Robustesse et Interprétabilité des modèles boîtes-noires

Utiliser les IA pour les Systèmes Critiques

[1]**EDF R&D - Lab Chatou** - PRISME Department
[2]**Institut de Mathématiques de Toulouse**
[3]**SINCLAIR AI Lab**

Webinar R&D
Nexialog Consulting - Paris, France.
30 Janvier 2024

Marouane Il Idrissi[123], Nicolas Bousquet[13], Fabrice Gamboa[2], Bertrand Iooss[123], Jean-Michel Loubes[2]

Pourquoi les algorithmes d'IA ne sont pas largement intégrés dans les systèmes critiques ?

Systèmes critiques : *Barrage hydroélectrique, centrale nucléaire, voiture, avion, train, marché financier, système judiciaire, un corps humain...*

Pourquoi les algorithmes d'IA ne sont pas largement intégrés dans les systèmes critiques ?

Systèmes critiques : *Barrage hydroélectrique, centrale nucléaire, voiture, avion, train, marché financier, système judiciaire, un corps humain...*

Raison principale : La **prise de décision** doit être **justifiable et justifiée**.

Or, les **arguments empiriques (SOTA)** ont **peu de poids** auprès des **autorités de sûreté/contrôle**.

Pourquoi les algorithmes d'IA ne sont pas largement intégrés dans les systèmes critiques ?

Systèmes critiques : *Barrage hydroélectrique, centrale nucléaire, voiture, avion, train, marché financier, système judiciaire, un corps humain...*

Raison principale : La **prise de décision** doit être **justifiable et justifiée**.

Or, les **arguments empiriques (SOTA)** ont **peu de poids** auprès des **autorités de sûreté/contrôle**.

Et ce, malgré la **plus-value évidente** que l'**apprentissage automatique** peut offrir...

La plus-value de l'IA

Par exemple, l'**utilisation de bases de données** pour :

- Modéliser finement des **risques de crédits** pour optimiser le provisionnement bancaire.
- Appréhender les **effets du changement climatique** sur les primes d'assurance.
- Repérer les **fraudes** et le **financement terroriste**.

La plus-value de l'IA

Par exemple, l'**utilisation de bases de données** pour :

- Modéliser finement des **risques de crédits** pour optimiser le provisionnement bancaire.
- Appréhender les **effets du changement climatique** sur les primes d'assurance.
- Repérer les **fraudes** et le **financement terroriste**.

Mais les prises de décision sont **régulées** et **contrôlées** (ABE, AMF, ACPR...) :

- **Éthique** (variables protégées).
- **Transparence de l'information**.
- **Stabilité des marchés**.

La plus-value de l'IA

Par exemple, l'**utilisation de bases de données** pour :

- Modéliser finement des **risques de crédits** pour optimiser le provisionnement bancaire.
- Appréhender les **effets du changement climatique** sur les primes d'assurance.
- Repérer les **fraudes** et le **financement terroriste**.

Mais les prises de décision sont **régulées** et **contrôlées** (ABE, AMF, ACPR...) :

- **Éthique** (variables protégées).
- **Transparence de l'information**.
- **Stabilité des marchés**.

Comment justifier une prise de décision soutenue par des modèles boîtes-noires (par ex, l'IA) ?

Deux littératures

L'**intelligence artificielle explicable (XAI)** cherche à étudier **cette question**.

S'inscrit dans la **littérature ML/DL**, avec une vision "computer science" importante.

Beaucoup de méthodes proposées (LIME, SHAP...), mais qui restent **principalement justifiées empiriquement**.

Deux littératures

L'**intelligence artificielle explicable (XAI)** cherche à étudier **cette question**.

S'inscrit dans la **littérature ML/DL**, avec une vision "computer science" importante.

Beaucoup de méthodes proposées (LIME, SHAP...), mais qui restent **principalement justifiées empiriquement**.

L'**analyse de sensibilité (SA)** à pour but la **compréhension de modèles boîtes-noires**.

Découle de la **quantification des incertitudes (UQ)** en milieu industriel, avec un fort **formalisme mathématique.**

Les méthodes proposées reposent sur un **cadre mathématique clair et bien posé** (garanties théoriques).

Nos travaux

Notre position :

Transposer le **formalisme mathématique de l'analyse de sensibilité** pour **mieux comprendre les IA**, par la construction de **méthodes justifiées théoriquement.**

Nos travaux

Notre position :

Transposer le **formalisme mathématique de l'analyse de sensibilité** pour **mieux comprendre les IA**, par la construction de **méthodes justifiées théoriquement.**

Pourquoi ?

S'inspirer de l'**avènement des outils de calcul scientifique (modèles numériques)** permettant de **simuler les phénomènes physiques**.

Ces **modèles "boîtes-noires"** servent aujourd'hui d'**appui pour la prise de décision sur des systèmes critiques** : *dossiers de sûreté, prévision de profitabilité, jumeaux numériques...*

Nos travaux

Notre position :

Transposer le **formalisme mathématique de l'analyse de sensibilité** pour **mieux comprendre les IA**, par la construction de **méthodes justifiées théoriquement.**

Pourquoi ?

S'inspirer de l'**avènement des outils de calcul scientifique (modèles numériques)** permettant de **simuler les phénomènes physiques**.

Ces **modèles "boîtes-noires"** servent aujourd'hui d'**appui pour la prise de décision sur des systèmes critiques** : *dossiers de sûreté, prévision de profitabilité, jumeaux numériques...*

Sujet de thèse :

*"**Développement de méthodes d'interprétabilité en apprentissage automatique** pour la **certification des intelligences artificielles** reliées aux **systèmes critiques**."*

Trois chantiers

Formalisation générale de l'interprétabilité des modèles boîtes-noires

Vision unificatrice *de la* ***quantification des incertitudes des modèles numériques****, et de* ***l'interprétation des IA****.*

Robustesse des IA aux perturbations

Comment va se comporter un modèle boîte-noire face à des perturbations sur ses entrées ?

Quantification d'influence en situation d'entrées dépendantes

Quelles sont les entrées les plus influentes/importantes pour une quantité d'intérêt reliée à un modèle boîte-noire ?

Robustesse aux perturbations

Étant donné des perturbations sur les entrées d'un modèle, comment celui-ci va-t-il se comporter ?

Robustesse aux perturbations

Étant donné des perturbations sur les entrées d'un modèle, comment celui-ci va-t-il se comporter ?

Prenons un exemple :

Supposons que nous ayons a disposition un modèle (IA ou non) qui permet de **prédire le niveau de l'eau d'une rivière.**

Une des entrées de ce modèle est le **coefficient de Strickler** K_s, qui représente le **degré de rugosité du fond de la rivière.**

Sur la rivière que l'on étudie, on sait que K_s **varie entre** 20 **et** 50 (calibration, dires d'experts...). Or, dû au changement climatique, **l'incertitude sur** K_s **augmente**, et on suspecte qu'il risque de varier dans une **plage de valeurs plus grande**.

Robustesse aux perturbations

Étant donné des perturbations sur les entrées d'un modèle, comment celui-ci va-t-il se comporter ?

Prenons un exemple :

Supposons que nous ayons a disposition un modèle (IA ou non) qui permet de **prédire le niveau de l'eau d'une rivière.**

Une des entrées de ce modèle est le **coefficient de Strickler** K_s, qui représente le **degré de rugosité du fond de la rivière.**

Sur la rivière que l'on étudie, on sait que K_s **varie entre** 20 **et** 50 (calibration, dires d'experts...). Or, dû au changement climatique, **l'incertitude sur** K_s **augmente**, et on suspecte qu'il risque de varier dans une **plage de valeurs plus grande**.

Question :

Quel va être l'impact de cette perturbation sur les prédictions du niveau de l'eau ?

Formalisation du problème

Soit $X \sim P$ une entrée initiale (par ex, K_s). On va chercher une distribution Q telle que :

$$Q \in \underset{P'}{\operatorname{argmin}} \quad \mathcal{D}\left(P, P'\right)$$
$$\text{s.t.} \quad P' \in \mathcal{C}$$

où $\mathcal{C}$ est un **ensemble de distributions** qui respecte des contraintes, et $\mathcal{D}$ est une discrépance entre mesures de probabilités.

Formalisation du problème

Soit $X \sim P$ une entrée initiale (par ex, K_s). On va chercher une distribution Q telle que :

$$Q \in \underset{P'}{\operatorname{argmin}} \quad \mathcal{D}\left(P, P'\right)$$
$$\text{s.t.} \quad P' \in \mathcal{C}$$

où $\mathcal{C}$ est un **ensemble de distributions** qui respecte des contraintes, et $\mathcal{D}$ est une discrépance entre mesures de probabilités.

Idée générale :

- A partir de Q, on construit des **entrées perturbée** $\widetilde{X} \sim Q$.
- Pour un modèle boîte-noire G, on **étudie les différences entre** $G(X)$ **et** $G(\widetilde{X})$.

Comment définir les perturbation ?

Contraintes :

Forcer des quantités statistiques des entrées à prendre des valeurs particulières :

- À but **prospectif** (i.e., anticiper des changements)
- À but **inspectif/exploratoire** (i.e., capacité de généralisation)
- Pour injecter de la **connaissance experte**.

Comment définir les perturbation ?

Contraintes :

Forcer des quantités statistiques des entrées à prendre des valeurs particulières :

- À but **prospectif** (i.e., anticiper des changements)
- À but **inspectif/exploratoire** (i.e., capacité de généralisation)
- Pour injecter de la **connaissance experte**.

Inspiration : **Perturbed Law Indices** (PLI) (Lemaître et al. 2015) et plus généralement les **perturbations entropiques** (Bachoc et al. 2020).

- **Choix de discrépance $\mathcal{D}$:** Divergence de Kullback-Leibler
- **Type de contraintes :** Moments généralisés (moyenne, variance...)

Les PLI ont été utilisés avec succès lors d'études de sûreté pour la quantification des marges en situation accidentelle remis à l'Autorité de Sûreté Nucléaire.

Notre proposition

Ce que l'on a étudié :

- **Choix de discrépance :** Distance de Wasserstein (transport optimal).
- **Type de contraintes :** Quantiles.

Notre proposition

Ce que l'on a étudié :

- **Choix de discrépance :** Distance de Wasserstein (transport optimal).
- **Type de contraintes :** Quantiles.

Pourquoi la distance de Wasserstein ?

- Métrique **intuitive**.
- Comparer un **plus grand ensemble de mesures de probabilité**.
- Produire **de nouvelles observations perturbées**.
- **Conserve la structure de dépendance** des entrées

Pourquoi les quantiles ?

- **Existent toujours**.
- Permettent de **contrôler le support**.
- Ont souvent **un sens pratique** (Value at Risk...).

Notre proposition

Ce que l'on a étudié :

- **Choix de discrépance :** Distance de Wasserstein (transport optimal).
- **Type de contraintes :** Quantiles.

Pourquoi la distance de Wasserstein ?

- Métrique **intuitive**.
- Comparer un **plus grand ensemble de mesures de probabilité**.
- Produire **de nouvelles observations perturbées**.
- **Conserve la structure de dépendance** des entrées

Pourquoi les quantiles ?

- **Existent toujours**.
- Permettent de **contrôler le support**.
- Ont souvent **un sens pratique** (Value at Risk...).

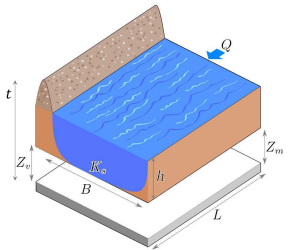
Tout en s'inscrivant dans le même cadre théorique et méthodologique solide.

Illustration : Validation d'un méta-modèle d'IA

Modèle physique (simplifié) : calcul du niveau d'eau dans une rivière (looss et Lemaître 2015).

$$Y = Z_v + \left(\frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)^{3/5}$$

- Q : River maximum annual water flow rate.
- K_s : Strickler riverbed roughness coefficient.
- Z_v : Downstream river level.
- Z_m : Upstream river level.
- L : River length.
- B : River width.



Structure probabiliste :

Entrée	Distribution	Domaine
Q	$\mathcal{G}(1013, 558)$ trunc.	$[500, 3000]$
K_s	$\mathcal{N}(30, 7)$ trunc.	$[20, 50]$
Z_v	$\mathcal{T}(49, 50, 51)$	$[49, 51]$
Z_m	$\mathcal{T}(54, 55, 56)$	$[54, 56]$
L	$\mathcal{T}(4990, 5000, 5010)$	$[4990, 5010]$
B	$\mathcal{T}(295, 300, 305)$	$[295, 305]$

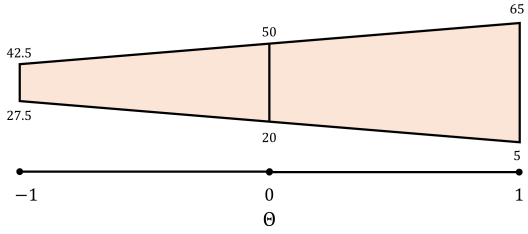
On se propose de perturber le coefficient de rugosité K_s.

Objectif prospectif : *Si le support de K_s augmente, comment se comportera notre modèle ?*

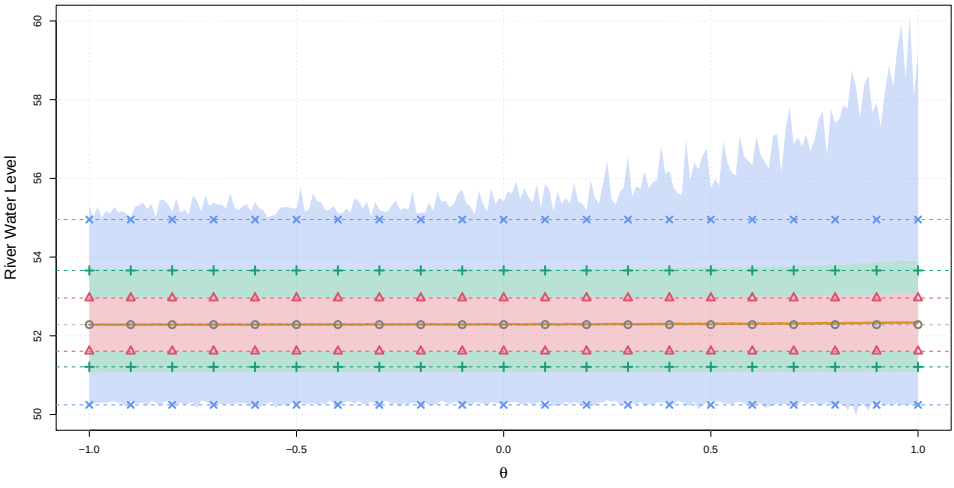
Stratégie : Perturber le **domaine d'application** de K_s.

Coefficient d'intensité :

- $\theta = -1$: entre une rivière naturelle et peu de végétation.
- $\theta = 0$: pas de modification.
- $\theta = 1$: entre des algues prolifiques et un fond en béton.



Effets des perturbations sur le modèle numérique



Méta-modélisation, et validation d'IA

Objectif exploratoire : *Si je remplace mon modèle numérique par une IA, est-ce que ce comportement persiste ?*

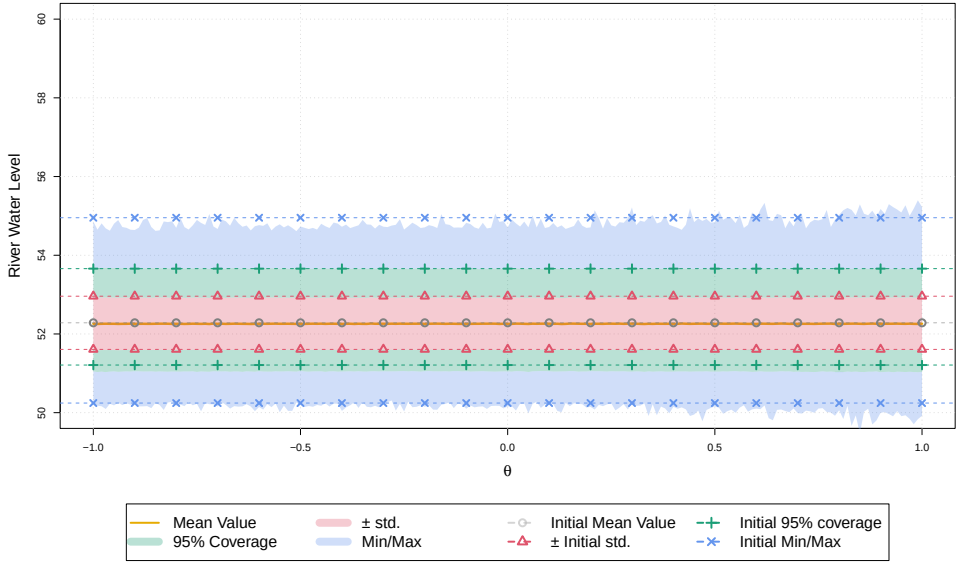
Méta-modèle : Réseau de neurone à 3 couches.

Volumétrie :

- **Entraînement :** 500.000 (non perturbées)
- **Validation :** 50.000 (non perturbées)
- **Critère de minimisation :** Erreur quadratique moyenne (MSE)

Données	R^2	Perte (MSE)
Entraînement	99.5%	0.0119
Validation	99.5%	0.0120

Effets des perturbations sur le réseau de neurones



Robustesse des IA : Conclusion

- Méthode de perturbation **générique** et **ancrée dans la théorie**.
- Méthode à la jonction de la **quantification d'incertitudes** et l'**XAI** (dataset ou densités).
- Permet la mise en évidence de **différences de comportements** pour la méta-modélisation.
- Permet de **prévenir le changement** et de **prendre en compte l'expertise métier**.
- Permet d'aller **plus loin que les métriques de validation usuelles.**

Trois chantiers

Formalisation générale de l'interprétabilité des modèles boîtes-noires

Vision unificatrice *de la* ***quantification des incertitudes des modèles numériques****, et de* ***l'interprétation des IA****.*

Robustesse des IA aux perturbations

Comment va se comporter un modèle boîte-noire face à des perturbations sur ses entrées ?

Quantification d'influence en situation d'entrées dépendantes

Quelles sont les entrées les plus influentes/importantes pour une quantité d'intérêt reliée à un modèle boîte-noire ?

Quantification de l'influence

*Quelles sont les entrées qui **influent** le plus sur une **quantité d'intérêt** reliée à un modèle boîte-noire ?*

Quantification de l'influence

*Quelles sont les entrées qui **influent** le plus sur une **quantité d'intérêt** reliée à un modèle boîte-noire ?*

- **Quantité d'intérêt (QoI) :** Indicateur de performance, variance, prédiction...
- **Influence :** Décomposer la QoI pour attribuer une part à chaque entrées/ensemble d'entrées.

Quantification de l'influence

*Quelles sont les entrées qui **influent** le plus sur une **quantité d'intérêt** reliée à un modèle boîte-noire ?*

- **Quantité d'intérêt (QoI) :** Indicateur de performance, variance, prédiction...
- **Influence :** Décomposer la QoI pour attribuer une part à chaque entrées/ensemble d'entrées.

Pourquoi ?

- Vérifier et valider la **cohérence du modèle avec l'attendu pratique.**
- Sélectionner les entrées **avec des arguments tangibles.**
- S'assurer de la **non-utilisation d'entrées protégées/sensibles.**
- Mieux comprendre les **rouages internes de la boîte-noire.**
- Prioriser l'acquisition de **certaines données pertinentes.**
- Aide au **design de l'IA** (bugs, dégénérescence...).

Dans la littérature

Analyse de sensibilité :

- Importance = Décomposition de la variance $\mathbb{V}(G(X))$.
- Passe par une **décomposition de** $G(X)$ (décomposition d'Hoeffding (1948) - FANOVA).
- **Nécéssite que les entrées soient mutuellement indépendantes**.

Dans la littérature

Analyse de sensibilité :

- Importance = Décomposition de la variance $\mathbb{V}(G(X))$.
- Passe par une **décomposition de** $G(X)$ (décomposition d'Hoeffding (1948) - FANOVA).
- **Nécéssite que les entrées soient mutuellement indépendantes**.

XAI :

- Beaucoup de méthodes cherchent à décomposer une **prédiction**.
- Repose sur une analogie avec la **théorie des jeux coopératifs**.
- **Ne nécéssite pas que les entrées soient mutuellement indépendantes**.

Dans la littérature

Analyse de sensibilité :

- Importance = Décomposition de la variance $\mathbb{V}(G(X))$.
- Passe par une **décomposition de** $G(X)$ (décomposition d'Hoeffding (1948) - FANOVA).
- **Nécéssite que les entrées soient mutuellement indépendantes**.

XAI :

- Beaucoup de méthodes cherchent à décomposer une **prédiction**.
- Repose sur une analogie avec la **théorie des jeux coopératifs**.
- **Ne nécéssite pas que les entrées soient mutuellement indépendantes**.

Owen (2014) : Utiliser la **théorie des jeux coopératifs** pour la **quantification de l'importance** avec **entrées dépendantes**.

“Théorie des jeux coopératifs = Partage de gâteau”

“Théorie des jeux coopératifs = Partage de gâteau”

Deux ingrédients :

- Un **ensemble de joueurs** $D = \{1, \dots, d\}$, et l'ensemble des **coalitions de joueurs** $\mathcal{P}_D$.
- Une **fonction de valeur** v, qui associe à **chaque coalition** une **quantité produite**.

En collaborant tous ensemble, **les joueurs produisent la quantité** $v(D)$ **= le gâteau.**

“Théorie des jeux coopératifs = Partage de gâteau”

Deux ingrédients :

- Un **ensemble de joueurs** $D = \{1, \dots, d\}$, et l'ensemble des **coalitions de joueurs** $\mathcal{P}_D$.
- Une **fonction de valeur** v, qui associe à **chaque coalition** une **quantité produite**.

En collaborant tous ensemble, **les joueurs produisent la quantité $v(D)$ = le gâteau.**

Grande question :

Comment redistribuer le gâteau $v(D)$ auprès des d joueurs ?

Allocations

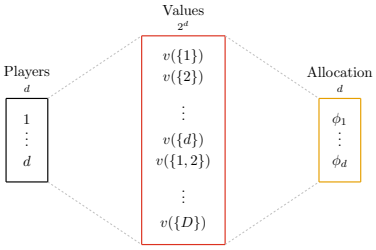
Règle d'allocation : Description du partage du gateau.

Règle d'allocation : Description du partage du gateau.

Elles doivent respecter la propriété suivant :

L'entièreté du gateau et **seulement le gateau** doit être partagé (**Efficacité, positivité**).

Mais comment faire ?

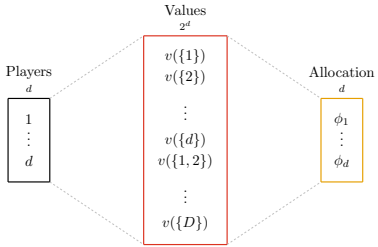


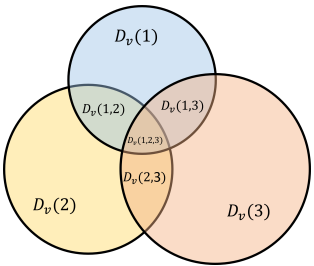
Règle d'allocation : Description du partage du gateau.

Elles doivent respecter la propriété suivant :

L'entièreté du gateau et **seulement le gateau** doit être partagé (**Efficacité, positivité**).

Mais comment faire ?





Les **dividendes d'Harsanyi (1963)** des coalitions :

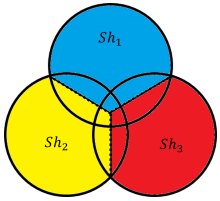
$$\mathcal{D}_v(A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} v(B)$$

permettent, de **manière mécanique**, de découper le gâteau en fonction des **coalitions**.

Répartition égalitaire : les valeurs de Shapley

Les **valeurs de Shapley (1951)** sont une **répartition égalitaire des dividendes**. Pour un joueur $i \in D$:

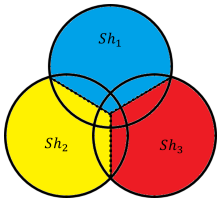
$$\mathrm{Sh}_i = \sum_{A \in \mathcal{P}_D : i \in A} \frac{\mathcal{D}_v(A)}{|A|}.$$



Répartition égalitaire : les valeurs de Shapley

Les **valeurs de Shapley (1951)** sont une **répartition égalitaire des dividendes**. Pour un joueur $i \in D$:

$$\mathrm{Sh}_i = \sum_{A \in \mathcal{P}_D : i \in A} \frac{\mathcal{D}_v(A)}{|A|}.$$



Owen (2014) a proposé d'utiliser ces quantités pour décomposer $\mathbb{V}(G(X))$, où les **entrées dépendantes** X sont les joueurs, et G est un modèle boîte-noire, avec le choix suivant de **fonction de valeur** :

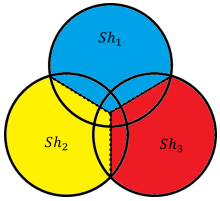
$$v(A) = \mathbb{V}(\mathbb{E}[G(X) \mid X_A]).$$

Ce qui permet de décomposer $v(D) = \mathbb{V}(G(X))$, **la quantité totale d'incertitudes du modèle**, pour chaque **entrée du modèle**. Ce sont les "**effets de Shapley**".

Répartition égalitaire : les valeurs de Shapley

Les **valeurs de Shapley (1951)** sont une **répartition égalitaire des dividendes**. Pour un joueur $i \in D$:

$$\mathrm{Sh}_i = \sum_{A \in \mathcal{P}_D : i \in A} \frac{\mathcal{D}_v(A)}{|A|}.$$



Owen (2014) a proposé d'utiliser ces quantités pour décomposer $\mathbb{V}(G(X))$, où les **entrées dépendantes** X sont les joueurs, et G est un modèle boîte-noire, avec le choix suivant de **fonction de valeur** :

$$v(A) = \mathbb{V}(\mathbb{E}[G(X) \mid X_A]).$$

Ce qui permet de décomposer $v(D) = \mathbb{V}(G(X))$, **la quantité totale d'incertitudes du modèle**, pour chaque **entrée du modèle**. Ce sont les "**effets de Shapley**".

SHAP (Lundberg et Lee 2017) : $v(A) = \mathbb{E}[G(X) \mid X_A = x_A]$, et le gâteau devient une prédiction $v(D) = G(x)$.

Blague de Shapley et les valeurs proportionnelles

Mais, les effets de Shapley présentent un défaut...

Blague de Shapley et les valeurs proportionnelles

Mais, les effets de Shapley présentent un défaut...

Une variable qui n'est **pas dans le modèle**, mais qui est **corrélée aux autres**, peut **se voir octroyer de l'importance**.

C'est la **blague de Shapley**

$$G(X) = X_1 + X_2, \quad X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix} \right)$$

$$\mathrm{Sh}_1 = 0.5 - \rho^2/4, \quad \mathrm{Sh}_2 = 0.5, \quad \mathrm{Sh}_3 = \rho^2/4.$$

Blague de Shapley et les valeurs proportionnelles

Mais, les effets de Shapley présentent un défaut...

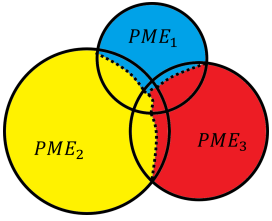
Une variable qui n'est **pas dans le modèle**, mais qui est **corrélée aux autres**, peut **se voir octroyer de l'importance**.

C'est la **blague de Shapley**

$$G(X) = X_1 + X_2, \quad X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix} \right)$$

$$\mathrm{Sh}_1 = 0.5 - \rho^2/4, \quad \mathrm{Sh}_2 = 0.5, \quad \mathrm{Sh}_3 = \rho^2/4.$$

Les **effets proportionnels marginaux (PME)** (Herin et al. 2022) sont une allocation qui se base sur **un partage proportionnel des dividendes d'Harsanyi**. Avec la **garantie théorique** :



Proposition *(Détection de l'exogénéité (Herin et al. 2022)).*

$$PME_i = 0 \iff X_i \text{ n'est pas dans le modèle.}$$

De plus, elles permettent de **mieux discriminer entre les entrées** en situation de **forte corrélation**.

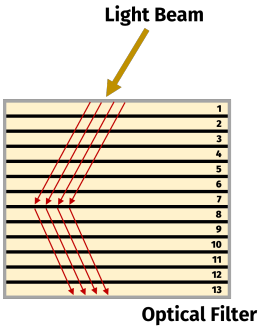
Étude d'un filtre optique

On veut étudier la **performance de transmission** d'un **filtre optique** composé de 13 couches consécutives (Vasseur et al. 2010).

Les entrées $I_1, \ldots, I_{13}$ représentent **des (petites) perturbations (erreurs)** sur les **indices de réfraction de chacun des filtres** ($\mathcal{U}([-0.05, 0.05])$)

Ces entrées sont **(très) corrélées** dû au processus de fabrication (corrélations de 0.95).

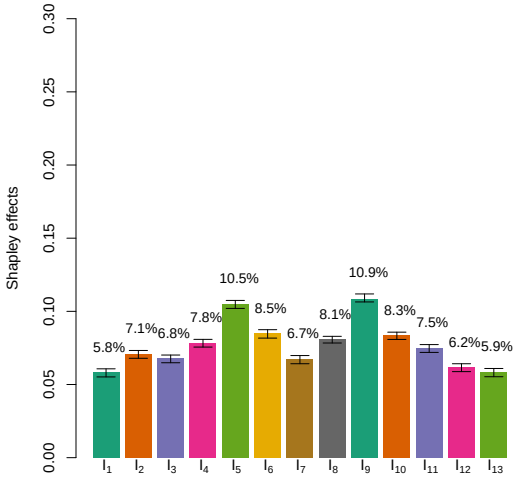
Le modèle (numérique) boîte-noire calcule l'**erreur de transmission globale du filtre**.

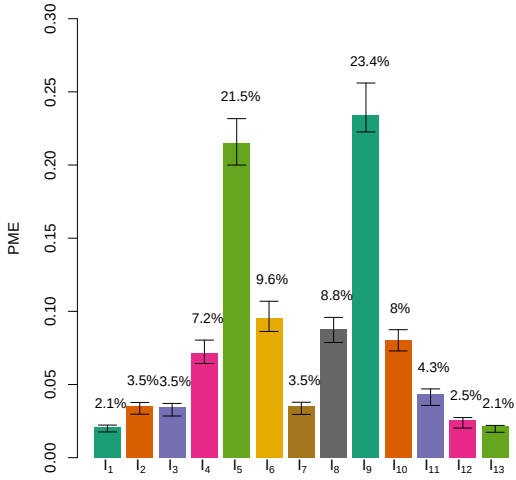


Contrainte : On a seulement accès à un jeu d'entrées-sorties de taille $n = 1000$.

On peut **estimer** les effets de Shapley et les PME via une approche **"plus-proches-voisins"** (Broto, Bachoc et Depecker 2020) via notre implémentation dans le package R `sensitivity`.

Étude d'un filtre optique : Quantification de l'importance





Étude d'un filtre optique - Méta-modélisation et sélection de variables

Scénario : Entraînement d'un méta-modèle (processus Gaussien*) pour remplacer ce modèle numérique.

Performance avec toutes les features : $Q^2 = 99.48\%$.

Sélection de variables :

- Premier seuil : 2.5% d'importance.
  - **Effets de Shapley :** Pas de retrait.
  - **PME :** I_1 et I_{13} sont retirées, $Q^2 = 99.14\%$.
- Deuxième seuil : 5% d'importance.
  - **Effets de Shapley :** Pas de retrait.
  - **PME :** 7 entrées sont retirées, $Q^2 = 98.79\%$.

* Noyau de covariance Matérn 5/2 et tendance constante.

On a maintenant **deux** manières de quantifier l'importance...

... Mais laquelle est **la** bonne ?

Le problème des jeux coopératifs

On a maintenant **deux** manières de quantifier l'importance...

... Mais laquelle est **la** bonne ?

Choisir une fonction de valeur v $\Longrightarrow$ **Dividendes d'Harsanyi**

Et les allocations sont **une aggrégation de ces dividendes** selon des règles (axiomes...).

Le problème des jeux coopératifs

On a maintenant **deux** manières de quantifier l'importance...

... Mais laquelle est **la** bonne ?

Choisir une fonction de valeur v $\Longrightarrow$ **Dividendes d'Harsanyi**

Et les allocations sont **une aggrégation de ces dividendes** selon des règles (axiomes...).

La question devient :

Comment trouver la bonne fonction de valeur v **?**

Décomposition de la variance avec entrées dépendantes

Trouver la bonne fonction de valeur revient à pouvoir **généraliser la décomposition de** $G(X)$ **(Hoeffding-FANOVA)** pour des **entrées dépendantes.**

Décomposition de la variance avec entrées dépendantes

Trouver la bonne fonction de valeur revient à pouvoir **généraliser la décomposition de** $G(X)$ **(Hoeffding-FANOVA)** pour des **entrées dépendantes.**

Pas mal de monde s'est déjà posé la question (Hooker 2007 ; Kuo et al. 2009 ; Chastaing, Gamboa et Prieur 2012 ; Hart et Gremaud 2018).

Décomposition de la variance avec entrées dépendantes

Trouver la bonne fonction de valeur revient à pouvoir **généraliser la décomposition de** $G(X)$ **(Hoeffding-FANOVA)** pour des **entrées dépendantes.**

Pas mal de monde s'est déjà posé la question (Hooker 2007 ; Kuo et al. 2009 ; Chastaing, Gamboa et Prieur 2012 ; Hart et Gremaud 2018).

Mais personne n'a pu apporter de réponse définitive et suffisamment générale à ce problème...

Décomposition de la variance avec entrées dépendantes

Trouver la bonne fonction de valeur revient à pouvoir **généraliser la décomposition de** $G(X)$ **(Hoeffding-FANOVA)** pour des **entrées dépendantes.**

Pas mal de monde s'est déjà posé la question (Hooker 2007 ; Kuo et al. 2009 ; Chastaing, Gamboa et Prieur 2012 ; Hart et Gremaud 2018).

Mais personne n'a pu apporter de réponse définitive et suffisamment générale à ce problème...

Jusqu'à il y a quelques mois :)

Généralisation de l'ANOVA fonctionnelle

On a pu **montrer** que **cette décomposition est vraie** :

- **Pour n'importe quel type d'entrée** (images, texte, séries temporelles, dataset...).
- **Pour n'importe quel modèle boîte-noire à sortie réelle**.
- Sous **deux hypothèses peu restrictives** :
  1. **Les entrées ne doivent pas être fonction les unes des autres.**
  2. **La dépendance stochastique entre les entrées ne doit pas être parfaite.**

Généralisation de l'ANOVA fonctionnelle

On a pu **montrer** que **cette décomposition est vraie** :

- **Pour n'importe quel type d'entrée** (images, texte, séries temporelles, dataset...).
- **Pour n'importe quel modèle boîte-noire à sortie réelle**.
- Sous **deux hypothèses peu restrictives** :
  1. **Les entrées ne doivent pas être fonction les unes des autres.**
  2. **La dépendance stochastique entre les entrées ne doit pas être parfaite.**

Ce résultat à permis d'avoir une **décomposition de la variance** qui :

- Est **naturelle**, **intuitive**, et **avec des garanties théoriques**.
- **Généralise** les outils **déjà utilisés en GSA pour l'étude des modèles numériques**.
- Est **viable pour les problématiques d'apprentissage automatique**.
- Permet de **séparer les effets dûs à la dépendance**, des **effets d'interaction**.
- **Mais, on ne sait pas (encore) les estimer...**

Généralisation de l'ANOVA fonctionnelle

On a pu **montrer** que **cette décomposition est vraie** :

- **Pour n'importe quel type d'entrée** (images, texte, séries temporelles, dataset...).
- **Pour n'importe quel modèle boîte-noire à sortie réelle**.
- Sous **deux hypothèses peu restrictives** :
  1. **Les entrées ne doivent pas être fonction les unes des autres.**
  2. **La dépendance stochastique entre les entrées ne doit pas être parfaite.**

Ce résultat à permis d'avoir une **décomposition de la variance** qui :

- Est **naturelle**, **intuitive**, et **avec des garanties théoriques**.
- **Généralise** les outils **déjà utilisés en GSA pour l'étude des modèles numériques**.
- Est **viable pour les problématiques d'apprentissage automatique**.
- Permet de **séparer les effets dûs à la dépendance**, des **effets d'interaction**.
- **Mais, on ne sait pas (encore) les estimer...**

Candidat prometteur d'outil pertinent pour quantifier l'importance :)

Quelques messages

En résumé :

- Pour **démocratiser l'utilisation du ML pour des systèmes critiques**, il faut développer des **outils pertinents de validation**, pour **justifier la prise de décision**.

Quelques messages

En résumé :

- Pour **démocratiser l'utilisation du ML pour des systèmes critiques**, il faut développer des **outils pertinents de validation**, pour **justifier la prise de décision**.
- Ces outils doivent **s'inscrire dans la théorie**, car la **pratique n'est pas suffisante**.

Quelques messages

En résumé :

- Pour **démocratiser l'utilisation du ML pour des systèmes critiques**, il faut développer des **outils pertinents de validation**, pour **justifier la prise de décision**.
- Ces outils doivent **s'inscrire dans la théorie**, car la **pratique n'est pas suffisante**.
- Pour la question de la **quantification d'importance** :
  - Les **approches "théorie des jeux"** **peuvent être critiquables**, mais **offrent une réelle solution pratique**.
  - La **généralisation de l'ANOVA fonctionnelle** permet de proposer des **outils candidats à la pertinence**, mais pour l'instant **on ne sait pas (encore) les calculer**.

References i

Bachoc, F., F. Gamboa, M. Halford, J-M. Loubes et L. Risser. 2020. "Explaining Machine Learning Models using Entropic Variable Projection" [en en]. *arXiv :1810.07924 (submitted)* (décembre). http://arxiv.org/abs/1810.07924.

Broto, B., F. Bachoc et M. Depecker. 2020. "Variance Reduction for Estimation of Shapley Effects and Adaptation to Unknown Input Distribution". *SIAM/ASA Journal on Uncertainty Quantification* 8 (2) : 693-716. ISSN : 2166-2525. https://doi.org/10.1137/18M1234631. https://epubs.siam.org/doi/10.1137/18M1234631.

Chastaing, G., F. Gamboa et C. Prieur. 2012. "Generalized Hoeffding-Sobol decomposition for dependent variables - application to sensitivity analysis". Publisher : Institute of Mathematical Statistics and Bernoulli Society, *Electronic Journal of Statistics* 6, nº none (janvier) : 2420-2448. ISSN : 1935-7524, 1935-7524. https://doi.org/10.1214/12-EJS749. https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-6/issue-none/Generalized-Hoeffding-Sobol-decomposition-for-dependent-variables---application/10.1214/12-EJS749.full.

Harsanyi, J. C. 1963. "A Simplified Bargaining Model for the n-Person Cooperative Game". Publisher : [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University], *International Economic Review* 4 (2) : 194-220. ISSN : 0020-6598. https://doi.org/10.2307/2525487. https://www.jstor.org/stable/2525487.

Hart, J., et P. A. Gremaud. 2018. "An approximation theoretic perspective of Sobol' indices with dependent variables" [en English]. Publisher : Begel House Inc. *International Journal for Uncertainty Quantification* 8 (6). ISSN : 2152-5080, 2152-5099. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2018026498. https://www.dl.begellhouse.com/journals/52034eb04b657aea,23dc16a4645b89c9,61d464a51b6bf191.html.

References ii

Herin, Margot, Marouane Il Idrissi, Vincent Chabridon et Bertrand Iooss. 2022. "Proportional marginal effects for global sensitivity analysis". Working paper or preprint, octobre. https://hal.archives-ouvertes.fr/hal-03825935.

Hoeffding, W. 1948. "A Class of Statistics with Asymptotically Normal Distribution". Publisher : Institute of Mathematical Statistics, *The Annals of Mathematical Statistics* 19, n° 3 (septembre) : 293-325. ISSN : 0003-4851, 2168-8990. https://doi.org/10.1214/aoms/1177730196. https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-19/issue-3/A-Class-of-Statistics-with-Asymptotically-Normal-Distribution/10.1214/aoms/1177730196.full.

Hooker, G. 2007. "Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables" [en en]. *Journal of Computational and Graphical Statistics* 16 (3) : 709-732. http://www.jstor.org/stable/27594267.

Iooss, B., et P. Lemaître. 2015. "A Review on Global Sensitivity Analysis Methods". In *Uncertainty Management in Simulation-Optimization of Complex Systems : Algorithms and Applications*, sous la direction de G. Dellino et C. Meloni, 101-122. Springer US. https://doi.org/10.1007/978-1-4899-7547-8_5. https://doi.org/10.1007/978-1-4899-7547-8_5.

Kuo, F. Y., I. H. Sloan, G. W. Wasilkowski et H. Woźniakowski. 2009. "On decompositions of multivariate functions" [en en]. *Mathematics of Computation* 79, n° 270 (novembre) : 953-966. ISSN : 0025-5718. https://doi.org/10.1090/S0025-5718-09-02319-9. http://www.ams.org/journal-getitem?pii=S0025-5718-09-02319-9.

Lemaître, P., E. Sergienko, A. Arnaud, N. Bousquet, F. Gamboa et B. Iooss. 2015. "Density modification-based reliability sensitivity analysis". *Journal of Statistical Computation and Simulation* 85 (6) : 1200-1223. https://doi.org/10.1080/00949655.2013.873039. eprint : https://doi.org/10.1080/00949655.2013.873039. https://doi.org/10.1080/00949655.2013.873039.

References iii

Lundberg, S., et S. Lee. 2017. “A Unified Approach to Interpreting Model Predictions”. Décembre. https://www.researchgate.net/publication/317062430_A_Unified_Approach_to_Interpreting_Model_Predictions.

Owen, Art B. 2014. “Sobol’ Indices and Shapley Value” [en en]. *SIAM/ASA Journal on Uncertainty Quantification* 2, n^{O} 1 (janvier) : 245-251. ISSN : 2166-2525. https://doi.org/10.1137/130936233. http://epubs.siam.org/doi/10.1137/130936233.

Shapley, L. S. 1951. *Notes on the n-Person Game – II : The Value of an n-Person Game* [en English]. Researach Memorandum ATI 210720. Santa Monica, California : RAND Corporation, août. https://www.rand.org/content/dam/rand/pubs/research_memoranda/2008/RM670.pdf.

Vasseur, O., M. Claeys-Bruno, M. Cathelinaud et M. Sergent. 2010. “High-dimensional sensitivity analysis of complex optronic systems by experimental design : applications to the case of the design and the robustness of optical coatings”. *Chinese Optics Letters* 8(s1) : 21-24.

Merci de votre attention !

Des questions ?