamond n}\right)$. This process is much more time-consuming than the first estimator where it is just needed to compute the quantile and then plug it. Thus, the $\widehat{P}_1$ estimator will be used in the sequel.

The O term of the QOSA index is trickier to estimate because a good approximation of the conditional distribution of Y given X_i is required. Both existing estimators of the QOSA index currently provided in Browne et al. (2017); Maume-Deschamps and Niang (2018) handle this feature thanks to kernel-based methods. But in practice, with these methods, we are faced with determining the optimal bandwidth parameter or using large sample sizes in order to have a sufficiently low estimation error when employing a non optimal bandwidth. Thus, when dealing with costly computational models, a precise enough estimation of these indices can be difficult to achieve or even unfeasible.

We propose in this paper to address these issues by using the random forest method for estimating the conditional distribution. Therefore, several statistical estimators for the O term of the first-order QOSA index will be defined in Section 4. Let us first recall the random forest algorithm.

3 Random forests

Random forests are ensemble learning methods, first introduced by Breiman (2001), which can be used in classification or regression problems. We only focus on their use for regression task and assume to be given a training sample $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$ of i.i.d random variables distributed as the prototype pair $(\mathbf{X}, Y)$.

Breiman's forest grows a collection of k regression trees based on the CART procedure described in Breiman et al. (1984). Building several different trees from a single dataset requires to randomize the tree building process. Randomness injected in each tree is denoted by Θ_ℓ where $(\Theta_\ell)_{\ell=1,\dots,k}$ are independent random variables distributed as Θ (independent of $\mathcal{D}_n$). $\Theta = (\Theta^1, \Theta^2)$ contains indices of observations selected to build the tree and indices of splitting candidate directions in each cell.

In more detail, the ℓ-th tree is built using a bootstrap sample $\mathcal{D}_n^\star(\Theta_\ell)$ from the original dataset. Only these observations are used to construct the tree and to make the tree prediction. Once the observations have been selected, the algorithm forms a recursive partitioning of the input space. In each cell, a number $max\_features$ of variables is selected uniformly at random among all inputs. Then, the best split is chosen as the one optimizing the CART splitting criterion only along the $max\_features$ preselected directions. This process is repeated in each cell. A stopping criterion, often implemented, is that a split point at any depth will only be considered if it leaves at least $min\_samples\_leaf$ samples in each of the left and right child nodes. After tree partition has been completed, the prediction of the ℓ-th tree denoted by $m_n^b(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ at a new point $\mathbf{x}$ is computed by averaging the $N_n^b(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ observations falling into the cell $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ of the new point.
Hence, the random forest prediction is the average of the k predicted values:

$$m_{k,n}^b(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n) = \frac{1}{k}\sum_{\ell=1}^{k} m_n^b(\mathbf{x}; \Theta_\ell, \mathcal{D}_n) = \frac{1}{k}\sum_{\ell=1}^{k}\left(\sum_{j\in\mathcal{D}_n^\star(\Theta_\ell)} \frac{\mathbb{1}_{\{\mathbf{X}^j\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^b(\mathbf{x};\Theta_\ell,\mathcal{D}_n)} Y^j\right) . \tag{3.1}$$

By defining the random variable $B_j(\Theta_\ell^1, \mathcal{D}_n)$ as the number of times that the observation $(\mathbf{X}^j, Y^j)$ has been used from the original dataset for the ℓ-th tree construction, the conditional mean estimator in Equation (3.1) is rewritten as follows

$$m_{k,n}^b(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n) = \sum_{j=1}^{n} w_{n,j}^b(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n)\, Y^j , \tag{3.2}$$

where the weights $w_{n,j}^b(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n)$ are defined by

$$w_{n,j}^b(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n) = \frac{1}{k}\sum_{\ell=1}^{k} \frac{B_j(\Theta_\ell^1, \mathcal{D}_n)\, \mathbb{1}_{\{\mathbf{X}^j\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^b(\mathbf{x};\Theta_\ell,\mathcal{D}_n)} . \tag{3.3}$$

A variant of the Equation (3.2) provides another estimator of the conditional mean. Trees are still grown as in the standard random forest algorithm being based on the bootstrap samples but, for the tree prediction, the original dataset $\mathcal{D}_n$ is used instead of the bootstrap sample $\mathcal{D}_n^\star(\Theta_\ell)$ associated to the ℓ-th tree and we get

$$m_{k,n}^o(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n) = \sum_{j=1}^{n} w_{n,j}^o(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n)\, Y^j , \tag{3.4}$$

where the weights $w^o_{n,j}(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n)$ are defined by

$$w^o_{n,j}(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n)=\frac{1}{k}\sum_{\ell=1}^{k}\frac{\mathbb{1}_{\{\mathbf{X}^j\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N^o_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}\ . \tag{3.5}$$

It has to be noted that contrary to Equation (3.3) where $N^b_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$ refers to the number of elements of $\mathcal{D}^\star_n(\Theta_\ell)$ falling into $A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$, in Equation (3.5), $N^o_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$ is the number of elements of the original dataset $\mathcal{D}_n$ that fall into $A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$.

Thus, both weighted approaches using, either the bootstrap samples (Equation (3.2)) or the original dataset (Equation (3.4)), allow to see the random forest method as a local averaging estimate (Lin and Jeon, 2006; Scornet, 2016) and will be at the heart of the strategies proposed for estimating the O term of the first-order QOSA index. In the following, to lighten the notation, the dependence to Θ and $\mathcal{D}_n$ in the weights will be omitted.

4 Estimation of the O term of the QOSA index

By using the random forest method aforementioned, five estimators of the O term may be defined. The first two rely on the expression $\mathbb{E}\left[\psi_\alpha\left(Y,q^\alpha\left(Y|\,X_i\right)\right)\right]$ and the others on $\mathbb{E}\left[\min_{\theta\in\mathbb{R}}\mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\,X_i\right]\right]$. Since our aim is to estimate conditional expressions with respect to one input variable, say X_i, we shall consider forests driven by X_i, i.e. the random forest is built with the observations $\mathcal{D}^i_n=\left(X^j_i,Y^j\right)_{j=1,\ldots,n}$ from $\mathcal{D}_n$, which means that Y is explained with X_i only. When needed, we shall denote by $\mathcal{D}^{\star i}_n$ a bootstrap sample from $\mathcal{D}^i_n$ and $\mathcal{D}^{\diamond i}_n=(X^{\diamond j}_i,Y^{\diamond j})_{j=1,\ldots,n}$ an independent copy of $\mathcal{D}^i_n$.

4.1 Quantile-based O term estimators

In this section, the estimations of the O term of the QOSA index are based on the quantity $\mathbb{E}\left[\psi_\alpha\left(Y,q^\alpha\left(Y|\,X_i\right)\right)\right]$. Using two training samples $\mathcal{D}^i_n$ and $\mathcal{D}^{\diamond i}_n$, we define

$$\widehat{R}_i=\frac{1}{n}\sum_{j=1}^{n}\psi_\alpha\left(Y^{\diamond j},\widehat{q}^\alpha\left(Y|\,X_i=X^{\diamond j}_i\right)\right)\ ,$$

where the sample $\mathcal{D}^i_n$ is used to get $\widehat{q}^\alpha\left(Y|\,X_i=x_i\right)$, an estimator of the conditional quantile $q^\alpha\left(Y|\,X_i=x_i\right)$. It is obtained thanks to two approaches based on the random forests, described in the sequel.

4.1.1 Quantile estimation with a weighted approach

We consider the estimator of the Conditional Cumulative Distribution Function (C_CDF), using $\mathcal{D}^i_n$ to construct the forest, introduced in Meinshausen (2006) and whose the consistency has been showed in Elie-Dit-Cosaque and Maume-Deschamps (2022b). The C_CDF estimator used to estimate the conditional quantile is

$$F^o_{k,n}\left(y|\,X_i=x_i\right)=\sum_{j=1}^{n}w^o_{n,j}\left(x_i\right)\mathbb{1}_{\{Y^j\leqslant y\}}\ ,$$

where the $w^o_{n,j}\left(x_i\right)$'s are defined in Equation (3.5).

Hence, given a level $\alpha \in [0,1]$, the conditional quantile estimator $\widehat{q}^{\alpha}(Y|X_i = x_i)$ is defined by plugging $F^{o}_{k,n}(y|X_i = x_i)$ instead of $F(y|X_i = x_i)$ as follows

$$\widehat{q}^{\alpha}(Y|X_i = x_i) = \inf_{p=1,\dots,n} \left\{ Y^p : F^{o}_{k,n}(Y^p|X_i = x_i) \geqslant \alpha \right\} .$$

As a result, the estimator of $\mathbb{E}[\psi_\alpha(Y, q^{\alpha}(Y|X_i))]$ based on this method is denoted $\widehat{R}^{1,o}_i$.

4.1.2 Quantile estimation within a leaf

Let us consider a set of k trees indexed by $\ell = 1,\dots,k$ constructed with the sample $\mathcal{D}^i_n$. Once the forest is built with the bootstrap samples of $\mathcal{D}^i_n$, the estimator $\widehat{q}^{o,\alpha}_{\ell}(Y|X_i = x_i)$ of $q^{\alpha}(Y|X_i = x_i)$ for the ℓ-th tree is obtained with the original observations from $\mathcal{D}^i_n$ falling into $A_n(x_i;\Theta_\ell,\mathcal{D}^i_n)$ as follows

$$\widehat{q}^{o,\alpha}_{\ell}(Y|X_i = x_i) = \inf_{p=1,\dots,n} \Big\{ Y^p,\ (X^p_i, Y^p) \in \mathcal{D}^i_n \text{ and } X^p_i \in A_n(x_i;\Theta_\ell,\mathcal{D}^i_n) : \sum_{j=1}^{n} \frac{\mathbb{1}_{\left\{X^j_i \in A_n(x_i;\Theta_\ell,\mathcal{D}^i_n)\right\}} \cdot \mathbb{1}_{\{Y^j \leqslant Y^p\}}}{N^o_n(x_i;\Theta_\ell,\mathcal{D}^i_n)} \geqslant \alpha \Big\} .$$

The values from the k randomized trees are then aggregated to obtain the following random forest estimate

$$\widehat{q}^{o,\alpha}(Y|X_i = x_i) = \frac{1}{k}\sum_{\ell=1}^{k} \widehat{q}^{o,\alpha}_{\ell}(Y|X_i = x_i) .$$

Thus, this method allows us to propose the following estimator $\widehat{R}^{2,o}_i$ of $\mathbb{E}[\psi_\alpha(Y, q^{\alpha}(Y|X_i))]$ using the original sample.

4.2 Minimum-based O term estimators

The estimators developped in Subsection 4.1, based on $\mathbb{E}[\psi_\alpha(Y, q^{\alpha}(Y|X_i))]$, require to first approximate the conditional quantile and then plug it to estimate the O term. As mentioned before, a run of the model f could be time-consuming. Therefore, they may be inappropriate as two training samples are necessary. Hence, we propose in this part to develop estimators of the O term taking advantage from the expression $\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}[\psi_\alpha(Y,\theta)|X_i]\right]$ for which we only need to find the minimum instead of plugging the quantile.

4.2.1 Minimum estimation with a weighted approach

First of all, a random forest is built with the observations $\mathcal{D}^i_n$. Then, by considering an additional sample $(\mathbf{X}^{\diamond j})_{j=1,\dots,n}$ independent of $\mathcal{D}_n$, the O term may be estimated as follows

$$\widehat{Q}^{1,o}_i = \frac{1}{n}\sum_{m=1}^{n} \min_{p=1,\dots,n} \sum_{j=1}^{n} w^{o}_{n,j}(X^{\diamond m}_i)\, \psi_\alpha\left(Y^j, Y^p\right) .$$

Let us notice that the conditional expectation $\mathbb{E}[\psi_\alpha(Y,\theta)|X_i = x_i]$ is estimated with $\sum_{j=1}^{n} w^{o}_{n,j}(x_i)\,\psi_\alpha(Y^j,\theta)$ whose minimum is reached for θ equals one of the elements of $(Y^j)_{j=1,\dots,n}$.

4.2.2 Minimum estimation within a leaf

In this subsection, we are going to take advantage of the tree structure in order to propose a new estimator. To begin with, let us consider that a random forest is built with the observations $\mathcal{D}_n^i$.

Then, the key point is that an additional sample is no longer required in order to process the outer expectation of the O term. Indeed, for the ℓ-th tree, the observations falling into its m-th leaf node denoted by $A_n\left(m;\Theta_\ell,\mathcal{D}_n^i\right)$ approximate the conditional distribution of Y given a certain point $X_i=x_i$, which allows to estimate the minimum of the conditional expectation $\min_{\theta\in\mathbb{R}}\mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\,X_i=x_i\right]$. Then, we make the average over all the leaves of the ℓ-th tree to deal with the outer expectation. Hence, let $N_n^o(m;\Theta_\ell,\mathcal{D}_n^i)$ be the number of observations of the original sample $\mathcal{D}_n^i$ falling into the m-th leaf node and N_{leaves}^ℓ be the number of leaves in the ℓ-th tree. We define the following tree estimator for the O term

$$\frac{1}{N_{leaves}^\ell}\sum_{m=1}^{N_{leaves}^\ell}\left(\min\left\{p=1,\ldots,n,\ \left(X_i^p,Y^p\right)\in\mathcal{D}_n^i \text{ and } X_i^p\in A_n\left(m;\Theta_\ell,\mathcal{D}_n^i\right)\right\}\right.$$
$$\left.\sum_{j=1}^{n}\frac{\psi_\alpha\left(Y^j,Y^p\right)\cdot\mathbb{1}_{\left\{\left(X_i^j,Y^j\right)\in\mathcal{D}_n^i,\ X_i^j\in A_n(m;\Theta_\ell,\mathcal{D}_n^i)\right\}}}{N_n^o(m;\Theta_\ell,\mathcal{D}_n^i)}\right).$$

The approximations of the k randomized trees are then averaged to obtain the following random forest estimate

$$\widehat{Q}_i^{2,o}=\frac{1}{k}\sum_{\ell=1}^{k}\left[\frac{1}{N_{leaves}^\ell}\sum_{m=1}^{N_{leaves}^\ell}\left(\min\left\{p=1,\ldots,n,\ \left(X_i^p,Y^p\right)\in\mathcal{D}_n^i \text{ and } X_i^p\in A_n\left(m;\Theta_\ell,\mathcal{D}_n^i\right)\right\}\right.\right.$$
$$\left.\left.\sum_{j=1}^{n}\frac{\psi_\alpha\left(Y^j,Y^p\right)\cdot\mathbb{1}_{\left\{\left(X_i^j,Y^j\right)\in\mathcal{D}_n^i,\ X_i^j\in A_n(m;\Theta_\ell,\mathcal{D}_n^i)\right\}}}{N_n^o(m;\Theta_\ell,\mathcal{D}_n^i)}\right)\right].$$

It should be noted that looking for the minimum in the leaves directly implies that they are sufficiently sampled for the method to be valid.

4.2.3 Minimum estimation with a weighted approach and complete trees

In Subsections 4.2.1 and 4.2.2, the conditional distribution of Y given X_i is obtained from trees grown with $\mathcal{D}_n^i$. Instead of using this approach, it is proposed in this part to build a forest with complete trees, i.e. grown with all the model's inputs and then adjust the weights to recover the conditional expectation $\mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\,X_i\right]$.

Thus, as noticed, a full random forest is constructed with the whole dataset $\mathcal{D}_n$. Then, by using an additional sample $\left(\mathbf{X}^{\diamond j}\right)_{j=1,\ldots,n}$ independent of $\mathcal{D}_n$, the conditional expectation $\mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\,X_i=x_i\right]$ is estimated as follows

$$\begin{aligned}\mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\,X_i=x_i\right]&=\mathbb{E}\left[\mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\,X_1,\ldots,x_i,\ldots,X_d\right]|\,X_i=x_i\right]\\&\approx\frac{1}{n}\sum_{\ell=1}^{n}\left(\sum_{j=1}^{n}w_{n,j}^o\left(X_1^{\diamond\ell},\ldots,X_{i-1}^{\diamond\ell},x_i,X_{i+1}^{\diamond\ell},\ldots,X_d^{\diamond\ell}\right)\psi_\alpha\left(Y^j,\theta\right)\right)\\&\approx\sum_{j=1}^{n}w_{n,j}^{o,i}\left(x_i\right)\psi_\alpha\left(Y^j,\theta\right),\end{aligned}$$

where the suitable weights $w_{n,j}^{b,i}(x_i)$ are defined by

$$w_{n,j}^{o,i}(x_i) = \frac{1}{n}\sum_{\ell=1}^{n} w_{n,j}^{o}\left(\mathbf{X}_{-i}^{\diamond\ell}, x_i\right) . \tag{4.1}$$

The notation $\mathbf{X}_{-i}$ indicates the set of all variables except X_i and we note that the conditional expectation given $X_i = x_i$ is recovered by averaging over the components $\mathbf{X}_{-i}$. Thus, having independent inputs is required for this estimator compared to the previous ones and very convenient. Otherwise, it would be necessary to know the dependency structure in order to generate the observations $\left(\mathbf{X}_{-i}^{\diamond l}\right)_{l=1,\dots,n}$ for each new point $X_i = x_i$, which would make this estimator very cumbersome.

In addition to being used to recover the conditional expectation given $X_i = x_i$, the sample $\left(\mathbf{X}^{\diamond j}\right)_{j=1,\dots,n}$ is also used to estimate the outer expectation and we finally obtain the following estimator for the O term

$$\widehat{Q}_i^{3,o} = \frac{1}{n}\sum_{m=1}^{n} \min_{p=1,\dots,n} \sum_{j=1}^{n} w_{n,j}^{o,i}\left(X_i^{\diamond m}\right) \psi_\alpha\left(Y^j, Y^p\right) .$$

For the sake of clarity, only the version of the estimators making use of the original dataset has been presented within this section. Their analogous version based on the bootstrap samples is postponed in Appendix A.1, which leads to ten estimators of the O term. Nonetheless, numerical performances of both versions will be assessed in the following. Estimators based on bootstrap samples will be denoted in the same way as those using original dataset except the superscript o will be replaced by b, for instance, $\widehat{Q}_i^{3,b}$ instead of $\widehat{Q}_i^{3,o}$.

5 Overall estimation procedure

After defining the respective estimators for each term of the first-order QOSA index in Sections 2 and 4, the overall estimators are set in the following. In order to improve their accuracy, different strategies are also presented to tune hyperparameters of the random forest.

5.1 Issues with the leaf size

When using a random forest method for a regression task, a prediction is generally obtained by using the default values, proposed in the packages, for the $max\_features$ and $min\_samples\_leaf$ hyperparameters. There are some empirical studies on the impact of these hyperparameters such as Díaz-Uriarte and De Andres (2006); Scornet (2017); Duroux and Scornet (2018) but no theoretical guarantee to support the default values.

Concerning the estimation methods of the O term of the QOSA index proposed in Section 4, except for $\widehat{Q}_i^{3,b}$ and $\widehat{Q}_i^{3,o}$, it turns out that the values of the hyperparameters must be chosen carefully.
First of all, as a forest explaining Y by X_i is built for each model's input, the $max\_features$ hyperparameter has no impact in our procedures because it equals 1. Regarding the $min\_samples\_leaf$ hyperparameter, its impact on the quality of the estimators is investigated through the following toy example

$$Y = X_1 - X_2 , \tag{5.1}$$

with $X_1, X_2 \sim \mathcal{E}(1)$, where $\mathcal{E}(\lambda)$, $\lambda > 0$ stands for the exponential distribution with parameter λ. This standard example is commonly used in Sensitivity Analysis literature to assess the quality

of QOSA index estimators such as in Fort et al. (2016); Browne et al. (2017); Maume-Deschamps and Niang (2018). It should be noted that the estimation procedures presented in this article are valid for a model with error (i.e. such that $Y = f(\mathbf{X}) + \varepsilon$). However, the theoretical values are not explicitly calculable for such a model. Therefore, the performance of the various estimators is assessed on a model of the form $Y = f(\mathbf{X})$. Nevertheless, one simulation is performed for the noisy model $Y = X_1 - X_2 + \varepsilon$. In this case, the theoretical QOSA indices are computed with a Monte-Carlo method on a large sample (size 10^7), see Table 2 in Section 6.

In order to illustrate the influence of the hyperparameter *min_samples_leaf*, the boxplot of $\widehat{R}_1^{1,o}$ made with 100 values for different leaf sizes is presented in Figure 1. For each value of *min_samples_leaf*, an estimation $\widehat{R}_1^{1,o}$ is computed using two samples of size $n = 10^4$ and a forest grown with $n_{trees} = 500$. Then, the boxplots are compared with the analytical value given below and represented with the dotted orange line on each graph in Figure 1:

$$\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y|\, X_1\right)\right)\right] = -\alpha \log\left(\alpha\right) \ .$$

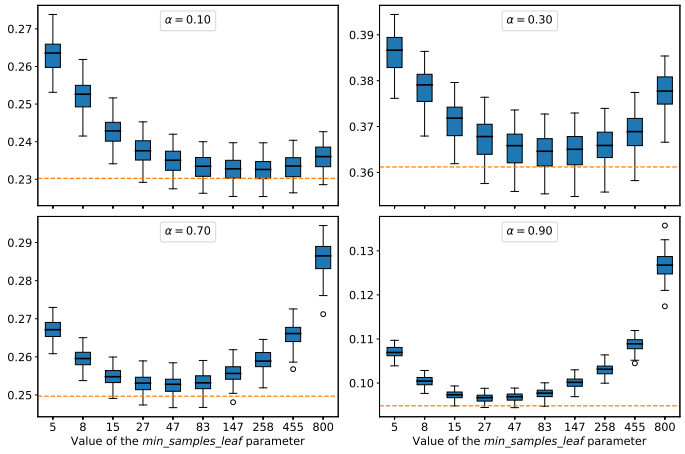

Figure 1: For several levels α: distribution of $\widehat{R}_1^{1,o}$, the estimation of the O term associated to the variable X_1 for different leaf sizes. The dotted orange line represents the true value on each plot.

Based on the results obtained in Figure 1, we see that for each level α, the performance of $\widehat{R}_1^{1,o}$ depends highly on the choice of the *min_samples_leaf* hyperparameter. Indeed, with the grid proposed for the values of *min_samples_leaf*, the optimum value seems to be 258 for $\alpha = 0.1$, 83 for $\alpha = 0.3$, 47 for $\alpha = 0.7$ and 27 for $\alpha = 0.9$.
This issue about the leaf size is only highlighted for $\widehat{R}_i^{1,o}$ but is also encountered for both methods, stated in Subsection 4.1, computing the conditional quantile with either the bootstrap samples or the original sample.

By using the same setting as in Figure 1, the distribution of $\widehat{Q}_1^{1,o}$ is presented in Figure 2 in order to assess the impact of the *min_samples_leaf* hyperparameter for a method where the minimum is estimated instead of plugging the quantile. The quality of $\widehat{Q}_1^{1,o}$ also seems to depend on the leaf size and the optimum value, allowing to well estimate $\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\, X_1\right]\right]$ for each level α, is the same as in Figure 1.
As before, this concern about the leaf size was only emphasized for $\widehat{Q}_i^{1,o}$ but is also encountered for both methods, detailed in Subsections 4.2.1 and 4.2.2, approximating the minimum with either the bootstrap samples or the original sample.

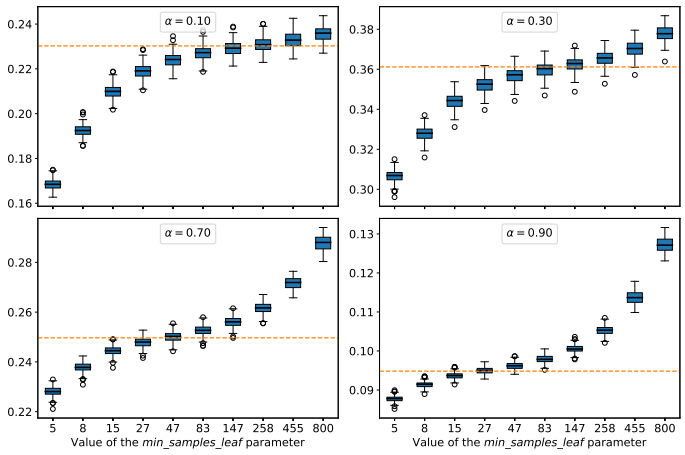

Figure 2: For several levels α: distribution of $\widehat{Q}_1^{1,o}$, the estimation of the O term associated to the variable X_1 for different leaf sizes. The dotted orange line represents the true value on each plot.

For the methods $\widehat{Q}_i^{3,b}$ and $\widehat{Q}_i^{3,o}$, based on complete trees, it seems that the tuning of the leaf size is less important as observed in Figure 3. Indeed, whatever the α level, the best results are observed for almost fully developed trees.

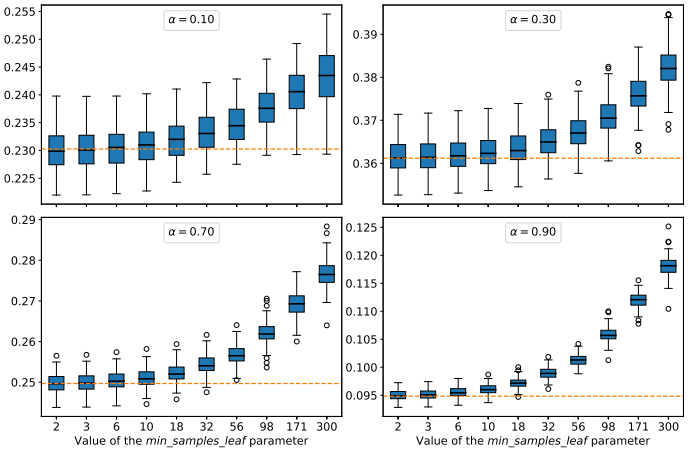

Figure 3: For several levels α: distribution of $\widehat{Q}_1^{3,o}$, the estimation of the O term associated to the variable X_1 for different leaf sizes. The dotted orange line represents the true value on each plot.

Thus, for all other estimators of the O term proposed in Section 4, a method giving us the optimal value of the leaf size for each level α is required to properly estimate the first-order QOSA index.

5.2 Tuning the leaf size

In order to tune the leaf size of our estimators, two methods are presented in this part. They lead to significatively improve the efficiency of the estimation. The first one rests on a classical cross-validation procedure and the second one uses the Out-Of-Bag samples.

5.2.1 Cross-validation procedure

The estimators of the O term developed in Subsection 4.1 are part of the conditional quantile estimation problem. Indeed, in a regression scheme, the conditional mean minimizes the expected squared error loss, while the conditional quantile $q^{\alpha}\left(Y|\, X_i = x_i\right)$ minimizes the following expected loss

$$q^{\alpha}\left(Y|\, X_i\right) = \underset{h:\mathbb{R}\to\mathbb{R}}{\arg\min}\, \mathbb{E}\left[\psi_{\alpha}\left(Y, h\left(X_i\right)\right)\right] \ .$$

Thus, estimators of $\mathbb{E}\left[\psi_{\alpha}\left(Y, q^{\alpha}\left(Y|\, X_i\right)\right)\right]$ established in Subsection 4.1 allow to assess the quality of the approximation of the true conditional quantile function. The smaller they are, the better the estimate of the conditional quantile function is. That is verified in Figure 1 and explains why we have this convex shape depending on the leaf size. As a matter of fact, when the value of the *min_samples_leaf* hyperparameter is incorrectly chosen, the approximation of the true conditional quantile function is wrong and so, this of $\mathbb{E}\left[\psi_{\alpha}\left(Y, q^{\alpha}\left(Y|\, X_i\right)\right)\right]$ too.

Hence, in order to estimate well the conditional quantile function $q^{\alpha}\left(Y|\, X_i\right)$ and therefore, $\mathbb{E}\left[\psi_{\alpha}\left(Y, q^{\alpha}\left(Y|\, X_i\right)\right)\right]$ (which is our goal), the optimum value of the leaf size will be chosen within a predefined grid containing potential values as being the one minimizing the empirical generalization error computed with a K-fold cross-validation procedure. A detailed description of this process is given in Algorithm 1 with $\widehat{R}_i^{1,o}$ for instance. The principle is the same for all estimators defined in Subsection 4.1.

It has to be noted that the number of folds K should be chosen carefully. Indeed, a lower value of K results in a more biased estimation of the generalization error, and hence undesirable. In contrast, a larger value of K is less biased, but can suffer from large variability. The choice of K is usually 5 or 10, but there is no formal rule.

Regarding the minimum-based estimators, using a similar approach with a K-fold cross-validation procedure is unsuitable due to the behavior of these ones depending on the leaf size (cf. Figure 2). Consequently, we propose to get the optimal value with one of the estimators plugging the quantile, in conjunction with the cross-validation process detailed in Algorithm 1. Once done, the estimator based on the minimum is computed with the optimal value obtained.

5.2.2 Out-Of-Bag quantile error

The estimators detailed in Subsection 4.1 deserve special attention. Indeed, another less cumbersome approach than cross-validation can be used to tune the leaf size. It is based on an adaptation to our context of the widespread "Out-Of-Bag" (OOB) error (Breiman, 1996) in regression and classification to estimate the generalization error.

• We first adapt the calculation of the OOB error for the conditional quantiles estimated with local averaging estimate of the C_CDF proposed in Subsection 4.1.1. For this purpose, we start by defining the OOB quantile error for $\widehat{R}_i^{1,o}$.
Let us fix an observation (X_i^m, Y^m) from $\mathcal{D}_n^i$ and consider $\mathcal{I}^m$ as the set of trees built with the bootstrap samples not containing this observation, i.e. for which this one is "Out-Of-Bag". The conditional quantile given that $X_i = X_i^m$ is estimated through $F_{k,n}^{o}\left(y|\, X_i = x_i\right) = \sum_{\substack{j=1 \\ j\neq m}}^{n} w_{n,j}^{o}\left(x_i\right) \mathbb{1}_{\{Y^j \leqslant y\}}$ where the weights are tailored to our context as follows

$$w_{n,j}^{o}\left(x_i; \Theta_1, \ldots, \Theta_{|\mathcal{I}^m|}, \mathcal{D}_n^i\right) = \frac{1}{|\mathcal{I}^m|} \sum_{\ell \in \mathcal{I}^m} \frac{\mathbb{1}_{\left\{X_i^j \in A_n\left(x_i; \Theta_\ell, \mathcal{D}_n^i\right)\right\}}}{N_n^o\left(x_i; \Theta_\ell, \mathcal{D}_n^i\right) - 1}, \ j = 1, \ldots, n, j \neq m \ .$$

Algorithm 1: K-fold cross-validation procedure explained with $\widehat{R}_i^{1,o}$

Input:

- Datasets: $\mathcal{D}_n^{\diamond i} = \left(X_i^{\diamond j}, Y^{\diamond j}\right)_{j=1,\ldots,n}$ from $\mathcal{D}_n^{\diamond}$ and $\mathcal{D}_n^i = \left(X_i^j, Y^j\right)_{j=1,\ldots,n}$ from $\mathcal{D}_n$
- Number of trees: $k \in \mathbb{N} \setminus \{0\}$
- The order where estimating $\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y|\, X_i\right)\right)\right] : \alpha \in]0,1[$
- Grid where looking for the best parameter: *grid_min_samples_leaf*
- Number of folds: $K \in \{2, \ldots, n\}$

Output: Estimated value of $\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y|\, X_i\right)\right)\right]$ at the α-level with $\widehat{R}_i^{1,o}$

begin Cross-validation procedure
  Randomly split the dataset $\mathcal{D}_n^i$ into K folds.
  foreach $\ell \in$ *grid_min_samples_leaf* **do**
    foreach *fold* **do**
      Take the current fold as a test set.
      Take the remaining groups as a training set.
      Fit a random forest model on the training set with the current ℓ as *min_samples_leaf* hyperparameter.
      Evaluate the conditional quantiles at the observations X_i in the test dataset and then compute $\widehat{R}_i^{1,o}$ on the test set.
      Retain the estimation obtained.
    end
    Summarize the quality related to the current ℓ by averaging the K estimated values and save the mean.
  end
end
Select as optimal value ℓ_{opt} for the *min_samples_leaf* hyperparameter, this one with the smallest mean.
Fit a random forest model on the complete dataset $\mathcal{D}_n^i$ by fixing the *min_samples_leaf* hyperparameter to ℓ_{opt}.
Compute $\widehat{R}_i^{1,o}$ with $\mathcal{D}_n^{\diamond i}$.

Then, $q^\alpha\left(Y|\, X_i = X_i^m\right)$ is estimated by plugging $F_{k,n}^o\left(y|\, X_i = X_i^m\right)$ instead of $F\left(y|\, X_i = X_i^m\right)$

$$\widehat{q}_{oob}^{o,\alpha}\left(Y|\, X_i = X_i^m\right) = \inf_{\substack{p=1,\ldots,n \\ p \neq m}} \left\{Y^p : F_{k,n}^o\left(Y^p|\, X_i = X_i^m\right) \geqslant \alpha\right\} \ .$$

After this operation is carried out for all data in $\mathcal{D}_n^i$, we calculate the error related to the approximation of the true conditional quantile function, i.e. the empirical generalization error

$$\widehat{OOB_i}^o = \frac{1}{n}\sum_{m=1}^{n} \psi_\alpha\left(Y^m, \widehat{q}_{oob}^{o,\alpha}\left(Y|\, X_i = X_i^m\right)\right) \ .$$

We may use the bootstrap samples (rather than the original one) in the definition of the weights:

$$w_{n,j}^b\left(x_i; \Theta_1, \ldots, \Theta_{|\mathcal{I}^m|}, \mathcal{D}_n^i\right) = \frac{1}{|\mathcal{I}^m|}\sum_{\ell \in \mathcal{I}^m} \frac{B_j\left(\Theta_\ell^1, \mathcal{D}_n^i\right) \mathbb{1}_{\left\{X_i^j \in A_n\left(x_i; \Theta_\ell, \mathcal{D}_n^i\right)\right\}}}{N_n^b\left(x_i; \Theta_\ell, \mathcal{D}_n^i\right)}, \ j = 1, \ldots, n \ .$$

This leads to define $F^b_{k,n}$ as follows

$$F^b_{k,n}\left(y|\, X_i = x_i\right) = \sum_{j=1}^{n} w^b_{n,j}\left(x_i\right) \mathbb{1}_{\{Y^j \leqslant y\}} \ .$$

The estimations of the conditional quantile $\widehat{q}^{b,\alpha}_{oob}$ and of the OOB quantile error $\widehat{OOB}^b_i$ follow.

• Secondly, we adapt the calculation of the OOB error for conditional quantiles estimated directly in tree leaves as introduced in Subsection 4.1.2. In that sense, define OOB quantile error for $\widehat{R}^{2,o}_i$.
Let us fix an observation (X^m_i, Y^m) from $\mathcal{D}^i_n$ and consider the set of trees built with the bootstrap samples not containing this observation. We then aggregate only the predictions of these trees to make our prediction $\widehat{q}^{o,\alpha}_{oob}\left(Y|\, X_i = X^m_i\right)$ of $q^\alpha\left(Y|\, X_i = X^m_i\right)$. After this operation carried out for all the data in $\mathcal{D}^i_n$, we calculate the error related to the approximation of the true conditional quantile function, i.e. the empirical generalization error

$$\widehat{OOB}^o_i = \frac{1}{n}\sum_{m=1}^{n} \psi_\alpha\left(Y^m, \widehat{q}^{o,\alpha}_{oob}\left(Y|\, X_i = X^m_i\right)\right) \ .$$

Again, using the bootstrap samples instead of the original one lead to define $\widehat{OOB}^b_i$.

The advantage of these methods, compared to cross-validation techniques, is that they do not require cutting out the training sample $\mathcal{D}^i_n$ and take place during the forest construction process.

Thus, given the dataset $\mathcal{D}^i_n$ and a grid containing potential values of the *min_samples_leaf* hyperparameter, a random forest is built for each one and the OOB quantile error associated is computed. Then, the optimal hyperparameter is chosen as the one with the smallest OOB error.

5.3 Full estimation procedure

Now, we have all the components in order to set the estimators of the first-order QOSA index S^α_i. These are separated in two classes according to the estimation method adopted for the O term. First of all, with the methods plugging the quantile, we define

$$\widehat{S}^\alpha_i = 1 - \frac{\widehat{R}_i}{\widehat{P}_1} \text{ with } \widehat{R}_i \in \left\{\widehat{R}^{1,b}_i, \widehat{R}^{1,o}_i, \widehat{R}^{2,b}_i, \widehat{R}^{2,o}_i\right\} \ .$$

The whole procedure integrating the cross-validation process for these methods is detailed in Algorithm 2 (see Appendix A.2).

On the other hand, regarding the methods based on the minimum to compute the O term, we set

$$\widehat{S}^\alpha_i = 1 - \frac{\widehat{Q}_i}{\widehat{P}_1} \text{ with } \widehat{Q}_i \in \left\{\widehat{Q}^{1,b}_i, \widehat{Q}^{1,o}_i, \widehat{Q}^{2,b}_i, \widehat{Q}^{2,o}_i, \widehat{Q}^{3,b}_i, \widehat{Q}^{3,o}_i\right\} \ .$$

The estimation process based on the minimum is formalized in Algorithms 3, 4 and 5. For the sake of clarity, they are all gathered in Appendix A.2 . Algorithm 3 (resp. 5) estimating the QOSA index with $\widehat{Q}^{1,b}_i$ or $\widehat{Q}^{1,o}_i$ (resp. $\widehat{Q}^{3,b}_i$ or $\widehat{Q}^{3,o}_i$), needs a full training sample $\mathcal{D}_n$ as well as a partial one $\left(\mathbf{X}^{\diamond j}\right)_{j=1,\dots,n}$. While estimating the QOSA index with $\widehat{Q}^{2,b}_i$ or $\widehat{Q}^{2,o}_i$ only requires one training sample $\mathcal{D}_n$. This is a major advantage over methods plugging the quantile that need two full training samples.

So far, no consistency result has been proved for $\widehat{S}_i^{\alpha}$. These various estimators are reviewed in the next section in order to establish their efficiency in practice. Moreover, all these algorithms are implemented within a `python` package named `qosa-indices` available at Elie-Dit-Cosaque (2020), it can be also freely downloaded on the PyPI website.

6 Numerical illustrations

Let us now carry out some simulations in order to investigate the influence of the hyperparameter optimization algorithm, the impact of the number of trees on our estimators and compare the decrease of the estimation error of each one in function of the train sample-size. From these results, the performance of the two best estimators as well as those based on kernel methods defined in Browne et al. (2017); Maume-Deschamps and Niang (2018) is assessed. Then, their scalability is tested on a toy example.

6.1 Comparison of hyperparameter optimization algorithms

We start by studying the influence of the hyperparameter optimization algorithm on the performance of our estimators plugging the conditional quantile (i.e. using $\widehat{R}_i^{1,b}, \widehat{R}_i^{1,o}, \widehat{R}_i^{2,b}$ or $\widehat{R}_i^{2,o}$). This survey is carried out with the model introduced in Equation (5.1) and the following setting.

The estimators of the QOSA index are computed with samples of size $n = 10^4$ (i.e. $2n$ runs of the model for estimators using two training samples). The leaf size is tuned for each estimator over a grid with 20 numbers evenly spaced ranging from 5 to 300 by using either the strategy based on the OOB quantile error developed in Subsection 5.2.2 or a 3-fold cross-validation procedure. Then, to assess the efficiency of each method (CV vs OOB), the experiment is repeated $s = 100$ times and the following metrics are computed

$$\begin{aligned} RMSE_i^{\alpha} &= \sqrt{\frac{1}{s}\sum_{j=1}^{s}\left(\widehat{S}_i^{\alpha,j} - S_i^{\alpha}\right)^2}, \\ Bias_i^{\alpha} &= \left|\frac{1}{s}\sum_{j=1}^{s}\widehat{S}_i^{\alpha,j} - S_i^{\alpha}\right|, \\ Variance_i^{\alpha} &= \frac{1}{s}\sum_{j=1}^{s}\left(\widehat{S}_i^{\alpha,j} - \frac{1}{s}\sum_{j=1}^{s}\widehat{S}_i^{\alpha,j}\right)^2, \end{aligned} \tag{6.1}$$

with S_i^{α}, the analytical values that were provided in Fort et al. (2016).

In Figure 4, for three levels α, we present the evolution of the different metrics related to the variable X_1 of our toy example in function of the number of trees ranging from 1 to 200 (in log scale). More precisely, sub-figures at the top of Figure 4 show the Root Mean Square Error (RMSE), in the middle, the bias and the variance at the bottom.

We observe that regardless of the level α and the number of trees, our estimators plugging the quantile have globally the same performance when calculated with either the OOB strategy or the cross-validation procedure. But, the run time is faster when using the OOB strategy rather than the cross-validation procedure.

6.2 Convergence with the number of trees and the train sample-size

We analyze in this part the impact of the number of trees on the performance of all our estimators except for those using $\widehat{Q}_i^{3,b}$ and $\widehat{Q}_i^{3,o}$ because of the computational cost. This survey is also carried

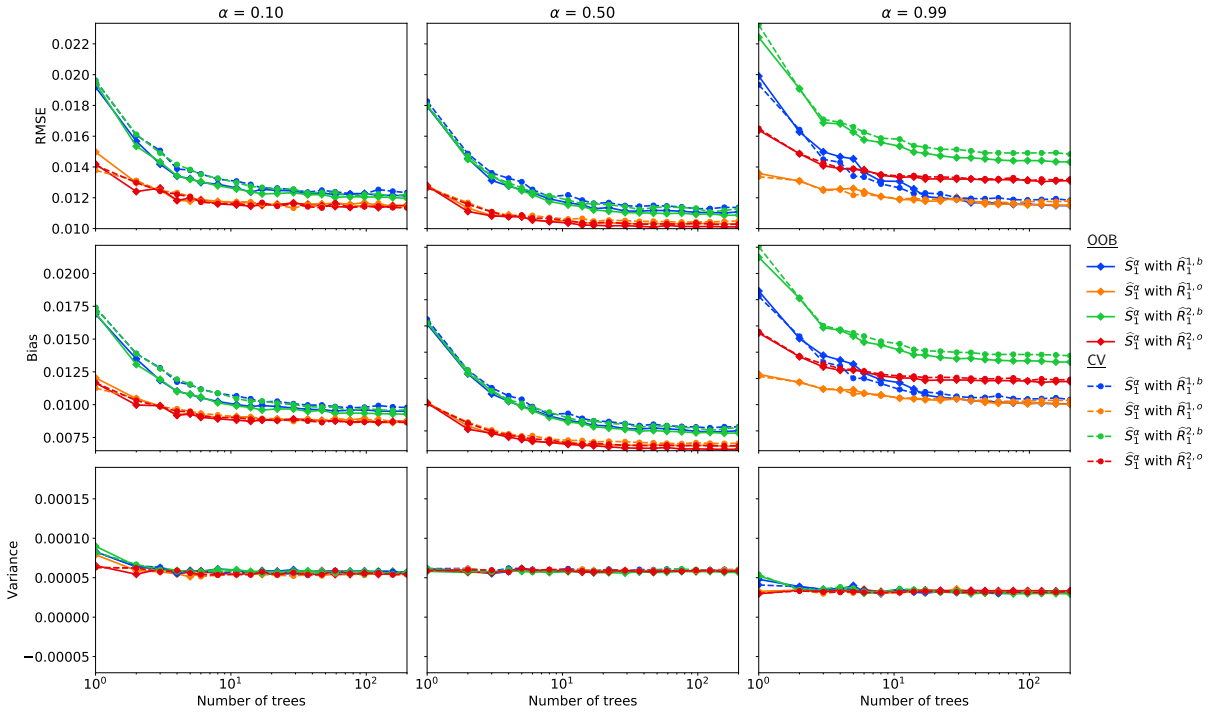

Figure 4: Evolution of RMSE, bias and variance of the estimators associated with X_1, calculated with either the OOB strategy or the Cross-Validation procedure, in function of the number of trees for three levels α.

out with the model introduced in Equation (5.1) and the following setting.

The estimators of the QOSA index are computed with samples of size $n = 10^4$. The leaf size is tuned over a grid with 20 numbers evenly spaced ranging from 5 to 300 by using a 3-fold cross-validation procedure for $\widehat{R}_i^{1,b}$ and $\widehat{R}_i^{1,o}$ while the strategy based on the OOB samples, developed in Subsection 5.2.2, is used for $\widehat{R}_i^{2,b}$ and $\widehat{R}_i^{2,o}$. Regarding the minimum based estimators, the optimal leaf size is obtained via $\widehat{R}_i^{1,o}$ during the 3-fold cross-validation process. Then, the efficiency of our estimators is assessed with the metrics introduced in Equation (6.1) by repeating the experiment $s = 200$.

In Figure 5, for three levels α, we present the evolution of the different metrics related to the variable X_1 of our toy example in function of the number of trees ranging from 1 to 200 (in log scale). More precisely, sub-figures at the top of Figure 5 show the Root Mean Square Error (RMSE), in the middle, the bias and the variance at the bottom.

We observe that regardless of the level α, RMSE of our estimators is small. The number of trees seems to have no impact for those using $\widehat{Q}_i^{1,o}$ and $\widehat{Q}_i^{2,o}$ as the RMSE value is almost always the same. RMSE of the others decreases in function of the number of trees until it reaches a threshold starting at about 50 trees. Indeed, it is well known that from a certain number, increasing the number of trees becomes useless but results in higher calculation costs. However, we did not expect to have a stable estimation error with so few trees.
Besides, still from the RMSE curves, it first appears that the estimators using the original sample (plain lines) have a lower error compared to those using the bootstrap samples (dotted lines). On the other hand, the performance of the minimum based estimators (green and red lines) seems better than those based on the quantile (blue and orange lines). That might be explained by the additional error due to the estimation of the conditional quantile.

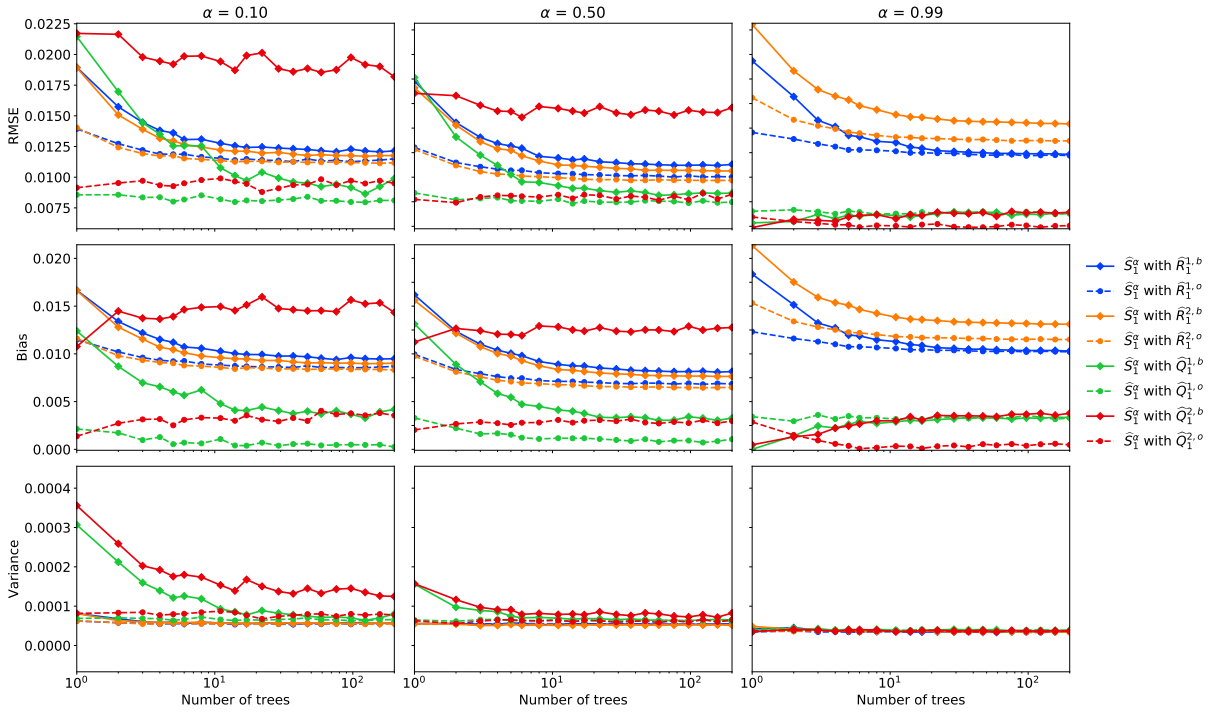

Figure 5: Evolution of RMSE, bias and variance of the estimators associated with X_1 in function of the number of trees for three levels α.

Variance of all estimators is close to 0 and the bias curves have the same behavior as RMSE curves. This means that bias is the main/only source of error in the RMSE. This bias could be reduced by taking a larger grid where looking for the optimal leaf size during the cross-validation or using another more efficient method to find the optimum. Also, on Figure 5, it seems that the Q-estimators have more bias when the bootstrap sample is reused in the estimation (instead of the original sample). This could be an over-fitting effect which is also observed with less amplitude for the bootstrap R-estimator.

Let us now compare the decrease of the estimation error in function of the train sample-size. As observed in Figure 5, take a very large number of trees is not required in order to have a stable estimation error. Thus, we take $n_{trees} = 100$ and the same setting as before for other parameters in the next study and observe the evolution of the metrics introduced in Equation (6.1) in function of the sample size.

Figure 6 presents RMSE, bias and variance of our estimators for different sample sizes. We observe that all the metrics associated with the various estimators converge to 0 at different rates. Indeed, the convergence rates of the metrics of the quantile-based estimators are slower than those based on the minimum.

Hence, from our experiments, it turns out that the minimum-based estimators give the best results. This is an interesting feature because they need less data than those plugging the quantile. Furthermore, few trees are necessary in order to reduce the estimation error. It therefore allows to get a good estimation of the indices with a reasonable computational cost.

6.3 Comparison with kernel methods

In this subsection, we compare on the toy example introduced in Equation (5.1):

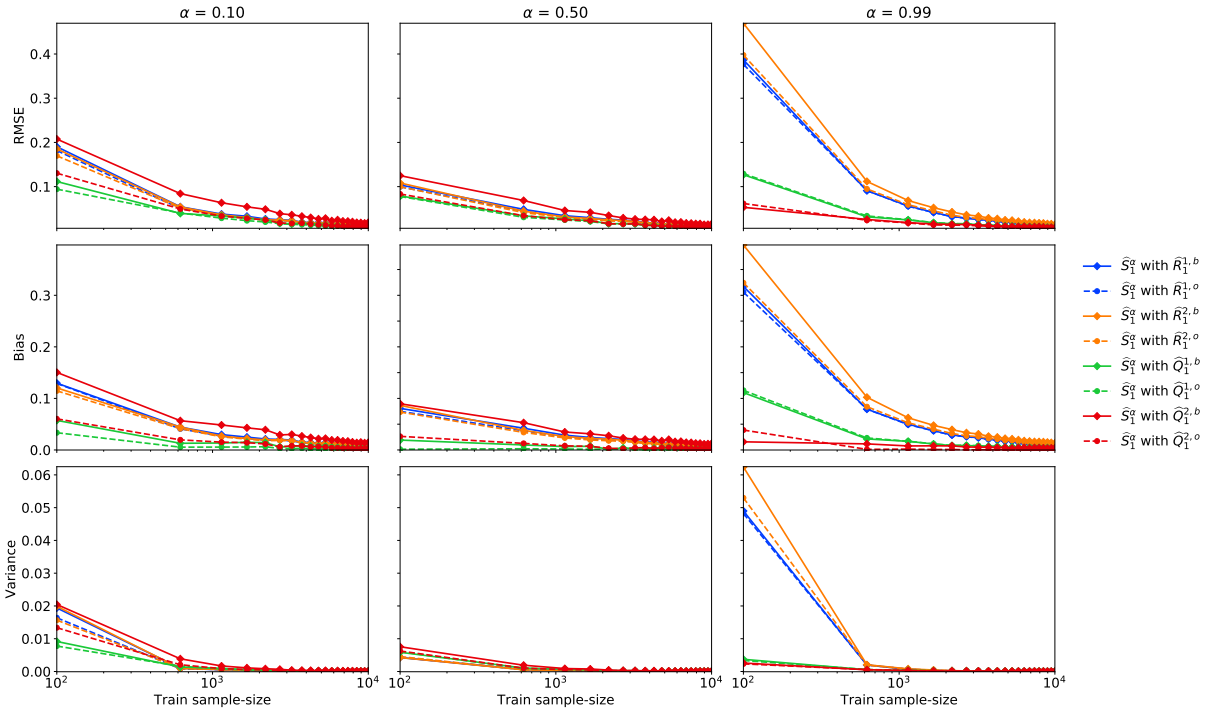

Figure 6: Evolution of RMSE, bias and variance of the estimators associated with X_1 in function of the train sample-size for three levels α.

- the kernel-based estimators proposed in Browne et al. (2017); Maume-Deschamps and Niang (2018) denoted by $\check{S}_i^\alpha$ and $\widetilde{S}_i^\alpha$,
- the minimum-based QOSA index estimators building one forest for each input and using the original sample,
- and the minimum-based QOSA index estimators using a forest grown with trees fully developed.

The estimators of the QOSA indices are computed with samples of size $n = 10^4$. Forest methods are grown with $n_{trees} = 100$. The optimal leaf size for the minimum-based estimators building one forest for each input is obtained with $\widehat{R}_i^{1,o}$ during the 3-fold cross-validation process over a grid containing 20 numbers evenly spaced ranging from 5 to 300. Regarding the minimum-based estimators using a forest grown with trees fully developed, the *min_samples_leaf* hyperparemeter equals 2.
In order to have comparable methods, a cross-validation procedure is also implemented for the kernel-based estimators to choose the optimal bandwidth parameter. It is selected within over a grid containing 20 potential values ranging from 0.001 to 1. Then, we assess the performance of the different estimators by computing their empirical root mean squared error with 100 experiments.

Table 1 contains the empirical root mean squared error of the different estimators associated to each input as well as the overall run time requested to obtain them. About their performance, it seems that the random forest-based estimators are better than the kernel methods. Nevertheless, as regards the methods using $\widehat{Q}_i^{3,b}$ and $\widehat{Q}_i^{3,o}$, while they have a low error and do not need to tune the leaf size, their run time with the current implementation is too long to be used in practice. Accordingly, we recommend to compute the indices with $\widehat{Q}_i^{1,o}$ and $\widehat{Q}_i^{2,o}$ in order to get

	$\widehat{S}_i^\alpha$ with $\widehat{Q}_i^{1,o}$		$\widehat{S}_i^\alpha$ with $\widehat{Q}_i^{2,o}$		$\widehat{S}_i^\alpha$ with $\widehat{Q}_i^{3,b}$		$\widehat{S}_i^\alpha$ with $\widehat{Q}_i^{3,o}$		$\check{S}_i^\alpha$		$\widetilde{S}_i^\alpha$	
	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
$\alpha = 0.1$	0.007	0.006	0.009	0.006	0.017	0.006	0.017	0.006	0.020	0.044	0.061	0.006
$\alpha = 0.25$	0.008	0.006	0.009	0.006	0.013	0.007	0.013	0.007	0.013	0.036	0.042	0.012
$\alpha = 0.5$	0.008	0.006	0.008	0.007	0.010	0.009	0.010	0.009	0.019	0.021	0.027	0.025
$\alpha = 0.75$	0.008	0.007	0.008	0.008	0.008	0.014	0.008	0.014	0.035	0.012	0.014	0.042
$\alpha = 0.99$	0.006	0.016	0.006	0.018	0.006	0.032	0.006	0.032	0.084	0.071	0.013	0.11
run time	1 hr		18 min 24 sec		10 hr 41 min		8 hr 18 min		1 hr 55 min		1 min 51 sec	

Table 1: RMSE and run time, for the toy example, of the random forest based estimators: $\widehat{S}_i^\alpha$ computed with $\widehat{Q}_i^{1,o}$, $\widehat{Q}_i^{2,o}$, $\widehat{Q}_i^{3,b}$ and $\widehat{Q}_i^{3,o}$ as well as those based on kernel: $\widetilde{S}_i^\alpha$ and $\check{S}_i^\alpha$.

good estimations of the first-order QOSA indices in a reasonable time.

For completeness, one simulation is also performed on a noisy model: $Y = X_1 - X_2 + \varepsilon$ with X_i following an exponential distribution as above and ε a centered normal distribution with standard deviation 0.5, independent of $\mathbf{X}$. In order to compute the RMSE, the QOSA indices are estimated with a Monte Carlo approach on a sample of size 10^7. The other parameters and hyperparameters are as above. Table 2 below contains the empirical root mean squared error of the estimators $\widehat{S}^\alpha$ computed with $\widehat{Q}^{1,o}, \widehat{Q}^{2,o}$ and $\widetilde{S}^\alpha$. We remark that the RMSE have the same order of magnitude for the noisy and the un-noisy models.

	$\widehat{S}_i^\alpha$ with $\widehat{Q}_i^{1,o}$		$\widehat{S}_i^\alpha$ with $\widehat{Q}_i^{2,o}$		$\widetilde{S}_i^\alpha$	
	X_1	X_2	X_1	X_2	X_1	X_2
$\alpha = 0.1$	0.008	0.008	0.008	0.009	0.021	0.041
$\alpha = 0.25$	0.007	0.008	0.008	0.008	0.013	0.029
$\alpha = 0.5$	0.007	0.007	0.008	0.008	0.015	0.017
$\alpha = 0.75$	0.008	0.008	0.008	0.009	0.021	0.012
$\alpha = 0.99$	0.015	0.017	0.02	0.024	0.096	0.067

Table 2: RMSE, for the toy example with noise, of the random forest based estimators: $\widehat{S}_i^\alpha$ computed with $\widehat{Q}_i^{1,o}, \widehat{Q}_i^{2,o}$ and the kernel based estimator $\widetilde{S}_i^\alpha$.

6.4 Scalability of the methods

The influence of the model's dimension d over the performance of the estimators using $\widehat{Q}_i^{1,o}$ and $\widehat{Q}_i^{2,o}$ is investigated in this subsection with the following additive exponential framework

$$Y = \sum_{i=1}^{d} X_i \ . \tag{6.2}$$

Independent inputs $X_i, i = 1, \ldots, d$, follow an Exponential distribution $\mathcal{E}(\lambda_i)$, with distinct λ_i. The resulting output Y is a generalized Erlang distribution also called Hypoexponential distribution. By taking advantage of the other expression of the first-order QOSA index given in Maume-Deschamps and Niang (2018), we obtain the following semi closed-form analytical formula

$$S_i^\alpha = 1 - \frac{\alpha \mathbb{E}\left[Xs_{(-i)}\right] - \mathbb{E}\left[Xs_{(-i)} \mathbb{1}_{\left\{Xs_{(-i)} \leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}}\right]}{\alpha \mathbb{E}\left[Y\right] - \mathbb{E}\left[Y \mathbb{1}_{\left\{Y \leqslant q^\alpha(Y)\right\}}\right]} \ , \tag{6.3}$$

with $Xs_{(-i)} = \sum_{j \neq i} X_j$ that also follows a Hypoexponential distribution. Knowing the cumulative distribution function of the Hypoexponential distribution, quantiles $q^{\alpha}(Y)$ and $q^{\alpha}\left(Xs_{(-i)}\right)$ are computed by numeric inversion and the analytical expression of the truncated expectations is derived from Marceau (2013).

For a specific dimension d, d values evenly spaced are selected from the interval $[0.3, 1.25]$ and then each one represents the λ_i parameter of an input X_i, $i = 1, \ldots, d$. QOSA index estimations are then computed with samples of size $n = 10^4$, a forest grown with $n_{trees} = 100$ and the setting defined hereafter. The leaf size is tuned with $\widehat{R}_i^{1,o}$ over a grid with 20 numbers evenly spaced ranging from 5 to 300 by using a 3-fold cross-validation. Each experiment is done 100 times in order to compute the RMSE defined in Equation (6.1) for each input, and then we take the weighted mean by the analytical values of the QOSA indices over all dimensions in order to get a global measure.

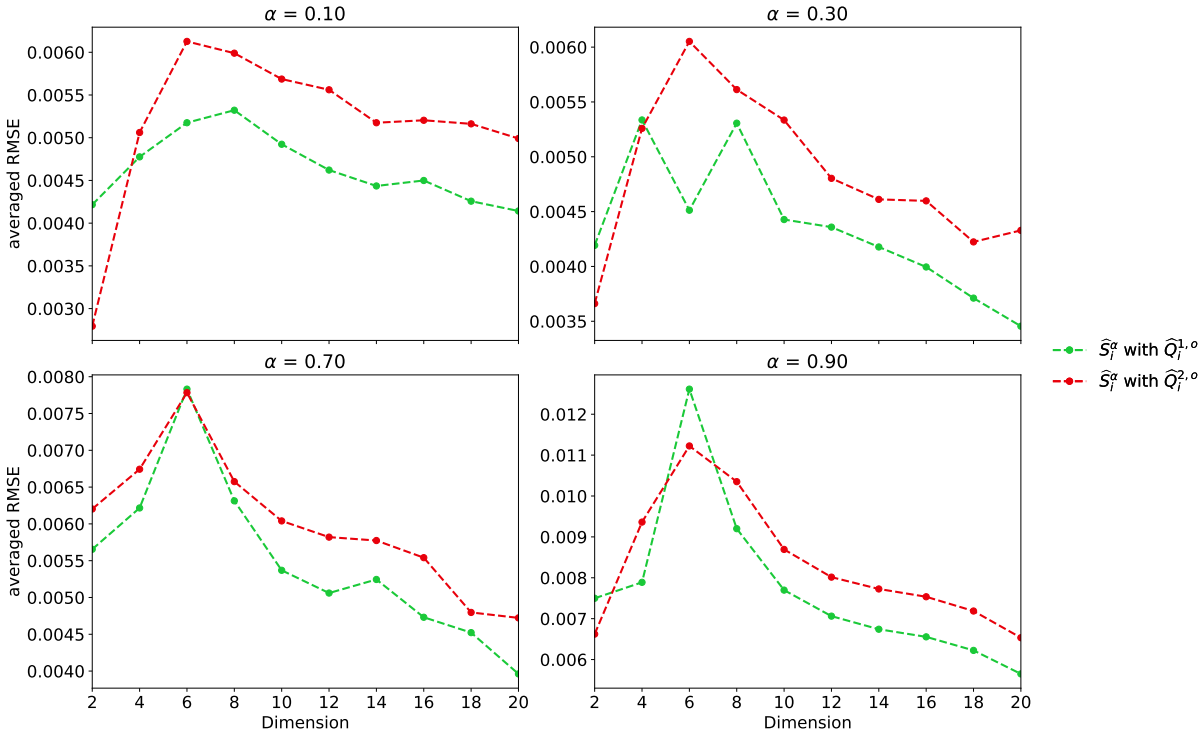

Figure 7: Evolution of the averaged RMSE over all dimensions of the estimators calculated with $\widehat{Q}_i^{1,o}$ and $\widehat{Q}_i^{2,o}$ in function of the model dimension for four levels α.

Figure 7 presents the weighted RMSE as a function of the increasing dimension of our model for several levels α. For each one, we observe that the error increases slowly at the beginning until the dimension 6 for both methods then decreases. This phenomenon is due to the chosen parametrization. Indeed, when increasing the dimension of the model, the respective impact of each input is reduced. Thus, from a certain dimension, all the analytical values of the first-order QOSA indices become small and even close to 0 for some inputs. Our estimators properly capture this trend as they decrease by increasing the dimension. However, the estimator using $\widehat{Q}_i^{1,o}$ seems better than this one employing $\widehat{Q}_2^{1,o}$ as its error is lower.

7 Practical case study

We propose to apply our methodology to a real dataset used in Besse et al. (2007) to improve the ozone concentration predicted by the fluid mechanics model named MOCAGE (Modèle de Chimie Atmosphérique à Grande Echelle)[1]. Indeed, predictions carried out by this model were biased. Besse et al. (2007) therefore proposed to correct this bias by building a statistical model between some input variables including the predicted ozone concentration by MOCAGE and the corresponding observed one. In our context, our goal will be to quantify the impact of each input on the α-quantile of the observed ozone concentration.

The "depSeuil.dat" dataset used for our study is available at `http://www.math.univ-toulouse.fr/~besse/Wikistat/data` and contains 10 variables with 1041 observations. We will consider that **O3obs**, observed ozone concentration, is explained by the 9 other variables described below.

JOUR: type of day (0 for holiday vs 1 for non holiday)	STATION: site of observations (5 different sites)
RMH2O: humidity ratio	NO2: nitrogen dioxide concentration
VentMOD: wind force	VentANG: wind direction
NO: nitric oxide concentration	TEMPE: officially predicted temperatures
MOCAGE: ozone concentration predicted by a fluid mechanics model	

The left-hand graph on Figure 8 below gives the QOSA estimations for several levels α, using the $\widehat{Q}_i^{2,o}$ estimator, since from our numerical study, it is the quicker, requires only one sample and is efficient. The inputs are ranked, for different values of alpha. On the right picture, the QOSA indices are in percentage (normalized by the sum of QOSA indices for all variables).

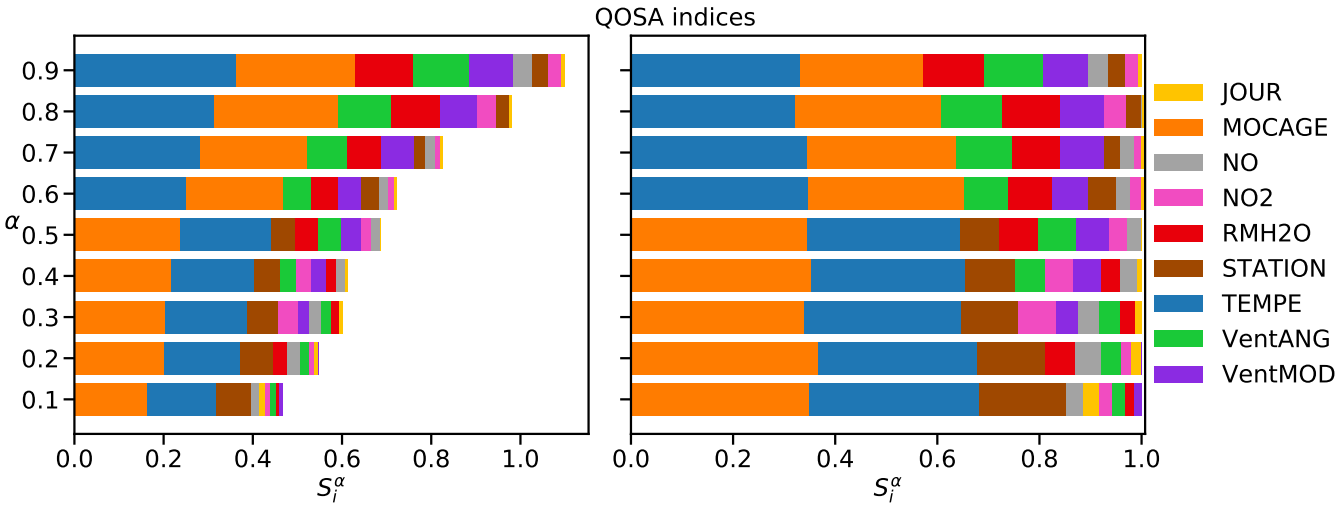

Figure 8: QOSA (resp. normalised QOSA) indices at different levels α on the left-hand (resp. on the right-hand) plot.

In Besse et al. (2007); Broto et al. (2020), the impact of the input variables on the expectation of the observed ozone concentration has been studied, through Shapley effects, and leads to consider MOCAGE and TEMPE as the most influencial variables followed by STATION and NO2. Our study gives consistent results with the previous one because MOCAGE and TEMPE

[1]Large Scale Atmospherical Chemestrial Model

are also the most important variables. But, we can note with QOSA indices that for quantile levels greater than 0.6, inputs related to wind and RMH2O are also important, more than STATION. This gives a relevant information to the practitioner and highlights the influence of these variables when there is a pollution peak.

8 Conclusion

In this paper, we introduced several estimators for the first-order QOSA index by using the random forest method. Some of them use the original sample while the others use the bootstrap samples generated during the forest construction. Both classes of estimator seem to be efficient even if we observe in our experiments that the methods using the original sample have a lower estimation error than those based on the bootstrap ones. Thus, supplementary studies should be conducted to inquire into this difference. Furthermore, the performance of these methods is highly dependent on the leaf size. This parameter could be compared to the bandwidth parameter of kernel estimators as it controls the bias of the method. But, it turns out to be easier to calibrate and we propose two methods to do this: K-fold cross-validation or Out-of-Bag samples based selection method.

It is also well known for random forest methods that the number of trees k should be chosen large enough to reach the desired statistical precision and small enough to make the calculations feasible as the computational cost increases linearly with k as mentioned in Scornet (2017). But, we have seen on our "toy example" that estimators proposed herein require few trees in order to have a low estimation error. This makes possible to estimate the indices correctly while maintaining a reasonable computation time.

Besides, we obtain in our application better results for our estimators when comparing with the kernel methods. A major advantage is that we have developed an estimator that requires only one training sample, whereas kernel methods require two training samples or a full one plus a partial. This feature is interesting when dealing with costly models. Another significant asset of our estimators is that their efficiency seems maintained when increasing the model dimension.

Despite these benefits, the proof for the estimators' consistency as well as the asymptotic analysis to establish the convergence rates and confidence intervals remains a major wish for the future. At last, it would be interesting to leverage these estimation methods in order to propose estimators of Quantile-Oriented Shapley Effects (QOSE) defined in Elie-Dit-Cosaque and Maume-Deschamps (2022a).

References

Antoniadis, A., Lambert-Lacroix, S., and Poggi, J.-M. (2021). Random forests for global sensitivity analysis: A selective review. *Reliability Engineering & System Safety*, 206.

Besse, P., Milhem, H., Mestre, O., Dufour, A., and Peuch, V.-H. (2007). Comparaison de techniques de «Data Mining» pour l'adaptation statistique des prévisions d'ozone du modèle de chimie-transport MOCAGE. *Pollution atmosphérique*, (195):285–292.

Breiman, L. (1996). Out-of-bag estimation.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees.

Broto, B., Bachoc, F., and Depecker, M. (2020). Variance reduction for estimation of Shapley effects and adaptation to unknown input distribution. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):693–716.

Browne, T., Fort, J.-C., Iooss, B., and Le Gratiet, L. (2017). Estimate of quantile-oriented sensitivity indices. Technical Report, hal-01450891.

Da Veiga, S., Gamboa, F., Iooss, B., and Prieur, C. (2021). *Basics and trends in sensitivity analysis: Theory and practice in R*.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3.

Duroux, R. and Scornet, E. (2018). Impact of subsampling and tree depth on random forests. *ESAIM: Probability and Statistics*, 22:96–128.

Elie-Dit-Cosaque, K. (2020). qosa-indices, a python package available at: `https://gitlab.com/qosa_index/qosa`.

Elie-Dit-Cosaque, K. and Maume-Deschamps, V. (2022a). Goal-oriented shapley effects with special attention to the quantile-oriented case. *SIAM/ASA Journal on Uncertainty Quantification*, 10(3):1037–1069.

Elie-Dit-Cosaque, K. and Maume-Deschamps, V. (2022b). Random forest estimation of conditional distribution functions and conditional quantiles. *Electronic Journal of Statistics*, 16(2):6553–6583.

Fort, J.-C., Klein, T., and Rachdi, N. (2016). New sensitivity analysis subordinated to a contrast. *Communications in Statistics-Theory and Methods*, 45(15):4349–4364.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *Annals of Mathematical Statistics*, 19(3):293–325.

Homma, T. and Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17.

Jansen, M. J., Rossing, W. A., and Daamen, R. A. (1994). Monte Carlo Estimation of Uncertainty Contributions from Several Independent Multivariate Sources. In *Predictability and Nonlinear Modelling in Natural Sciences and Economics*, pages 334–343. Springer.

Kala, Z. (2019). Quantile-oriented global sensitivity analysis of design resistance. *Journal of Civil Engineering and Management*, 25(4):297–305.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, 15(4):143–156.

Kucherenko, S., Song, S., and Wang, L. (2019). Quantile based global sensitivity measures. *Reliability Engineering & System Safety*, 185:35–48.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

Marceau, E. (2013). *Modélisation et évaltuation quantitative des risques en actuariat*. Springer Berlin.

Maume-Deschamps, V. and Niang, I. (2018). Estimation of quantile oriented sensitivity indices. *Statistics & Probability Letters*, 134:122–127.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons.

Scornet, E. (2016). Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500.

Scornet, E. (2017). Tuning parameters in random forests. *ESAIM: Proceedings and Surveys*, 60:144–162.

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414.

A Appendix

A.1 Bootstrap version of the estimators of the O term

A.1.1 Quantile estimation with a weighted approach

Another estimator of the C_CDF can be achieved by replacing the weights $w_{n,j}^{o}(x_i)$ based on the original dataset of the forest by those using the bootstrap samples $w_{n,j}^{b}(x_i)$ provided in Equation (3.3). That gives the following estimator which has been proposed in Elie-Dit-Cosaque and Maume-Deschamps (2022b),

$$F_{k,n}^{b}(y|\, X_i = x_i) = \sum_{j=1}^{n} w_{n,j}^{b}(x_i)\, \mathbb{1}_{\{Y^j \leqslant y\}} \ .$$

The conditional quantiles are then estimated by plugging $F_{k,n}^{b}(y|\, X_i = x_i)$ instead of $F(y|\, X_i = x_i)$. Accordingly, the associated estimator of $\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha(Y|\, X_i)\right)\right]$ based on these weights is denoted $\widehat{R}_i^{1,b}$.

A.1.2 Quantile estimation within a leaf

For the ℓ-th tree, the estimator $\widehat{q}_\ell^{\,b,\alpha}(Y|\, X_i = x_i)$ of $q^\alpha(Y|\, X_i = x_i)$ is obtained with the bootstrap observations falling into $A_n(x_i;\Theta_\ell,\mathcal{D}_n^i)$ as follows

$$\widehat{q}_\ell^{\,b,\alpha}(Y|\, X_i = x_i) = \inf_{p=1,\dots,n}\Bigg\{Y^p,\ (X_i^p,Y^p)\in\mathcal{D}_n^{i\star}(\Theta_\ell) \text{ and } X_i^p \in A_n(x_i;\Theta_\ell,\mathcal{D}_n^i) : \sum_{j=1}^{n}\frac{B_j\left(\Theta_\ell^1,\mathcal{D}_n^i\right)\cdot\mathbb{1}_{\left\{X_i^j\in A_n(x_i;\Theta_\ell,\mathcal{D}_n^i)\right\}}\cdot\mathbb{1}_{\{Y^j\leqslant Y^p\}}}{N_n^b(x_i;\Theta_\ell,\mathcal{D}_n^i)} \geqslant \alpha\Bigg\} \ .$$

That gives us the following random forest estimate of the conditional quantile

$$\widehat{q}^{\,b,\alpha}(Y|\, X_i = x_i) = \frac{1}{k}\sum_{\ell=1}^{k}\widehat{q}_\ell^{\,o,\alpha}(Y|\, X_i = x_i) \ .$$

Hence, we propose the estimator $\widehat{R}_i^{2,b}$ of $\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha(Y|\, X_i)\right)\right]$ using the bootstrap samples.

A.1.3 Minimum estimation with a weighted approach

Another estimator is obtained by replacing weigths $w_{n,j}^{o}(x_i)$ with the $w_{n,j}^{b}(x_i)$ version presented in Equation (3.3) using the bootstrap samples. The obtained estimator of the O term is denoted by $\widehat{Q}_i^{1,b}$.

A.1.4 Minimum estimation within a leaf

For the ℓ-th tree, let $N_n^b(m;\Theta_\ell,\mathcal{D}_n^i)$ be the number of observations of the bootstrap sample $\mathcal{D}_n^{i\star}(\Theta_\ell)$ falling into the m-th leaf node and N_{leaves}^{ℓ} be the number of leaves in the ℓ-th tree. We define the following tree estimator for the O term

$$\frac{1}{N_{leaves}^{\ell}}\sum_{m=1}^{N_{leaves}^{\ell}}\Bigg(\min\left\{p=1,\dots,n,\ (X_i^p,Y^p)\in\mathcal{D}_n^{i\star}(\Theta_\ell) \text{ and } X_i^p\in A_n\left(m;\Theta_\ell,\mathcal{D}_n^i\right)\right\} \sum_{j=1}^{n}\frac{B_j\left(\Theta_\ell^1,\mathcal{D}_n^i\right)\cdot\psi_\alpha\left(Y^j,Y^p\right)\cdot\mathbb{1}_{\left\{(X_i^j,Y^j)\in\mathcal{D}_n^{i\star}(\Theta_\ell),\ X_i^j\in A_n(m;\Theta_\ell,\mathcal{D}_n^i)\right\}}}{N_n^b(m;\Theta_\ell,\mathcal{D}_n^i)}\Bigg) \ .$$

The approximations of the k randomized trees are then averaged to obtain the following random forest estimate

$$\widehat{Q}_i^{2,b} = \frac{1}{k}\sum_{\ell=1}^{k}\left[\frac{1}{N_{leaves}^{\ell}}\sum_{m=1}^{N_{leaves}^{\ell}}\left(\min\left\{p=1,\ldots,n,\ (X_i^p,Y^p)\in\mathcal{D}_n^{i\star}(\Theta_\ell) \text{ and } X_i^p\in A_n\left(m;\Theta_\ell,\mathcal{D}_n^i\right)\right\}\right.\right.$$
$$\left.\left.\sum_{j=1}^{n}\frac{B_j\left(\Theta_\ell^1,\mathcal{D}_n^i\right)\cdot\psi_\alpha\left(Y^j,Y^p\right)\cdot\mathbb{1}_{\left\{\left(X_i^j,Y^j\right)\in\mathcal{D}_n^{i\star}(\Theta_\ell),\ X_i^j\in A_n(m;\Theta_\ell,\mathcal{D}_n^i)\right\}}}{N_n^b(m;\Theta_\ell,\mathcal{D}_n^i)}\right)\right].$$

A.1.5 Minimum estimation with a weighted approach and complete trees

By using the weights $w_{n,j}^b(\mathbf{x})$ instead of $w_{n,j}^o(\mathbf{x})$, we may define the estimator $\widehat{Q}_i^{3,b}$.

A.2 Algorithms for estimating the first-order QOSA index

Algorithm 2: QOSA index estimators plugging the quantile

Input:

- Datasets: $\mathcal{D}_n^{\diamond} = \left(\mathbf{X}^{\diamond j}, Y^{\diamond j}\right)_{j=1,\ldots,n}$ and $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\ldots,n}$
- Number of trees: $k \in \mathbb{N}^{\star}$
- Order where estimating the QOSA index : $\alpha \in]0,1[$
- Grid where looking for the best parameter: $grid\_min\_samples\_leaf$
- Number of folds: $K \in \{2,\ldots,n\}$

Output: Estimated value of the QOSA index at the α-order $\widehat{S}_i^{\alpha}$ for all inputs.

Compute $\widehat{P}$ thanks to Equation (2.1).
foreach $i = 1,\ldots,d$ **do**
  $\mathcal{D}_n^{\diamond i} = \left(X_i^{\diamond j}, Y^{\diamond j}\right)_{j=1,\ldots,n}$ from $\mathcal{D}_n^{\diamond}$ and $\mathcal{D}_n^i = \left(X_i^j, Y^j\right)_{j=1,\ldots,n}$ from $\mathcal{D}_n$
  Cross-validation as in Algorithm 1 with $\mathcal{D}_n^i$ to get the optimal leaf size ℓ_{opt}.
  Fit a random forest model with $\mathcal{D}_n^i$ by fixing the $min\_samples\_leaf$ hyperparameter to ℓ_{opt}.
  Compute the estimator $\widehat{R}_i$ with $\mathcal{D}_n^{\diamond i}$.
  Compute $\widehat{S}_i^{\alpha} = 1 - \widehat{R}_i/\widehat{P}$.
end

Algorithm 3: QOSA index estimators with the weighted minimum approach

Input:

- Datasets: $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$ and $\left(\mathbf{X}^{\diamond j}\right)_{j=1,\dots,n}$
- Number of trees: $k \in \mathbb{N}^\star$
- Order where estimating the QOSA index : $\alpha \in]0,1[$
- Grid where looking for the best parameter: *grid_min_samples_leaf*
- Number of folds: $K \in \{2,\dots,n\}$

Output: Estimated value of the QOSA index at the α-order $\widehat{S}_i^\alpha$ for all inputs.

Compute $\widehat{P}$ thanks to Equation (2.1).
foreach $i = 1,\dots,d$ **do**
  $\mathcal{D}_n^i = \left(X_i^j, Y^j\right)_{j=1,\dots,n}$ from $\mathcal{D}_n$ and $\left(X_i^{\diamond j}\right)_{j=1,\dots,n}$
  Cross-validation as in Algorithm 1 with $\mathcal{D}_n^i$ to get the optimal leaf size ℓ_{opt}.
  Fit a random forest model with $\mathcal{D}_n^i$ by fixing the *min_samples_leaf* hyperparameter to ℓ_{opt}.
  Compute the estimator $\widehat{Q}_i \in \left\{\widehat{Q}_i^{1,b}, \widehat{Q}_i^{1,o}\right\}$ with $\mathcal{D}_n^{\diamond i}$ and $\left(X_i^{\diamond j}\right)_{j=1,\dots,n}$.
  Compute $\widehat{S}_i^\alpha = 1 - \widehat{Q}_i/\widehat{P}$
end

Algorithm 4: QOSA index estimators computing the minimum in leaves

Input:

- Datasets: $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$
- Number of trees: $k \in \mathbb{N}^\star$
- Order where estimating the QOSA index : $\alpha \in]0,1[$
- Grid where looking for the best parameter: *grid_min_samples_leaf*
- Number of folds: $K \in \{2,\dots,n\}$

Output: Estimated value of the QOSA index at the α-order $\widehat{S}_i^\alpha$ for all inputs.

Compute $\widehat{P}$ thanks to Equation (2.1).
foreach $i = 1,\dots,d$ **do**
  $\mathcal{D}_n^i = \left(X_i^j, Y^j\right)_{j=1,\dots,n}$ from $\mathcal{D}_n$
  Cross-validation as in Algorithm 1 with $\mathcal{D}_n^i$ to get the optimal leaf size ℓ_{opt}.
  Fit a random forest model with $\mathcal{D}_n^i$ by fixing the *min_samples_leaf* hyperparameter to ℓ_{opt}.
  Compute the estimator $\widehat{Q}_i \in \left\{\widehat{Q}_i^{2,b}, \widehat{Q}_i^{2,o}\right\}$.
  Compute $\widehat{S}_i^\alpha = 1 - \widehat{Q}_i/\widehat{P}$
end

Algorithm 5: QOSA index estimators with the weighted minimum and fully grown trees

Input:

- Datasets: $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$ and $\left(\mathbf{X}^{\diamond j}\right)_{j=1,\dots,n}$
- Number of trees: $k \in \mathbb{N}^\star$
- Order where estimating the QOSA index : $\alpha \in]0,1[$
- Minimum number of samples required in a leaf node: $min\_samples\_leaf \in \{1, ..., n\}$

Output: Estimated value of the QOSA index at the α-order $\widehat{S}_i^\alpha$ for all inputs.

Compute $\widehat{P}$ thanks to Equation (2.1).
Fit a random forest model with $\mathcal{D}_n$ and the $min\_samples\_leaf$ hyperparameter.
foreach $i = 1, \dots, d$ **do**
    Compute the estimator $\widehat{Q}_i \in \left\{\widehat{Q}_i^{3,b}, \widehat{Q}_i^{3,o}\right\}$ with $\left(\mathbf{X}^{\diamond j}\right)_{j=1,\dots,n}$.
    Compute $\widehat{S}_i^\alpha = 1 - \widehat{Q}_i / \widehat{P}$
end