Étude de l'influence de l'échantillonnage statistique et des incertitudes d'inférence probabiliste sur la variance d'un estimateur

Charles Surget, Sylvain Dubreuil, Jérôme Morio, Cécile Mattrand, Jean-Marc Bourinet, Nicolas Gayton

Étude de l'influence de l'échantillonnage statistique et des incertitudes d'inférence probabiliste sur la variance d'un estimateur

C. SURGET[a,b], S. DUBREUIL[a], J. MORIO[a], C. MATTRAND[b], J-M. BOURINET[b], N. GAYTON[b]

a. ONERA/DTIS, F-31055 Toulouse, France
b. SIGMA Clermont, Institut Pascal, F-63000 Clermont–Ferrand, France
charles.surget@onera.fr

Résumé :

Les enjeux industriels actuels nécessitent des essais mécaniques pouvant être coûteux et limitant ainsi la quantité de données à partir desquelles une inférence statistique est réalisée. L'incertitude liée à l'apprentissage de ces données, tels que des paramètres matériaux ou géométriques aléatoires, est alors propagée involontairement au sein de codes de calcul (EF, CFD...). L'estimation d'une quantité d'intérêt en sortie de ce code peut alors être compromise si cette propagation d'incertitude est ommise. La précision de l'estimation peut être améliorée en augmentant le nombre de simulations et d'essais à disposition. Néanmoins, dans un contexte où ceux-ci sont coûteux, une réponse à apporter au compromis essai-simulation est d'une importance majeure. Dans cet article, une méthode d'estimation considérant la variabilité de l'apprentissage est proposée. De plus, une analyse de sensibilité est réalisée afin de répondre à ce compromis à l'aide d'indicateurs permettant de diriger le choix d'investissement dans la base de données ou bien dans l'échantillon de Monte-Carlo. La pertinence de la démarche est illustrée sur des exemples académiques.

Abstract :

Current industrial issues require mechanical tests that can be expensive and thus limit the amount of data from which statistical inference is made. Uncertainty related to the learning of these data, such as random material or geometric parameters, is then unintentionally propagated through mechanical analysis (FE, CFD...). The estimate of a quantity of interest at the output of this code can then be impaired if this propagation of uncertainty is omitted. The accuracy of the estimate can be improved by increasing the number of simulations and tests available. Nevertheless, in a context where they are expensive, an answer to the test-simulation trade-off is of a major importance. In this paper, an estimation method considering the variability of learning is implemented. Moreover, a sensitivity analysis is performed in order to answer this trade-off with indicators allowing to direct the choice of investment in the database or in the Monte-Carlo sample. The relevance of the approach is illustrated on academic examples.

Mots clefs : Small-data, Compromis essai-simulation, Analyse de sensibilité, Pick-Freeze.

1 Introduction

Contrastant avec l'ère moderne du "*big-data*", le concept de "*small-data*" prend de l'ampleur dans le domaine de la recherche [1] notamment en présence de freins technologiques et budgétaires. L'étude de son influence se manifeste au cours des dernières années aussi bien en estimation statistique [2–5] qu'au sein d'autres domaines tel que le *machine learning* [6–9]. Lorsqu'une inférence statistique est réalisée dans un tel contexte, le modèle probabiliste interprété est parfois trop spécifique à l'échantillon de données obtenu. De telles circonstances surviennent, par exemple, lorsque la complexité des applications industrielles limite le nombre d'essais mécaniques réalisés. Les données obtenues, qu'elles soient représentatives de paramètres géométriques ou matériaux, ne sont pas toujours capables de dépeindre la variance de la population dont elles sont issus. Dès lors que les modèles probabilistes inférés nécessitent d'être utilisés, la prise en compte de la propagation des incertitudes en devient primordiale. Les modèles probabilistes inférés sont définis en entrée d'une fonction dite boîte noire par le biais d'une simulation de Monte-Carlo (MCS). Dans un contexte mécanique, ce type de fonction est représentative d'un code de calcul dont les caractéristiques et la robustesse sont des critères indépendants de l'étude en cours. Sans la considération de ces incertitudes, l'interprétation de la quantité d'intérêt (QoI) de sortie à évaluer, telle qu'un effort ou encore un déplacement, s'en retrouve incomplète. L'incertitude au sein de l'estimateur de la QoI est la combinaison de deux niveaux d'incertitudes. Le premier niveau traduit l'incertitude liée à l'estimation statistique intrinsèque et dépend de la taille de l'échantillon MCS. Le second niveau est quant à lui la conséquence directe de l'inférence statistique dont l'incertitude dépend de la taille de l'échantillon de données initial. La précision de l'estimateur peut alors être améliorée grâce à un investissement de données. Néanmoins, lorsque le coût lié aux essais ou le coût d'appel à la boîte noire ne sont pas négligeables, il est souhaitable de déterminer où l'effort doit être investi. La recherche de réponses à ce compromis essai-simulation est d'un intérêt majeur pour la branche industrielle afin de contrôler la précision de l'estimation de la QoI.

Le présent article propose une méthodologie de réponse au compromis essai-simulation où la variance de l'estimateur donné est obtenue en prenant en considération chacune des deux sources d'incertitude précédemment évoquées. L'estimation d'indices de Sobol' [10] lors d'une analyse de sensibilité permet de quantifier la proportion de variance due à chaque échantillon. L'illustration de la méthode au travers d'un cas linéaire gaussien, ainsi que d'une application mécanique simplifiée, permet de mettre en évidence la pertinence d'un investissement guidé par l'indice prépondérant pour répondre au compromis essai-simulation.

L'article est organisé comme suit. La Section 2 présente le premier niveau d'incertitude lié à l'estimation statistique, sa combinaison avec le second niveau qui résulte de l'apprentissage des modèles probabilistes ainsi qu'un estimateur prenant en compte la variabilité de la base de données initiale. La Section 3 introduit ensuite l'approche de réponse au compromis essai-simulation par une analyse de sensibilité basée sur la variance de l'estimateur. La pertinence de la méthode est illustrée en Section 4 au travers d'un cas linéaire gaussien et d'une application mécanique sur poutre encastrée-libre, où l'intérêt pratique d'indicateurs pour diriger le choix d'investissement est exposé. Enfin, la Section 5 conclue le présent article et fournit des perspectives aux travaux effectués.

2 Incertitude à l'estimation et à l'inférence statistique

Le code de calcul mécanique est modélisé par une fonction boîte noire $\phi : \mathcal{X} \to \mathbb{R}$. Son entrée est un vecteur aléatoire continu $\mathbf{X}$, défini sur le domaine $\mathcal{X} \subseteq \mathbb{R}^d$, de densité de probabilité (PDF) $f_{\mathbf{X}}$. Sa sortie est une variable aléatoire Y telle que $\mathrm{Y} = \phi(\mathbf{X})$. Un intérêt spécifique est accordé à l'espérance d'une fonction particulière τ de Y telle qu'une moyenne ou une probabilité de défaillance. L'espérance $\mathbb{E}_{f_{\mathbf{X}}}$ peut être estimée par MCS :

$$\mathbb{E}_{f_{\mathbf{X}}}\left[\tau\left(\phi(\mathbf{X})\right)\right] = \int_{\mathcal{X}} \tau\left(\phi\left(\mathbf{x}\right)\right) f(\mathbf{x})\, \mathrm{d}\mathbf{x} \tag{2.1}$$

$$\approx \frac{1}{N_{\mathbf{X}}} \sum_{j=1}^{N_{\mathbf{X}}} \tau\left(\phi\left(\mathbf{X}^{(j)}\right)\right), \tag{2.2}$$

où $\mathbf{X}^{(j)} \overset{i.i.d.}{\sim} f_{\mathbf{X}}$. La précision de l'estimateur dépend de $N_{\mathbf{X}}$, la taille de l'échantillon MCS. Néanmoins, la connaissance de $f_{\mathbf{X}}$ en vue d'un échantillonnage est nécessaire.

Dans un contexte industriel, la connaissance de la loi de probabilité de l'entrée $\mathbf{X}$ est parfois restreinte à un $N_{\mathbf{D}}$-échantillon, avec $N_{\mathbf{D}}$ supposé faible. La base de données $\widetilde{\mathbf{D}} := \left(\mathbf{D}^{(1)}, ..., \mathbf{D}^{(N_{\mathrm{D}})}\right)$ constituée à partir d'essais mécaniques est définie sur le domaine $\mathcal{D} \subseteq \mathbb{R}^{d \times N_{\mathrm{D}}}$ et de PDF $f_{\widetilde{\mathbf{D}}}$. Elle regroupe $N_{\mathbf{D}}$ vecteurs aléatoires $\mathbf{D}^{(i)}$ qui sont indépendants et identiquement distribués (i.i.d.) selon une réelle PDF $f_{\mathbf{X}}$, néanmoins supposée inconnue.

La méconnaissance de la loi sous-jacente $f_{\mathbf{X}}$ nécessite alors son apprentissage. Son identification à partir d'une base de données consiste en l'approximation de la densité $f_{\mathbf{X}}$ par $\hat{f}_{\mathbf{X}|\widetilde{\mathbf{D}}}$. Les méthodes disponibles pour l'identification se classent principalement en deux familles : les approches paramétriques et non-paramétriques. Les approches paramétriques consistent à approcher la loi cible par une loi au sein d'une famille de distributions paramétriques [11]. Les approches non-paramétriques, quant à elles, s'affranchissent de cette sélection dans une famille de distribution tout en apportant une estimation robuste de la densité [12, 13]. La méthode d'estimation par noyau [14] a été retenue dans cette article pour sa facilité de mise en œuvre dans un cadre multivarié, tout en ayant pour ambition d'échantillonner aisément selon la densité estimée. Un noyau gaussien est utilisé et le paramètre de lissage est défini selon la règle de Silvermann [15]. L'échantillon constitué n'étant plus distribué directement selon $f_{\mathbf{X}}$, l'Equation (2.1) devient :

$$\mathbb{E}_{\hat{f}_{\mathbf{X}|\widetilde{\mathbf{D}}}}\left[\tau\left(\phi(\mathbf{X})\right)\right] = \int_{\mathcal{X}} \tau\left(\phi\left(\mathbf{x}\right)\right) \hat{f}_{\mathbf{X}|\widetilde{\mathbf{D}}}(\mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{2.3}$$

avec $\mathbf{X}^{(j)} \overset{i.i.d.}{\sim} \hat{f}_{\mathbf{X}|\widetilde{\mathbf{D}}}$ dans le cadre de son estimation. Néanmoins, dans un contexte "*small-data*", le risque d'obtenir une estimation $\hat{f}_{\mathbf{X}|\widetilde{\mathbf{D}}}$ spécifique à l'échantillon de données initial et comportant trop de dissimilarités avec $f_{\mathbf{X}}$ est important [16]. Il est donc nécessaire d'exprimer l'espérance relative au couple $(\mathbf{X}, \widetilde{\mathbf{D}})$ prenant en compte la variabilité de la base de données. Son expression est la suivante :

$$\begin{aligned}
\mathbb{E}_{f_{(\mathbf{X},\widetilde{\mathbf{D}})}}\left[\tau\left(\phi(\mathbf{X})\right)\right] &= \int_{\mathcal{D}} \int_{\mathcal{X}} \tau\left(\phi(\mathbf{x})\right) f_{(\mathbf{X},\widetilde{\mathbf{D}})}(\mathbf{x}, \widetilde{\mathbf{d}})\, \mathrm{d}\mathbf{x}\mathrm{d}\widetilde{\mathbf{d}} && (2.4)\\
&= \int_{\mathcal{D}} \left(\int_{\mathcal{X}} \tau\left(\phi(\mathbf{x})\right) f_{\mathbf{X}|\widetilde{\mathbf{D}}}(\mathbf{x}|\widetilde{\mathbf{d}})\, \mathrm{d}\mathbf{x} \right) f_{\widetilde{\mathbf{D}}}(\widetilde{\mathbf{d}})\, \mathrm{d}\widetilde{\mathbf{d}} \\
&\approx \int_{\mathcal{D}} \left(\int_{\mathcal{X}} \tau\left(\phi(\mathbf{x})\right) \hat{f}_{\mathbf{X}|\widetilde{\mathbf{D}}}(\mathbf{x}|\widetilde{\mathbf{d}})\, \mathrm{d}\mathbf{x} \right) f_{\widetilde{\mathbf{D}}}(\widetilde{\mathbf{d}})\, \mathrm{d}\widetilde{\mathbf{d}}, && (2.5)
\end{aligned}$$

L'estimation de ce type d'intégrale double est possible à l'aide d'estimateurs imbriqués (NRA) ou bien d'estimateurs de l'espace augmenté (ARA). Bien que les avantages d'un estimateur ARA au regard de l'estimateur NRA aient été présentés dans [17], l'estimateur NRA sera privilégié dans cet article pour l'aisance d'adaptation avec la méthodologie proposée. L'estimateur NRA de l'espérance en Equation (2.5) se présente ainsi :

$$\begin{aligned} \hat{\mu}^{NRA} &= \frac{1}{N}\sum_{k=1}^{N}\frac{1}{N_{\mathbf{X}}}\sum_{j=1}^{N_{\mathbf{X}}}\tau\left(\phi\left(\mathbf{X}_k^{(j)}\right)\right) \\ &= \frac{1}{N}\sum_{k=1}^{N}\hat{\mu}_k \end{aligned} \tag{2.6}$$

avec $\mathbf{X}_k^{(j)} \overset{i.i.d.}{\sim} \hat{f}_{\mathbf{X}|\widetilde{\mathbf{D}}_k}$ et N est la quantité de bases de données $\widetilde{\mathbf{D}}_k$ à disposition. Une estimation est alors possible dès lors qu'un échantillonnage selon $f_{\widetilde{\mathbf{D}}}$ est permis, ou bien qu'une base de données de taille suffisante est partitionnable.

En pratique et dans un cadre "*small-data*", il est déjà épineux de se procurer une unique base de données de taille $N_{\mathbf{D}}$ suffisante. Une solution consiste à exploiter les méthodes de ré-échantillonnage [18] qui permettent de simuler l'inférence sur la base de l'échantillon initial. Le *Bootstrap*, en plus d'être une technique plus permissive que ses pairs [19], permet de modéliser l'impact de la taille de l'échantillon initial. Ainsi, la variance de l'estimateur représente l'incertitude liée à la combinaison des deux sources et il est alors intéressant de se questionner sur la contribution de chacune d'elle.

3 Analyse de sensibilité

Une analyse de sensibilité est réalisée afin de répondre au compromis essai-simulation. Son objectif est l'estimation d'indices permettant de quantifier la proportion d'incertitudes liée au $N_{\mathbf{D}}$-échantillon $\widetilde{\mathbf{D}}$ et au $N_{\mathbf{X}}$-échantillon $\widetilde{\mathbf{X}} := \left(\mathbf{X}^{(1)}, ..., \mathbf{X}^{(N_{\mathbf{X}})}\right)$. La fonction Γ est introduite afin d'étudier l'influence de ses entrées sur sa sortie :

$$\Gamma : \left| \begin{array}{ccc} \mathcal{X}^{N_{\mathbf{X}}} \times \mathcal{D} & \longrightarrow & \mathbb{R} \\ (\widetilde{\mathbf{x}}, \widetilde{\mathbf{d}}) & \longmapsto & \hat{\mu} = \Gamma(\widetilde{\mathbf{x}}, \widetilde{\mathbf{d}}). \end{array} \right. \tag{3.1}$$

Un obstacle pour l'analyse de sensibilité réside dans la dépendance entre l'échantillon MCS et l'échantillon de données initial. En effet, les vecteurs aléatoires $\mathbf{X}^{(j)}$ sont directement distribués selon la PDF $\hat{f}_{\mathbf{X}|\widetilde{\mathbf{D}}}$, elle même estimée à partir de $\widetilde{\mathbf{D}}$. La fonction Γ s'exprime comme l'estimateur de l'Equation (2.3) et $\widetilde{\mathbf{D}}$ n'est alors plus une entrée directe de cette fonction. L'aisance de mise en œuvre de l'analyse de sensibilité semble alors compromise.

Pour contourner cet obstacle, il est possible d'envisager un remplacement de l'entrée $\widetilde{\mathbf{X}}$ par une autre variable, dont la facilité d'échantillonnage et l'indépendance avec $\widetilde{\mathbf{D}}$ est évidente. Par la suite, l'application d'une transformation iso-probabiliste [20–24] permet de réaliser le passage entre la variable désirée et une variable décorrelée spécifique à la méthode utilisée. L'indifférence du choix de la méthode est démontré par [25]. Ainsi, l'entrée $\widetilde{\mathbf{X}}$ de la fonction Γ est reconsidérée pour une entrée $\widetilde{\mathbf{U}}$ définie sur l'hypercube $[0,1]^{d\times N_{\mathbf{X}}}$. L'application en Equation (3.1) devient alors :

$$\Gamma : \left| \begin{array}{ccc} [0,1]^{d\times N_{\mathbf{X}}} \times \mathcal{D} & \longrightarrow & \mathbb{R} \\ \mathbf{z} & \longmapsto & \hat{\mu} = \Gamma(\mathbf{z}), \end{array} \right. \tag{3.2}$$

avec $\mathbf{z} = (\widetilde{\mathbf{u}}, \widetilde{\mathbf{d}})$. Les vecteurs aléatoires $\mathbf{U}^{(j)}$ sont distribués indépendemment de $\widetilde{\mathbf{D}}$ et l'analyse de sensibilité peut alors être réalisée avec des entrées indépendantes. Une transformation de Rosenblatt [23, 24] inverse permet par la suite de passer d'un échantillon $\widetilde{\mathbf{U}}$ à un échantillon $\widetilde{\mathbf{X}}$:

$$\mathcal{T}_{\widetilde{\mathbf{D}}} : \quad \mathbf{u} = \begin{bmatrix} \mathrm{u}_1 \\ \mathrm{u}_2 \\ \vdots \\ \mathrm{u}_d \end{bmatrix} \longmapsto \mathbf{x} = \begin{bmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \\ \vdots \\ \mathrm{x}_d \end{bmatrix} \text{ avec } \begin{cases} \mathrm{x}_1 = \hat{F}^{-1}_{\mathrm{X}_1|\widetilde{\mathbf{D}}}(\mathrm{u}_1) \\ \mathrm{x}_2 = \hat{F}^{-1}_{(\mathrm{X}_2|\mathrm{X}_1)|\widetilde{\mathbf{D}}}(\mathrm{u}_2|\mathrm{u}_1) \\ \vdots \\ \mathrm{x}_d = \hat{F}^{-1}_{(\mathrm{X}_d|\mathrm{X}_{d-1},\ldots,\mathrm{X}_1)|\widetilde{\mathbf{D}}}(\mathrm{u}_d|\mathrm{u}_{d-1},\ldots,\mathrm{u}_1) \end{cases} . \tag{3.3}$$

$$\mathbf{U} \sim \mathcal{U}\left([0,1]\right)^d \longmapsto \mathbf{X} \sim \hat{f}_{\mathbf{X}|\widetilde{\mathbf{D}}}$$

Les entrées sont dorénavant explicites dans la formule de l'estimation $\hat{\mu}$ qui se note :

$$\hat{\mu} = \frac{1}{N_{\mathbf{X}}} \sum_{j=1}^{N_{\mathbf{X}}} \tau \left(\phi \left(\mathcal{T}_{\widetilde{\mathbf{D}}} \left(\mathbf{U}^{(j)} \right) \right) \right). \tag{3.4}$$

L'objectif étant d'améliorer la précision de l'estimateur, une analyse basée sur la variance, aussi appelée ANOVA (*ANalyse Of VAriance*), est réalisée à l'aide des indices de Sobol clos [10]. Les indices $S_{\widetilde{\mathbf{D}}}$ et $S_{\widetilde{\mathbf{U}}}$ associés quantifient la proportion de variance due à chaque échantillon et sont alors estimés par la méthode *Pick-Freeze* [26]. L'obtention d'échantillons gelés passent par la définition d'une variable $\widetilde{\mathbf{Z}}$ et de son homologue $\widetilde{\mathbf{Z}}'$ de telle sorte à construire leurs version gelées $\widetilde{\mathbf{Z}}^{\widetilde{\mathbf{D}}}$ et $\widetilde{\mathbf{Z}}^{\widetilde{\mathbf{U}}}$:

$$\widetilde{\mathbf{Z}}^{\widetilde{\mathbf{D}}} = \begin{bmatrix} \widetilde{\mathbf{U}}'_1 & \widetilde{\mathbf{D}}_1 \\ \vdots & \vdots \\ \widetilde{\mathbf{U}}'_k & \widetilde{\mathbf{D}}_k \\ \vdots & \vdots \\ \widetilde{\mathbf{U}}'_N & \widetilde{\mathbf{D}}_N \end{bmatrix}, \quad \widetilde{\mathbf{Z}}^{\widetilde{\mathbf{U}}} = \begin{bmatrix} \widetilde{\mathbf{U}}_1 & \widetilde{\mathbf{D}}'_1 \\ \vdots & \vdots \\ \widetilde{\mathbf{U}}_k & \widetilde{\mathbf{D}}'_k \\ \vdots & \vdots \\ \widetilde{\mathbf{U}}_N & \widetilde{\mathbf{D}}'_N \end{bmatrix} \text{ avec } \begin{cases} \widetilde{\mathbf{Z}} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_k, \ldots, \mathbf{Z}_N) \\ \widetilde{\mathbf{Z}}' = (\mathbf{Z}'_1, \ldots, \mathbf{Z}'_k, \ldots, \mathbf{Z}'_N) \end{cases} . \tag{3.5}$$

Les sorties de la fonction Γ aux entrées $\mathbf{Z}_k$ avec $k = 1, ..., N$ sont calculées. Les estimations $\hat{\mu}^{\widetilde{\mathbf{D}}}$ et $\hat{\mu}^{\widetilde{\mathbf{U}}}$, étant respectivement les sorties dont les entrées $\widetilde{\mathbf{D}}$ et $\widetilde{\mathbf{U}}$ ont été gelées, se définissent par l'Equation (3.4) à partir de $\widetilde{\mathbf{Z}}^{\widetilde{\mathbf{D}}}$ et $\widetilde{\mathbf{Z}}^{\widetilde{\mathbf{U}}}$. Ils permettent ainsi l'estimation des indices de Sobol' suivant :

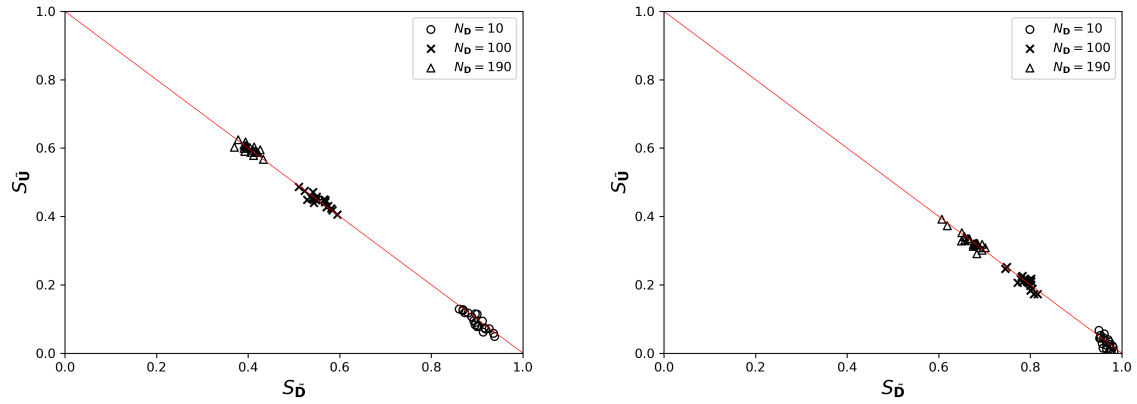
$$\begin{cases} S_{\widetilde{\mathbf{D}}} = \dfrac{Cov\left[\hat{\mu}, \hat{\mu}^{\widetilde{\mathbf{D}}}\right]}{Var\left[\hat{\mu}\right]} \\ S_{\widetilde{\mathbf{U}}} = \dfrac{Cov\left[\hat{\mu}, \hat{\mu}^{\widetilde{\mathbf{U}}}\right]}{Var\left[\hat{\mu}\right]} \end{cases}, \tag{3.6}$$

où la variance et la covariance sont estimées empiriquement au moyen de N réalisations. L'ordre de grandeur du nombre d'appel à ϕ est alors de $3 \times N \times N_{\mathbf{X}}$, bien inférieur au nombre d'appels exigés par la méthode double Monte-Carlo. Ces indices permettent une compréhension aisée de la proportion d'incertitude au sein de chaque échantillon. Un investissement dans la valeur la plus élevée a pour vocation de réduire la variance de la QoI relative à la source prédominante. Il est important de mentionner qu'un investissement de données peut occasionnellement augmenter la variance de l'estimateur. Ce phénomène survient lors de l'addition d'une réalisation dans une zone peu explorée telle que les queues de distributio au sein de la base de données initiale. Dans des circonstances particulières plus largement développées par [27], un investissement dans l'échantillon MCS peut provoquer un évènement similaire.

4 Résultats

Cette section vise à illustrer la pertinence de la méthode au travers de deux exemples numériques. Un cas linéaire gaussien est considéré pour la première application. Pour ce cas comme le suivant, la fonction particulière τ spécifiée dès l'Equation (2.1) est la fonction identité. La base de données initiale est constituée à partir d'un vecteur aléatoire $\mathbf{D} = (\mathrm{D}_1, \ldots, \mathrm{D}_d)^t$ composé de variables aléatoires D_i de telle sorte que $\mathrm{D}_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ pour $i = 1, \ldots, d$. Le cadre de cette application est restreint à une dimension $d = 2$ et la fonction ϕ est une fonction analytique simple consistant en la somme des entrées afin de garantir la linéarité du cas d'étude. La taille des échantillons *Pick-Freeze* présentés en Equation (3.5) est fixée à $N = 10^4$ et le couple $(N_{\mathbf{X}}, N_{\mathbf{D}})$ est initialisé à $(150, 10)$ afin de rester dans un contexte "*small-data*". À chaque état, $n = 20$ estimations à partir d'une même base de données initiale sont réalisées dans le but de visualiser la variance des indices.

Le comportement des indices de Sobol' après investissement dans la base de données est illustré graphiquement en Figure 1a. Dans un premier temps, il est intéressant de noter que les indices sont tous estimés autour d'une valeur dont la somme est unitaire. Cet aspect, représenté par la droite rouge, informe sur la faible ampleur ou l'absence d'impact d'une intéraction entre $\widetilde{\mathbf{D}}$ et $\widetilde{\mathbf{U}}$. Dans un second temps, les estimations à l'état initial indiquent une influence dominante sur la variance de l'estimateur par la base de données $\widetilde{\mathbf{D}}$. Ce résultat est le reflet de l'incertitude relative à l'apprentissage de modèles probabilistes lorsque la quantité de données observées est faible. Un investissement successif dans $N_{\mathbf{D}}$ montre



(a) (b)

FIGURE 1 – Évolution des indices de Sobol' pour le cas d'étude linéaire gaussien à $d = 2$ en fonction de $N_{\mathbf{D}}$ avec (a) $N_{\mathbf{X}} = 150$ et (b) $N_{\mathbf{X}} = 450$. Visualisation de $n = 20$ estimations pour chaque combinaison de couple $(\tilde{\mathbf{u}}, \tilde{\mathbf{d}})$.

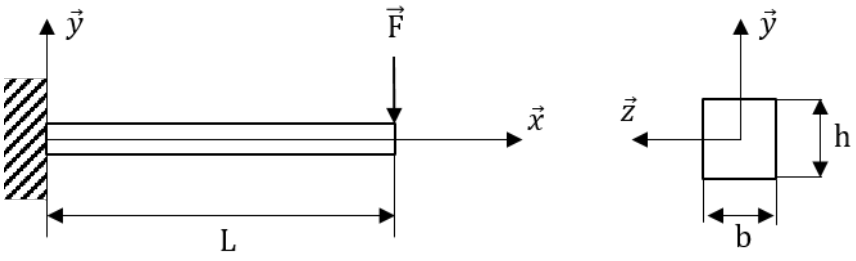
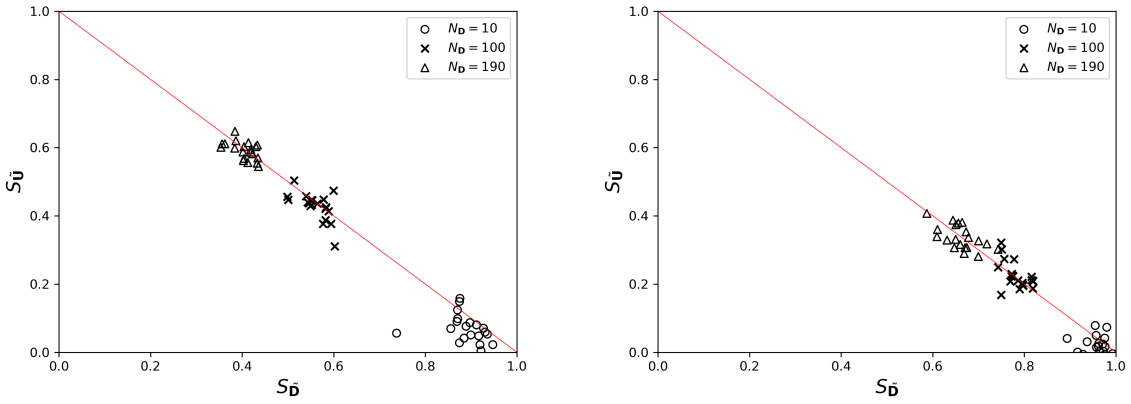


FIGURE 2 – Représentation d'une poutre encastrée-libre de longueur L, de module de Young E et de section b × h soumis à un chargement F vertical en son extrémité.

TABLE 1 – Distributions associées aux variables aléatoires d'entrée pour le cas test d'une poutre encastrée-libre soumis à un chargement en son extrémité.

Variable d'entrée	Distribution	Moyenne	Coefficient de variation
F	LogNormal	556.8 [N]	0.08
L	Normal	4290 [mm]	0.1
E	LogNormal	2.10^5 [MPa]	0.06
b	Normal	62 [mm]	0.1
h	Normal	98.7 [mm]	0.1



(a) (b)

FIGURE 3 – Évolution des indices de Sobol' pour le cas de la poutre encastrée-libre à $d = 5$ en fonction de $N_{\mathbf{D}}$ avec (a) $N_{\mathbf{X}} = 150$ et (b) $N_{\mathbf{X}} = 450$. Visualisation de $n = 20$ estimations pour chaque combinaison de couple $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{d}})$.

une diminution de cet indice et par complémentarité, une augmentation de l'influence de l'échantillon MCS. Les conséquences liées au contexte "*small-data*"sont constatable au couple $(150, 190)$ où la part de variance liée à $\widetilde{\mathbf{U}}$ domine celle due à la base de données initiale. C'est notamment lors de ces étapes en amont, où les données à disposition sont peu nombreuses, que le choix d'investissement n'est pas intuitif et que l'importance des ces indices se révèle. Dans le cas où $N_{\mathbf{X}}$ est défini à une valeur plus grande comme en Figure 1b, l'évolution de l'indice $S_{\widetilde{\mathbf{D}}}$ est plus lente. Un investissement intuitif sera alors guidé par l'indice le plus élevé avec pour finalité d'équilibrer les parts de variance.

Un second cas présentant l'étude de la flèche d'une poutre encastrée-libre soumis à un chargement en son extrémité est représentée en Figure 2. Il s'agit d'un problème en dimension $d = 5$ dont les lois de probabilité marginales sous-jacentes ont été introduites par [28] et son rappelées dans la Table 1. Bien que ce ne soit pas le cas en pratique, une hypothèse d'indépendance sur les variables est faite pour des facilités d'implémentation. L'estimation d'une copule et son utilisation dans le cadre d'une transformation iso-probabiliste sont détaillées par [2] mais ne sera pas développée dans cet article. La taille des échantillons est fixée cette fois-ci à $N = 10^3$. La QoI étudiée est la flèche moyenne dont l'expression suivante est rappelée :

$$\phi\left(\mathrm{F}, \mathrm{L}, \mathrm{E}, \mathrm{b}, \mathrm{h}\right) = \frac{4\mathrm{F}\mathrm{L}^3}{\mathrm{E}\mathrm{b}\mathrm{h}^3} \tag{4.1}$$

Après analyse, l'augmentation de la dimension impacte l'estimation des indices de Sobol' présentés en Figure 3. Les estimations obtenues sont plus dispersées que le cas précédent en conséquence de

la faible taille des échantillons. Cette dispersion diminue lors d'un apport de données en $N_{\mathbf{X}}$ et $N_{\mathbf{D}}$. L'interprétation des résultats est ici similaire au cas gaussien linéaire et traduit la capabilité des indices à répondre au compromis essai-simulation. Il est néanmoins important de noter que l'estimation par noyau directe sur la loi jointe n'est plus adéquate lors de la montée en dimension. En effet, les bases de données ré-échantillonnées étant de petites tailles, la répétition trop fréquente d'une même réalisation du vecteur aléatoire provoquerait une mauvaise estimation de la loi. La fréquence d'apparition d'erreurs d'estimation seraient alors nettement plus importante pour de faibles valeurs de $N_{\mathbf{D}}$.

5 Conclusion

Lorsque des modèles probabilistes définis en entrée d'un code de calcul ne sont pas connus, leur apprentissage par une inférence statistique sur une base de données est nécessaire. La quantité d'intérêt en sortie du code est donc soumise à une incertitude à deux niveaux originaire de l'échantillon de Monte-Carlo et de l'échantillon de données initial. Dans un contexte "*small-data*", la quantité de données à disposition est restreinte par la complexité des essais et des modèles de simulation. De telles circonstances impliquent de répondre à un compromis essai-simulation afin de favoriser la précision de l'estimateur statistique de la quantité d'intérêt.

Dans le présent article, une méthodologie de réponse en deux étapes est proposée. Dans un premier temps, un estimateur statistique prenant en compte la variabilité de la base de données est exprimé. L'apprentissage réalisé consiste en un ré-échantillonnage *Bootstrap* et une méthode d'estimation par noyau des lois de probabilité. Puis, dans un second temps, une analyse de sensibilité basée sur la variance de l'estimateur est réalisée. L'indépendance des sources à étudier est obtenue par une transformation iso-probabiliste et une méthode *Pick-Freeze* permet de réduire le nombre d'appel à la boîte noire. Les indices de Sobol' ainsi obtenus traduisent la proportion de variance de l'estimateur due à la base de données initiale et à l'échantillon de Monte-Carlo. L'investissement de données est réalisé dans la source donc l'indice prédomine. La pertinence de la méthode est illustrée par un cas linéaire gaussien à deux entrées puis par un cas mécanique de poutre encastrée-libre à cinq entrées.

La perspective principale à envisager pour cette approche concerne la rétro-activité de la méthode. En effet, un investissement dans la base de données initiale implique de réaliser de nouveaux appels à la boîte noire. L'éventail des méthodes associées à l'échantillonnage préférentiel [29] pourrait permettre de minimiser le nombre d'appel à la boîte noire par le recyclage des données aux étapes d'investissement antérieures. Une autre perspective consiste à prendre en compte l'aspect budgétaire des essais et du code de calcul dans le processus d'investissement. Il est alors intéressant d'observer l'évolution de la variance rapportée à la quantité de données ajoutées et de définir le choix d'investissement relatif à cette évolution. Enfin, la méthode d'apprentissage proposée est une parmi d'autres et son impact pourra être étudié dans le cadre de futurs travaux.

Références

[1] Rob Kitchin and Tracey P Lauriault. Small data in the era of big data. *GeoJournal*, 80(4) :463–475, 2015.

[2] Gabriel Sarazin. *Analyse de sensibilité fiabiliste en présence d'incertitudes épistémiques introduites par les données d'apprentissage*. PhD thesis, Toulouse, ISAE, 2021.

[3] Jiaxin Zhang and Michael D Shields. On the quantification and efficient propagation of imprecise probabilities resulting from small datasets. *Mechanical Systems and Signal Processing*, 98 :465–483, 2018.

[4] Zhenyu Gao. *A nonparametric-based approach on the propagation of imprecise probabilities due to small datasets*. PhD thesis, Georgia Institute of Technology, 2018.

[5] Rens Van De Schoot, Joris J Broere, Koen H Perryck, Mariëlle Zondervan-Zwijnenburg, and Nancy E Van Loey. Analyzing small data sets using bayesian estimation : The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European journal of psychotraumatology*, 6(1) :25216, 2015.

[6] Ying Zhang and Chen Ling. A strategy to apply machine learning to small datasets in materials science. *Npj Computational Materials*, 4(1) :1–8, 2018.

[7] Torgyn Shaikhina and Natalia A Khovanova. Handling limited datasets with neural networks in medical applications : A small-data approach. *Artificial intelligence in medicine*, 75 :51–63, 2017.

[8] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449, 2015.

[9] Luigi Fortuna, Salvatore Graziani, and Maria Gabriella Xibilia. Comparison of soft-sensor design methods for industrial plants using small data sets. *IEEE Transactions on instrumentation and measurement*, 58(8) :2444–2451, 2009.

[10] Ilya M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3) :271–280, 2001.

[11] James K Lindsey et al. *Parametric statistical inference*. Oxford University Press, 1996.

[12] Alan Julian Izenman. Review papers : Recent developments in nonparametric density estimation. *Journal of the american statistical association*, 86(413) :205–224, 1991.

[13] Larry Wasserman. Bayesian inference. In *All of Statistics*, pages 175–192. Springer, 2004.

[14] Bernard W Silverman. *Density estimation for statistics and data analysis*. London : Chapman and Hall, 1986.

[15] Nils-Bastian Heidenreich, Anja Schindler, and Stefan Sperlich. Bandwidth selection for kernel density estimation : a review of fully automatic selectors. *AStA Advances in Statistical Analysis*, 97(4) :403–433, 2013.

[16] Jiaxin Zhang and Michael D Shields. The effect of prior probabilities on quantification and propagation of imprecise probabilities resulting from small datasets. *Computer Methods in Applied Mechanics and Engineering*, 334 :483–506, 2018.

[17] Vincent Chabridon. *Analyse de sensibilité fiabiliste avec prise en compte d'incertitudes sur le modèle probabiliste-Application aux systèmes aérospatiaux*. PhD thesis, Université Clermont Auvergne(2017-2020), 2018.

[18] Chong Ho Yu. Resampling methods : concepts, applications, and justification. *Practical Assessment, Research, and Evaluation*, 8(1) :19, 2002.

[19] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[20] Andre Nataf. Determination des distribution don t les marges sont donnees. *Comptes Rendus de l Academie des Sciences*, 225 :42–43, 1962.

[21] Armen Der Kiureghian and Pei-Ling Liu. Structural reliability under incomplete probability information. *Journal of Engineering Mechanics*, 112(1) :85–104, 1986.

[22] Pei-Ling Liu and Armen Der Kiureghian. Multivariate distribution models with prescribed marginals and covariances. *Probabilistic engineering mechanics*, 1(2) :105–112, 1986.

[23] Murray Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3) :470–472, 1952.

[24] AE Brockwell. Universal residuals : A multivariate transformation. *Statistics & probability letters*, 77(14) :1473–1478, 2007.

[25] Régis Lebrun and Anne Dutfoy. Do rosenblatt and nataf isoprobabilistic transformations really differ ? *Probabilistic Engineering Mechanics*, 24(4) :577–584, 2009.

[26] Fabrice Gamboa, Alexandre Janon, Thierry Klein, A Lagnoux, and Clémentine Prieur. Statistical inference for sobol pick-freeze monte carlo method. *Statistics*, 50(4) :881–902, 2016.

[27] Pascal Pernot, Michèle Désenfant, and François Hennebelle. Model's output variance can increase when input variance decreases : a sensitivity analysis paradox ? In *17th International Congress of Metrology*, page 02004. EDP Sciences, 2015.

[28] LI Baoyu, Leigang Zhang, ZHU Xuejun, YU Xiongqing, and MA Xiaodong. Reliability analysis based on a novel density estimation method for structures with correlations. *Chinese Journal of Aeronautics*, 30(3) :1021–1030, 2017.

[29] Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449) :135–143, 2000.