Audit de biais algorithmiques

Etude de cas en recouvrement d'impayés

Jean-Michel Loubes, Benoit Rottembourg & Rémy Sourial @ REGALIA

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

Les lundis de l'IA et de la finance

Inria

Le recouvrement d'impayés

Plus de 2% des abonnements chaque mois sont impayés (téléphonie, eau, …)

Pour un seul opérateur téléphonique : plusieurs millions de dossiers impayés et > 100 millions d'euros / an

Un processus de recouvrement est mis en place, avec

- **Des règles métiers ("si récidiviste alors ….", "si montant de la dette > X alors …")**
- **Des algorithmes de scoring plus ou moins innovants (Machine Learning, Multi-stage Optimisation)**
- **Ce qui aboutit nécessairement à des traitements différenciés selon les dossiers (et les clients)**

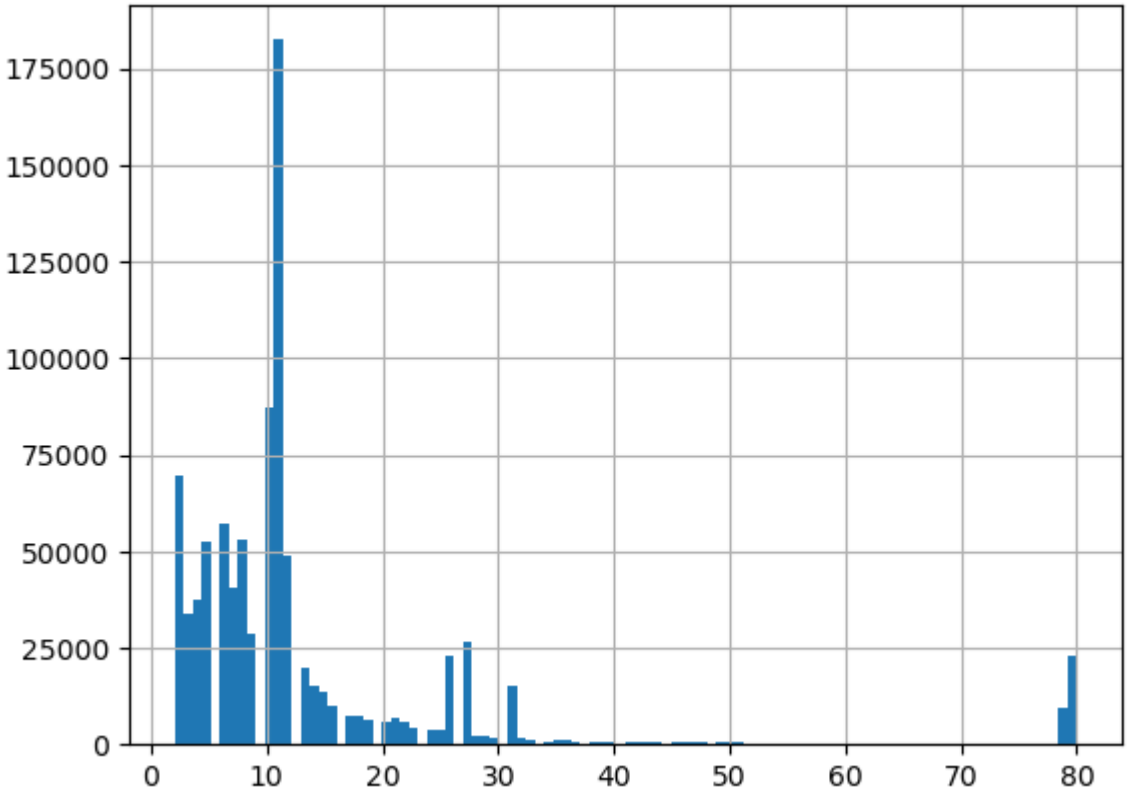
L'algorithme optimise la somme recouvrée (par unité de temps) en minimisant le coût des opérations de recouvrement (mail, SMS, téléphone, huissier).

Il peut profondément faire évoluer la rentabilité de la société de recouvrement (meilleurs achats, faible besoin en fond de roulement)

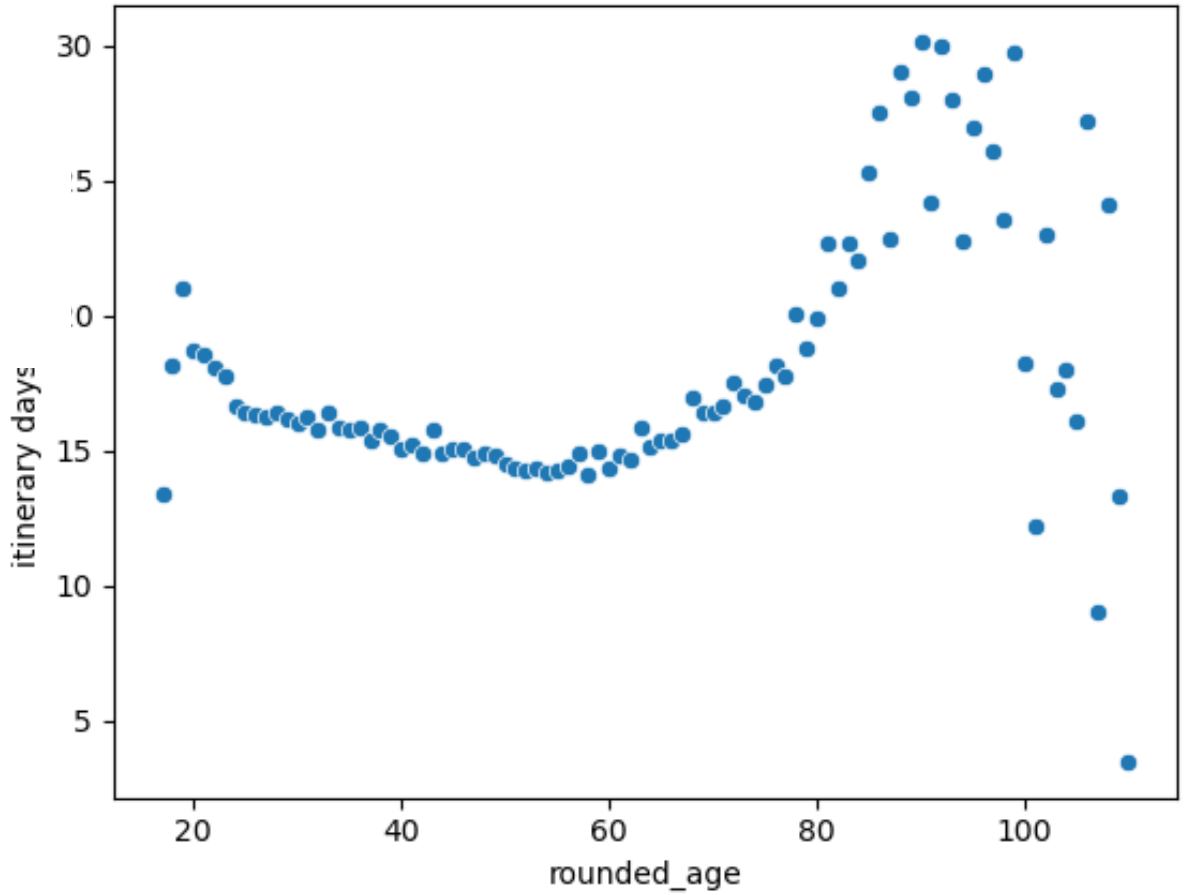
Mais il manipule des données comportementales « potentiellement sensibles »

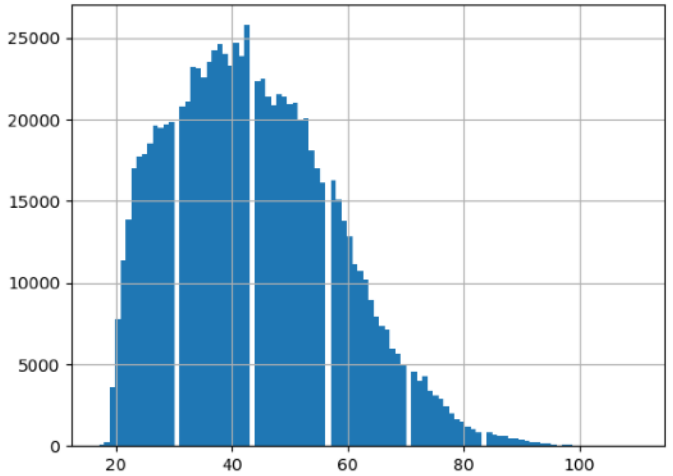
Notre use case (via un opérateur de recouvrement privé en France)

- **900 000 dossiers impayés**
- **Mis en recouvrement en Janvier/Février/Mars 2023**
- **Dossiers tous clos**
- **Temps de recouvrement très variable (de 0 à 80 jours)**
- **Coupure de la ligne à J+10 (c'est radical)**
- **Envoi en société de recouvrement à J+80 (radical aussi, mais autrement)**



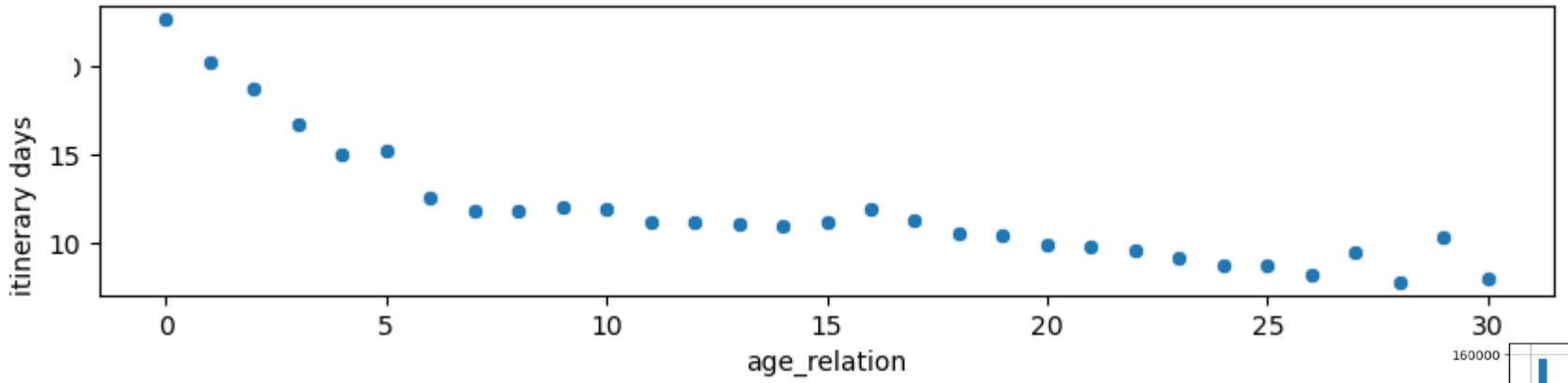
Influence de l’âge sur le temps de recouvrement

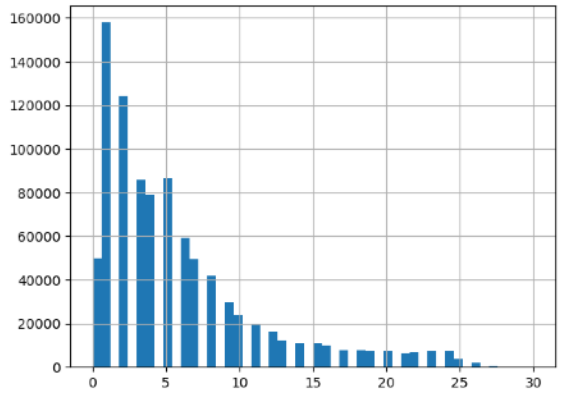




Influence de l’ancienneté sur le temps de recouvrement (évidemment c’est un peu lié à l’âge)

```
[83]:       : xlabel='age_relation', ylabel='itinerary days
```

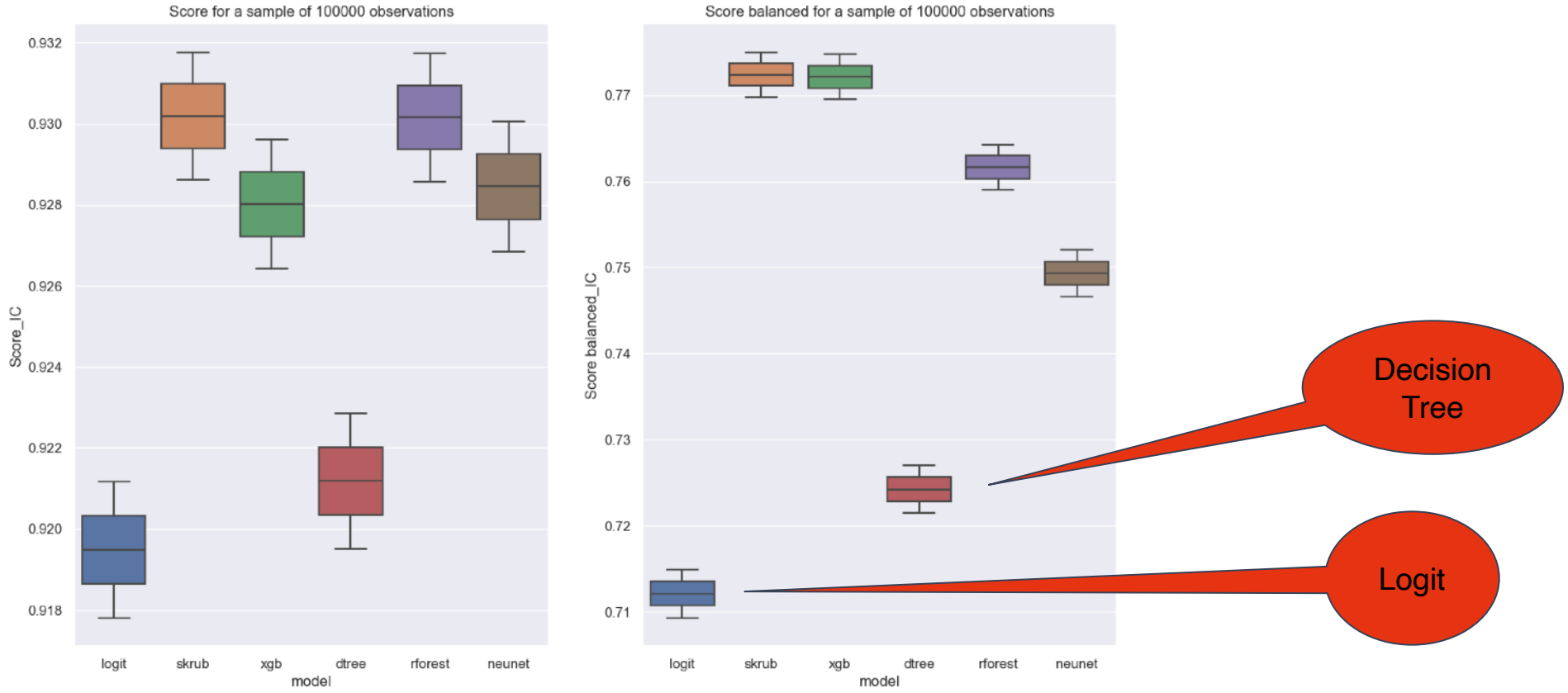




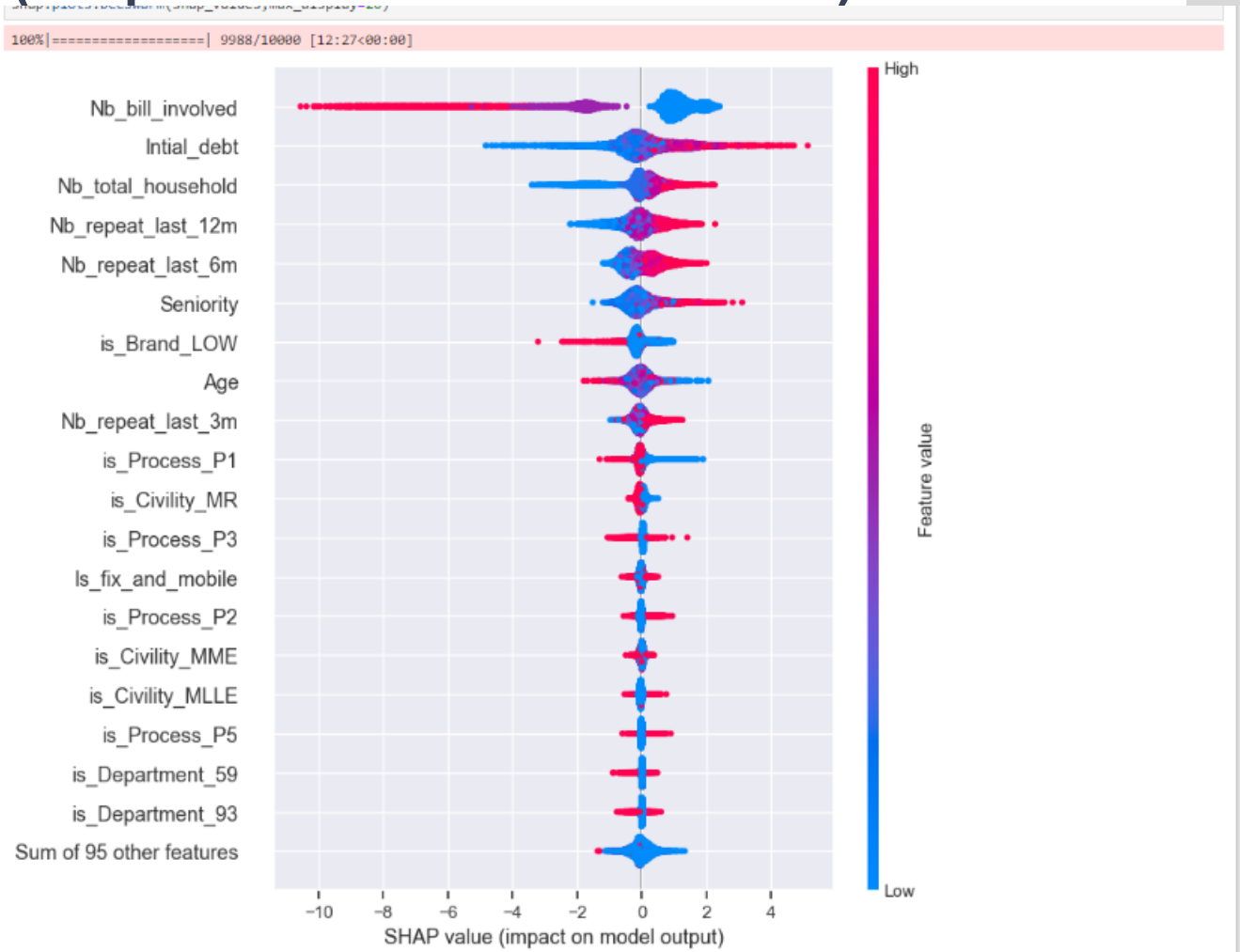
Influence de la civilité et de l'âge sur le temps de recouvrement

		itinerary days in	
		mean	count
LIBL_CIVL	age		
MLLE	20 - 30	12.970968	6510
	30 - 40	12.263838	21138
	40 - 50	11.080750	18613
	50 - 60	10.940803	8666
	< 20	13.923810	105
	>= 60	12.033589	2739
MME	20 - 30	16.610606	69241
	30 - 40	16.351028	84204
	40 - 50	15.433504	91261
	50 - 60	14.420497	75242
	< 20	27.499500	6997
	>= 60	16.746018	59512
MR	20 - 30	17.200119	98921
	30 - 40	16.227183	127043
	40 - 50	15.682902	118550
	50 - 60	15.317217	90342
	< 20	30.953550	10183
	>= 60	18.692519	66957

6 modèles prédictifs à l'étude (va-t-il rembourser avant J20 ?) : les « accuracy » ne se valent pas



Les valeurs de Shapley (impact sur la décision) :

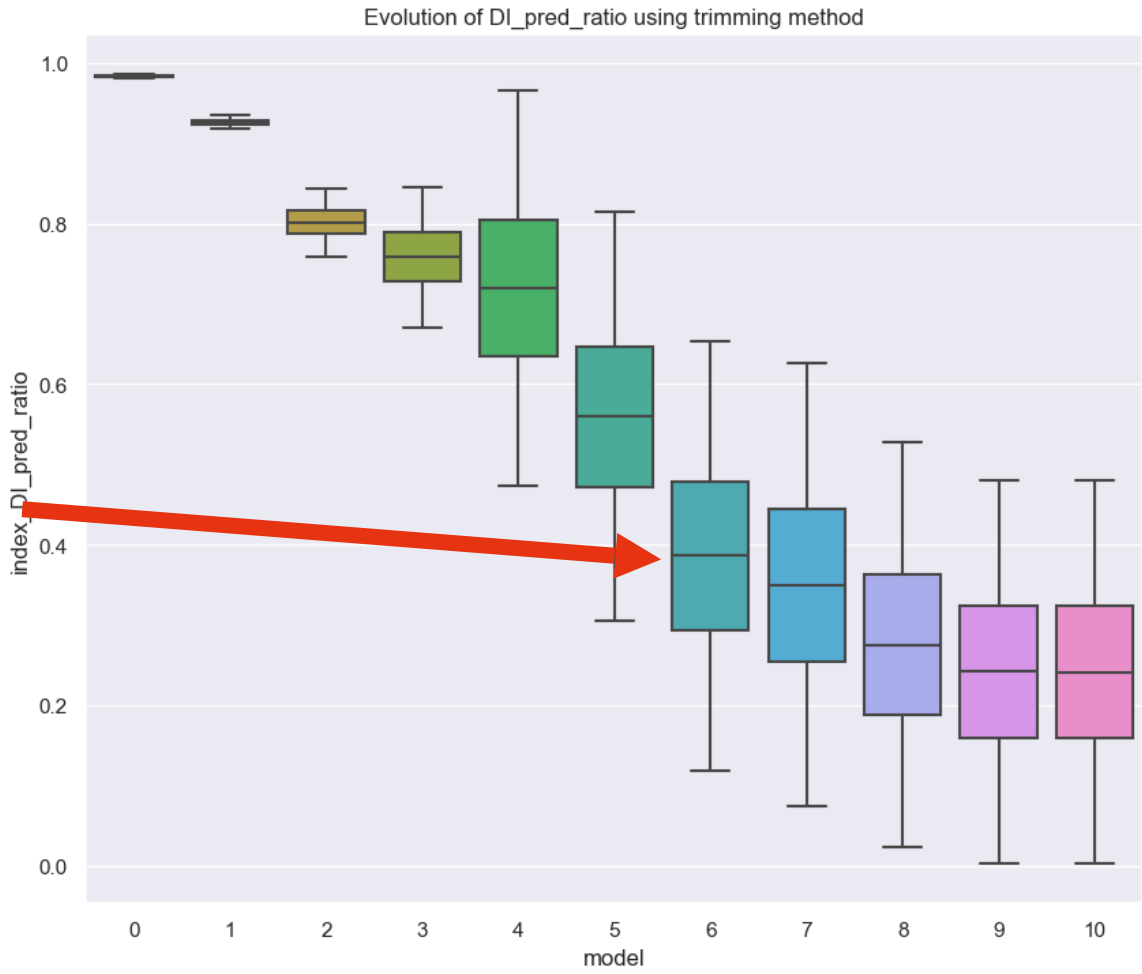


Recherche d’un biais « régional » sur le genre

- **Le disparate impact global sur la variable genre, pour notre meilleur modèle Skrub, est quasi parfait : 0,98 . Il n’y a pas de discrimination sur les 300000 échantillons de test. On pourrait s’arrêter là**

- **Mais en contextualisant, on peut identifier des zones à fort disparate impact , jusqu’à 0,4 (IC : 0,15-0,65), pour près de 3% de la population**

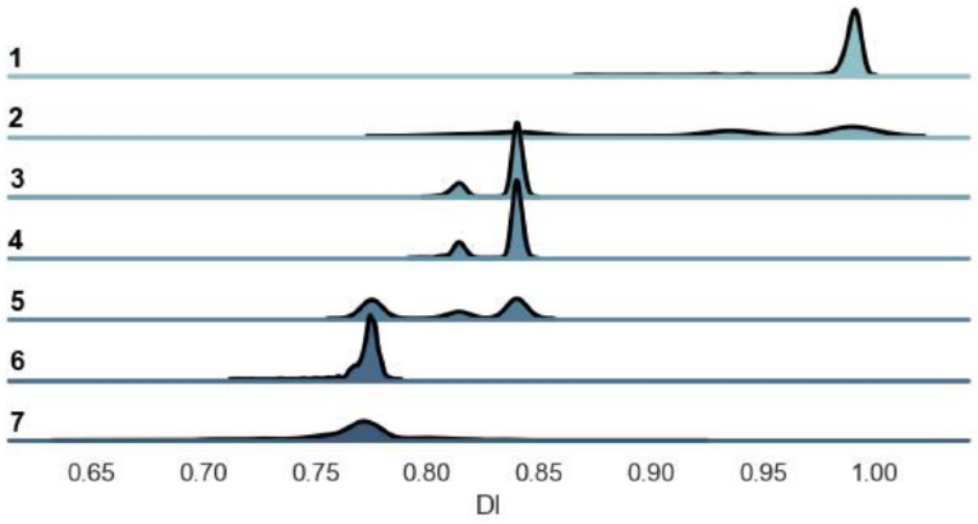
- **Un DI sous la barre de 0,8 est communément considéré comme un biais significatif . Journalistiquement, c’est suffisant.**



Quelle bonne stratégie de recherche adopter ?

- **Le problème de la recherche de la pire zone de biais (union d'intervalles des variables avec le plus faible disparate impact) est évidemment difficile à résoudre de manière générique**

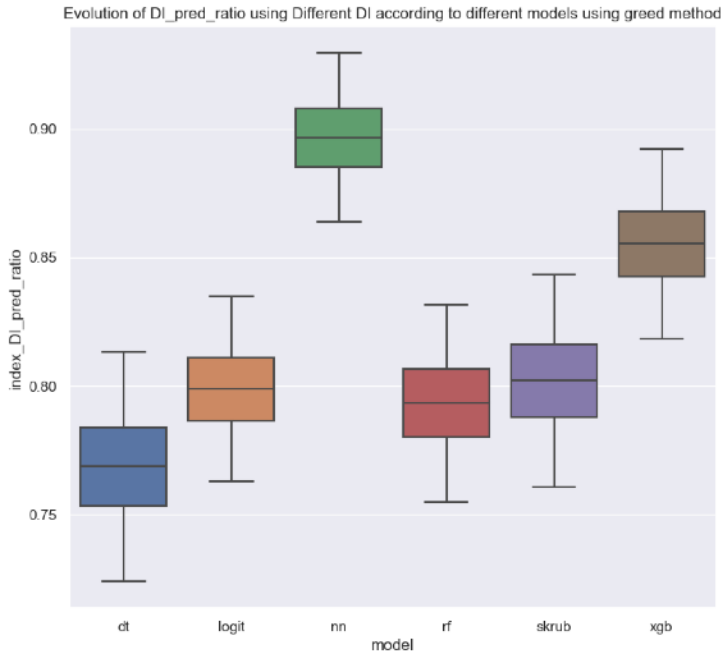
- **Mais, en une dizaine de secondes, sur des bases de données de million de clients, on peut ici identifier des biais sérieux, portant ici sur 3% de la base**
- **Pose à la fois des questions d'exploration d'un espace de grande dimension, de métriques locales de biais qui soient "représentatives" (cf. Jean-Michel Loubes) , et de respect des contraintes d'explicabilité à l'auditeur (un polyèdre facile à raconter)**
- **Questions au coeur de la librairie Regalia que nous voulons faire accoster à Scikit-learn**

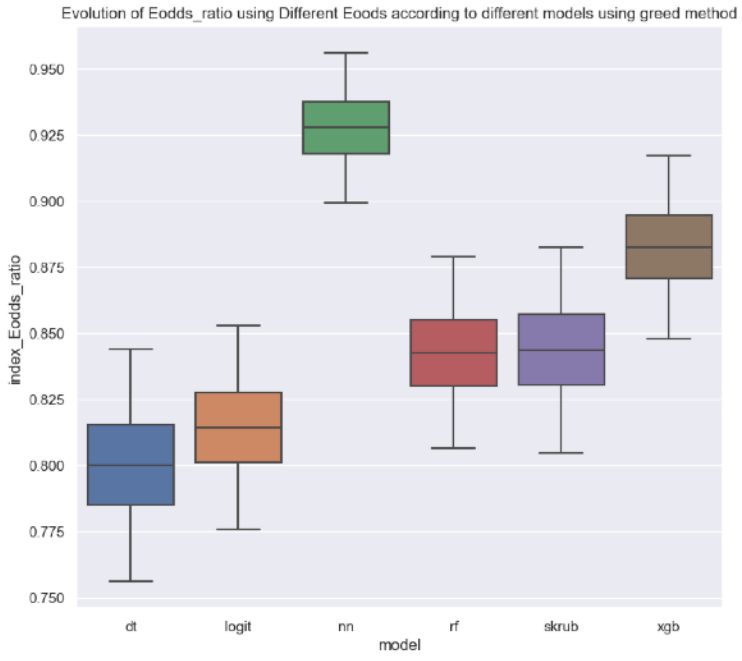


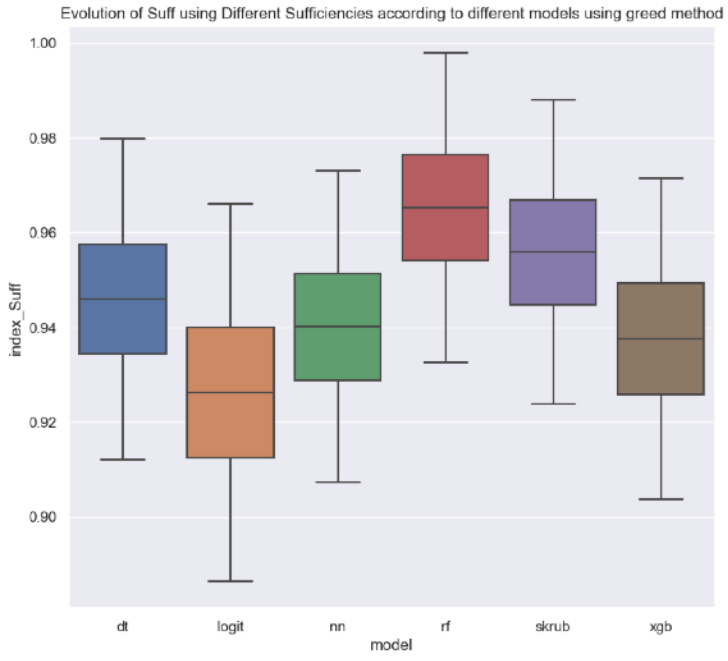
A chacune des 7 itérations j'observe le DI "local" de 1000 boules jetées au hasard dans le dataset, et je garde les meilleures.
Boule ? Norme ? DI local ?

Une fois qu'on a trouvé une « zone biaisée »

- **Dans la zone détectée**
- **Tous les critères de fairness ne sont pas identiques et les modèles performent différemment (ici le NN semble plus équilibré que les autres, sans trop de perte d'accuracy)**
- **On ne trouve pas les mêmes zones selon le critère de fairness retenu, ou selon le modèle choisi**

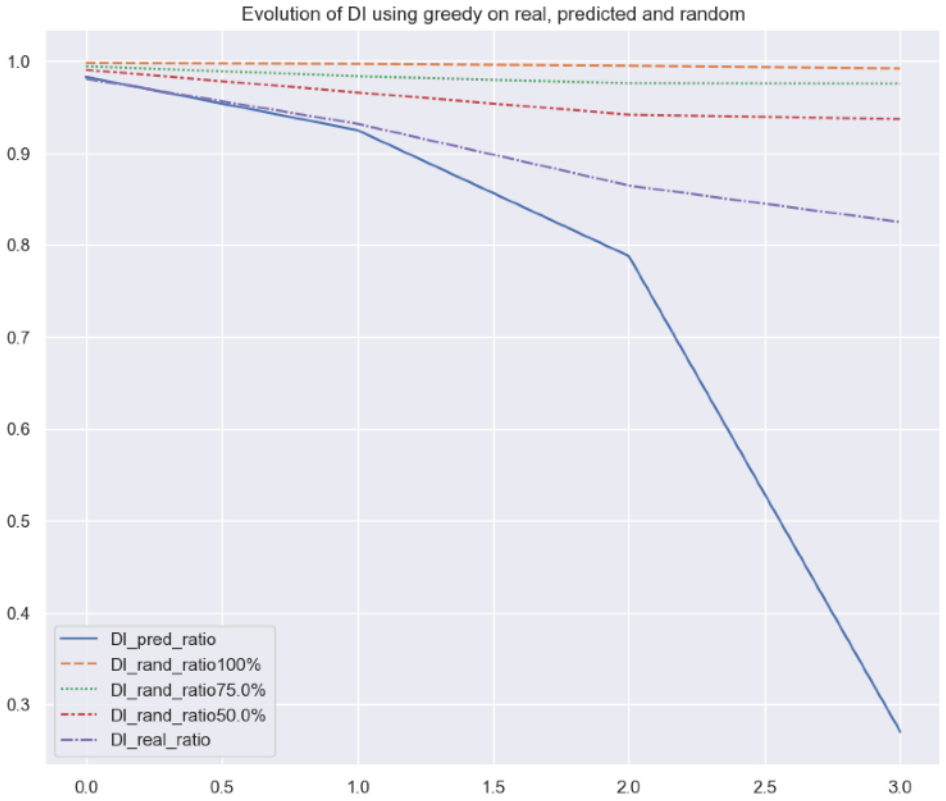






Quelques précautions

- **Vérifier que le hasard n'explique pas à lui seul ce biais**
- **Mesurer l'impact sur la population discriminée (ici les hommes) en évaluant quelle aurait été la décision de l'algorithme si la civilité n'était pas homme mais femme**
- **Mais cela ne suffit pas toujours car « être un homme » ou « être jeune » peut se déduire (ou induire) d'autres variables liées**
- **Par exemple le fait d'être une famille nombreuse est corrélé au nombre de lignes téléphoniques chez l'opérateur, ou que l'ancienneté est corrélée à l'âge**
- **Comparer avec l'approche manuelle préexistante, qui peut-être était elle-même très biaisée (les fameuses « règles métier »)**



Définir le biais pour quantifier et réparer les algorithmes (Depuis 2017 + packages python …)

1. Expliquer la **variabilité d'un comportement moyen** $\mathrm{Var}_A\mathbb{E}(\hat{Y}\,|\,A)$ $\mathrm{Var}_A\mathbb{E}(\ell(\hat{Y},Y)\,|\,A)$

Disparate Treatment $P(\hat{Y}=1\,|\,A=0)\,/\,P(\hat{Y}=1\,|\,A=1)$

Avoiding Disparate Treatment : $P(\hat{Y}=i\,|\,A=0,Y=j)\;-\;P(\hat{Y}\,|\,A=1,Y=j)$. ….

2. Mesurer l'**indépendance** $\mathscr{L}(\hat{Y}(X),A)=\mathscr{L}(\hat{Y}(X))\times\mathscr{L}(A)$

χ^2-test, Mutual Information, Covariance , Hilbert-Schmidt Independence Criterion for RKHS

$\mathscr{C}_{f(X)A}$ $\mathscr{C}_{f(X)A|Y}=\mathscr{C}_{f(XA}-\mathscr{C}_{f(X)Y}\mathscr{C}_{YY}^{-1}\mathscr{C}_{YA}$

3. Mesurer la similarité des **distributions** de l'algorithme ou de son erreur

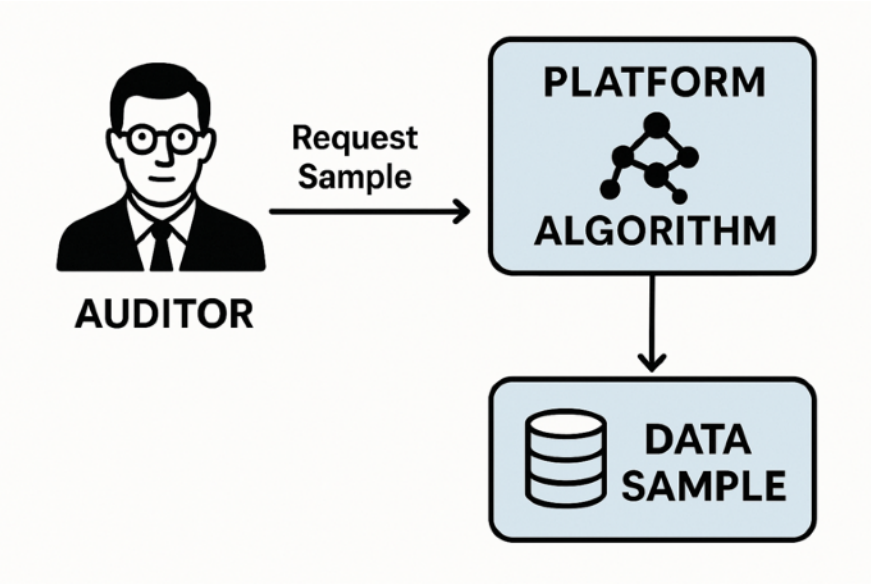
$d_{\mathscr{L}}\left(\mathscr{L}(\hat{Y}(X)\,|\,A=a),\mathscr{L}(\hat{Y}(X)\,|\,A=b)\right)$ avec transport optimal (Monge-Kantorovich a.ka Wasserstein)

D'un point de vue global ou individuel ….

Quel biais est mesuré ? Celui de la distribution ou celui de l'algorithme ? Ou celui de l'algorithme par rapport aux données ?

Cadre habituel pour l'audit d'un SIA : nécessité d'obtenir un échantillon valide et représentatif de données de test du système.

L'auditeur demande / collecte un échantillon de données pour vérifier la conformité du SIA

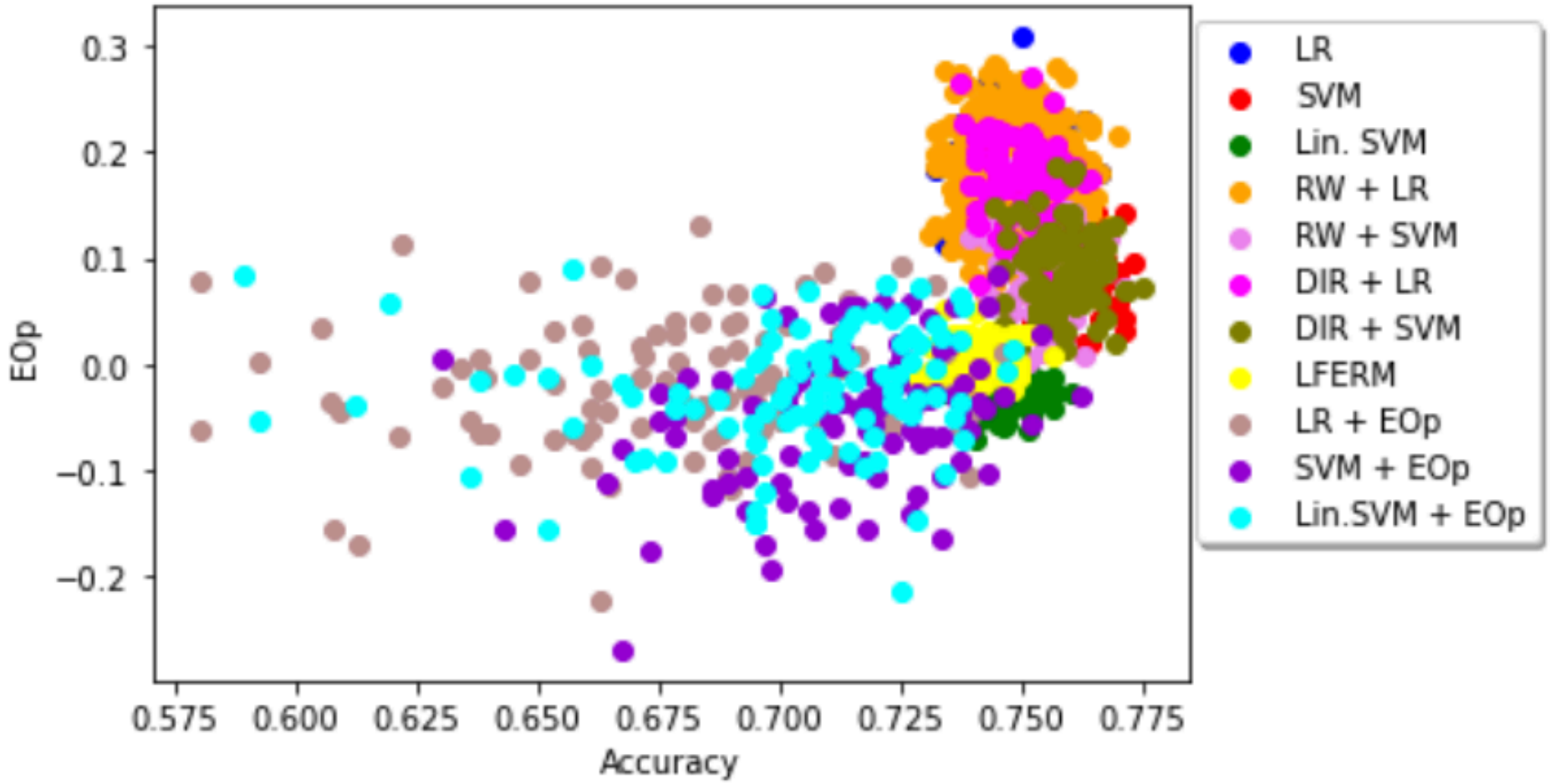


$\mathbb{P}$ Loi des vraies données

$= ?$

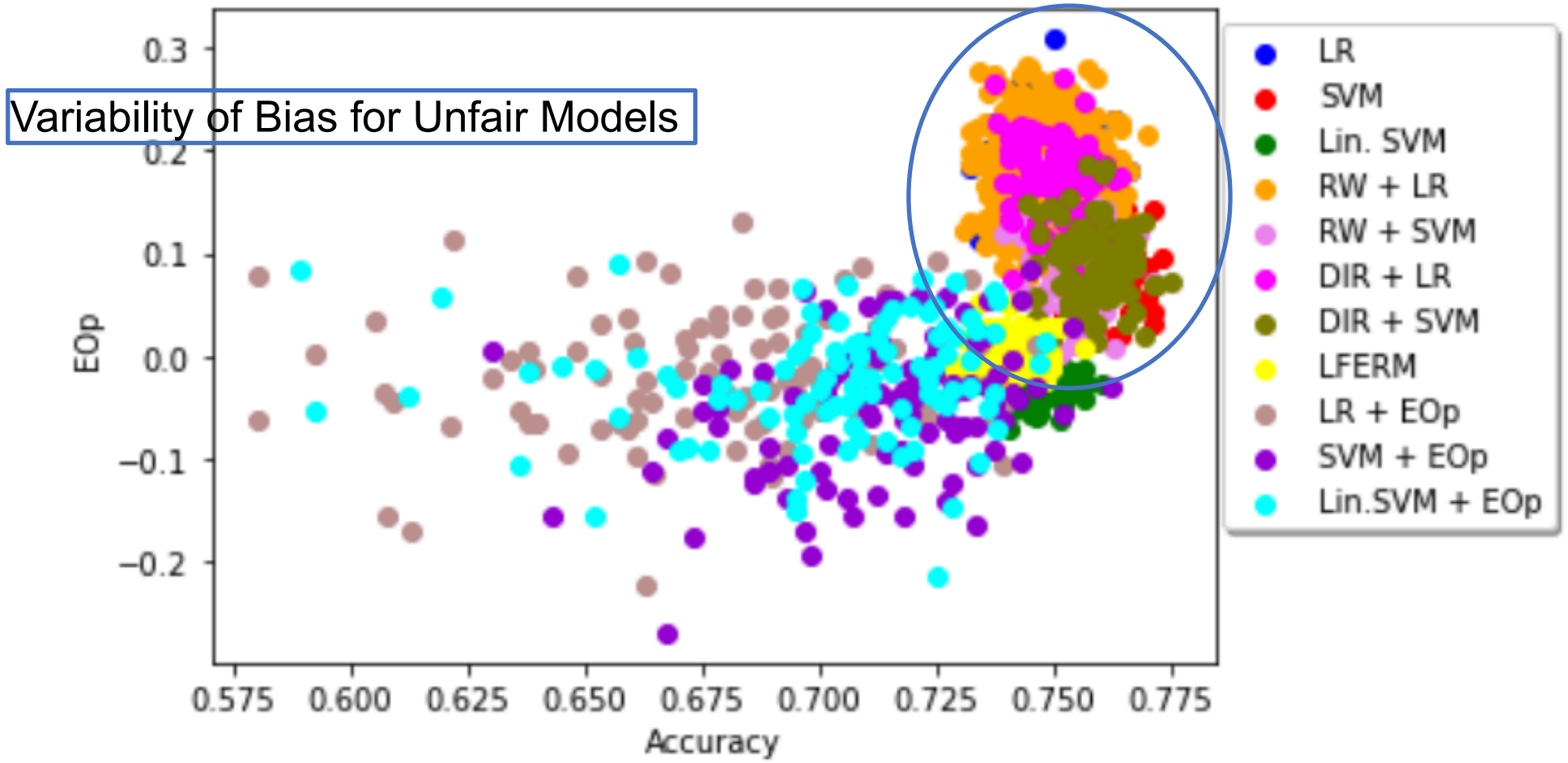
$\mathbb{Q}$ Loi du test

On échantillonne les données sur lesquelles on teste la fairness



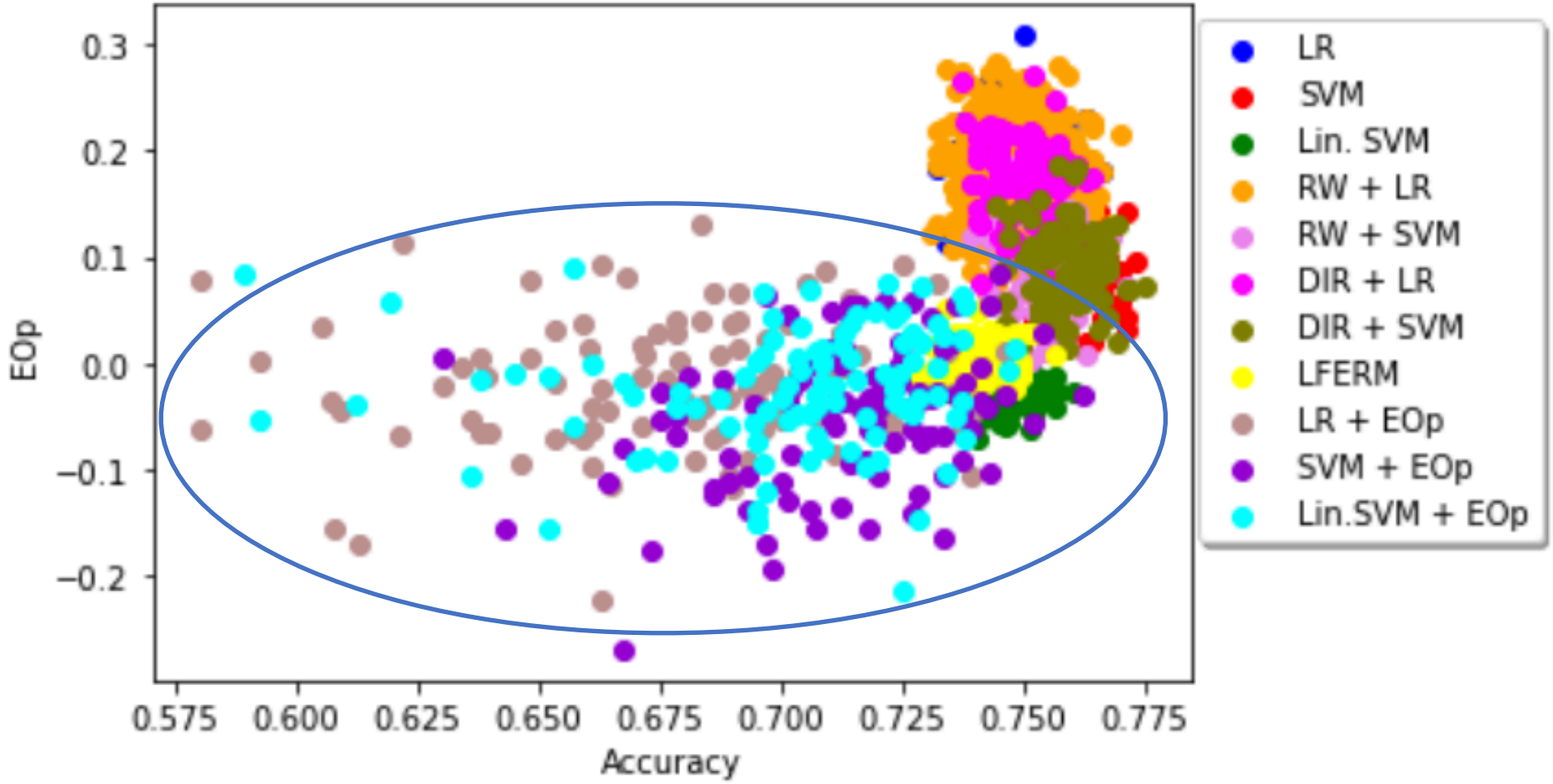
Différents sous-échantillons des données de Test aboutissent à des mesures de biais différentes :

Sous échantillonner ne permet pas de certifier le biais de l'algorithme



Les réparer en utilisant des contraintes par des mesures de biais entraine l'**instabilité des performances des algorithmes**

Variability of Bias & Performance for Mitigated Models



Déloyauté de l'audité.

Adapté de Globally Explaining models under Stress - GEMS-AI https://gems-ai.aniti.fr/

Cadre habituel pour l'audit d'un SIA : nécessité d'obtenir un échantillon valide et représentatif de données de test.

L'auditeur demande un échantillon de données pour vérifier la conformité du SIA

Le data scientist déloyal fournit un échantillon $Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}$

1. Proche de la distribution des données originales P
2. .. Mais construit pour satisfaire une contrainte d'équité c'est-à-dire

$d(Q_n, P) \leq \varepsilon$. but $\mathbb{E}_{Q_n}\Phi(Z) \leq \delta$. où Φ est une mesure de biais global

Des théorèmes pour garantir l'existence d'échantillons vérifiant des contraintes sur les mesures de biais globaux (publiés Information & Inference 2022, soumis ICML 2025, …)

Theorem : Let $\mathbb{P}_{\Phi,t}$ be the set of all probability measures P such that $\mathbb{E}_P\Phi(Z) = t$. Then $Q_t := \operatorname{arginf}_{Q_n \in \mathbb{P}_{\Phi,t}} \mathrm{KL}(P_n, Q_n)$ exists and is uniquely defined by

$$Q_t = \frac{1}{n}\sum_{i=1}^{n} \lambda_i^{(t)} \delta_{X_i, \hat{Y}_i, Y_i},$$ with $\lambda_i^{(t)}$ solutions to a feasible optimization problem.

Idem en remplaçant la distance KL par la distance de Monge-Kantorovich (a.k.a Wasserstein)

$$\arg\min_{P} W_2(Q_n, P) \quad \text{such that} \quad \mathbb{E}_P\Phi(Z) = t.$$

Conclusion : on peut créer un faux échantillon « représentatif » mais fair selon une mesure de biais classique.

Entrée $\mathbb{P}_n$

$\mathbb{P}_n^{\sigma}$ Echantillon

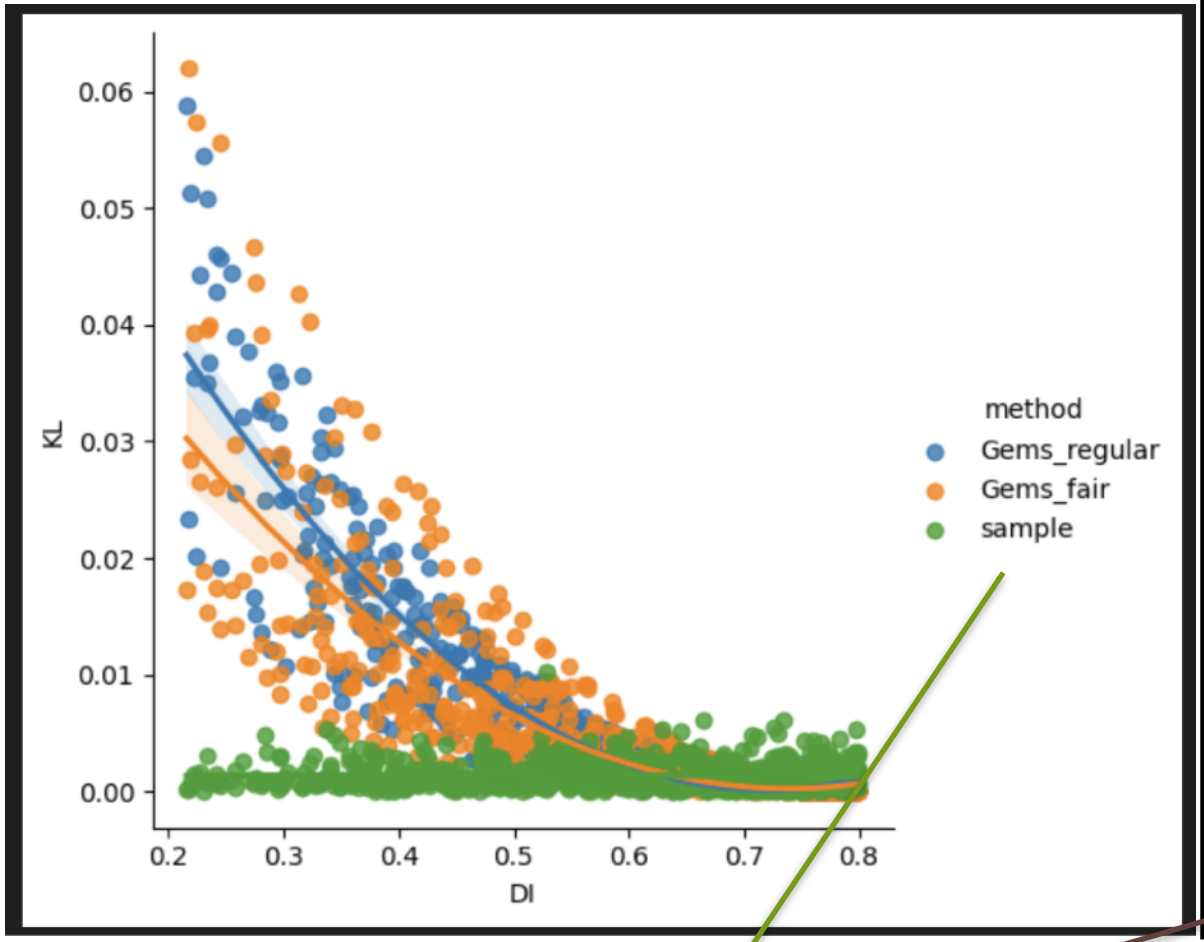
$\mathbb{Q}_n^{\mathrm{Fair}}$ Manipulation

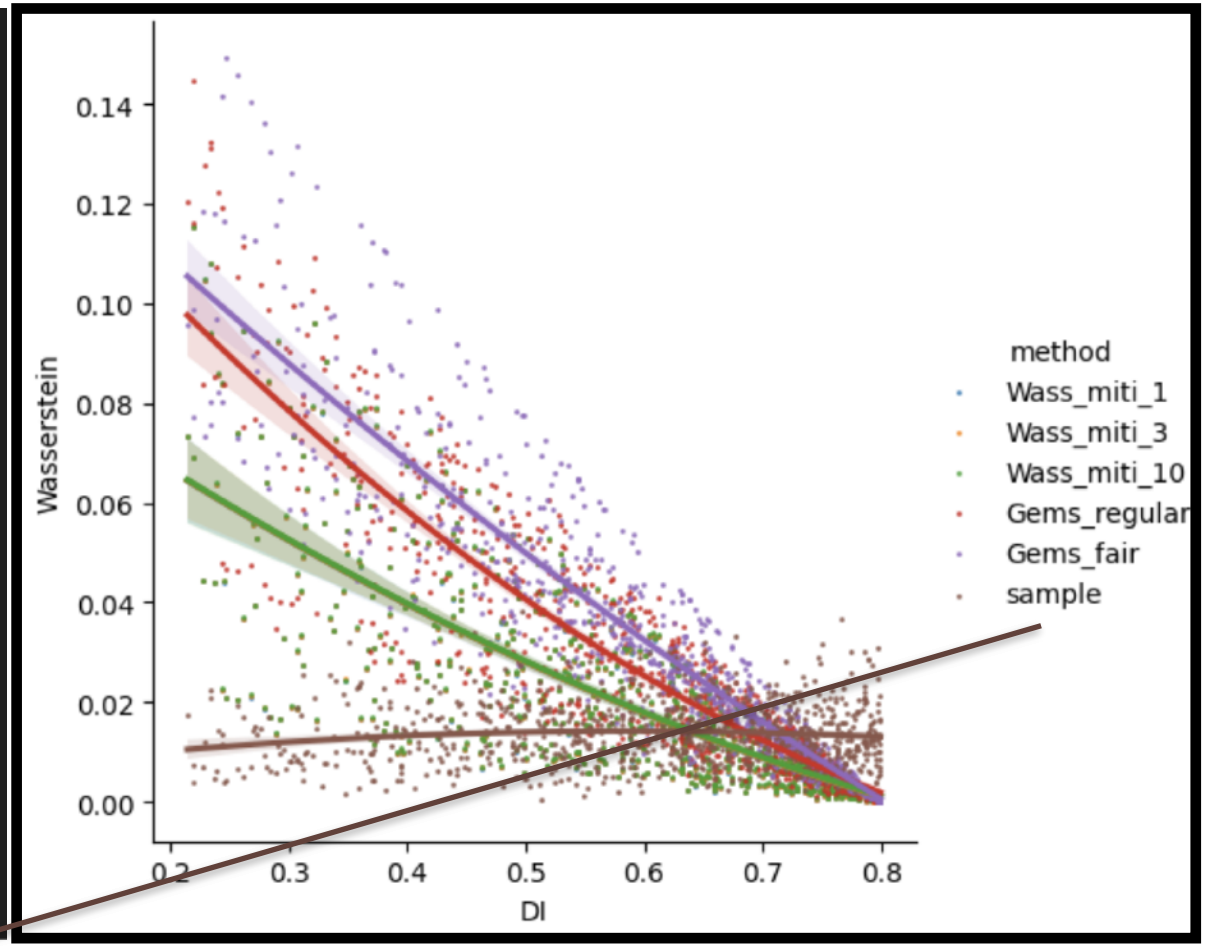
en contrôlant la distance minimale pour passer de la loi non conforme à loi conforme (i.e DI=0.8)

Inria

FAIR WASHING EN IMPOSANT DI =0.8

Globally Explaining models under Stress - GEMS-AI https://gems-ai.aniti.fr/



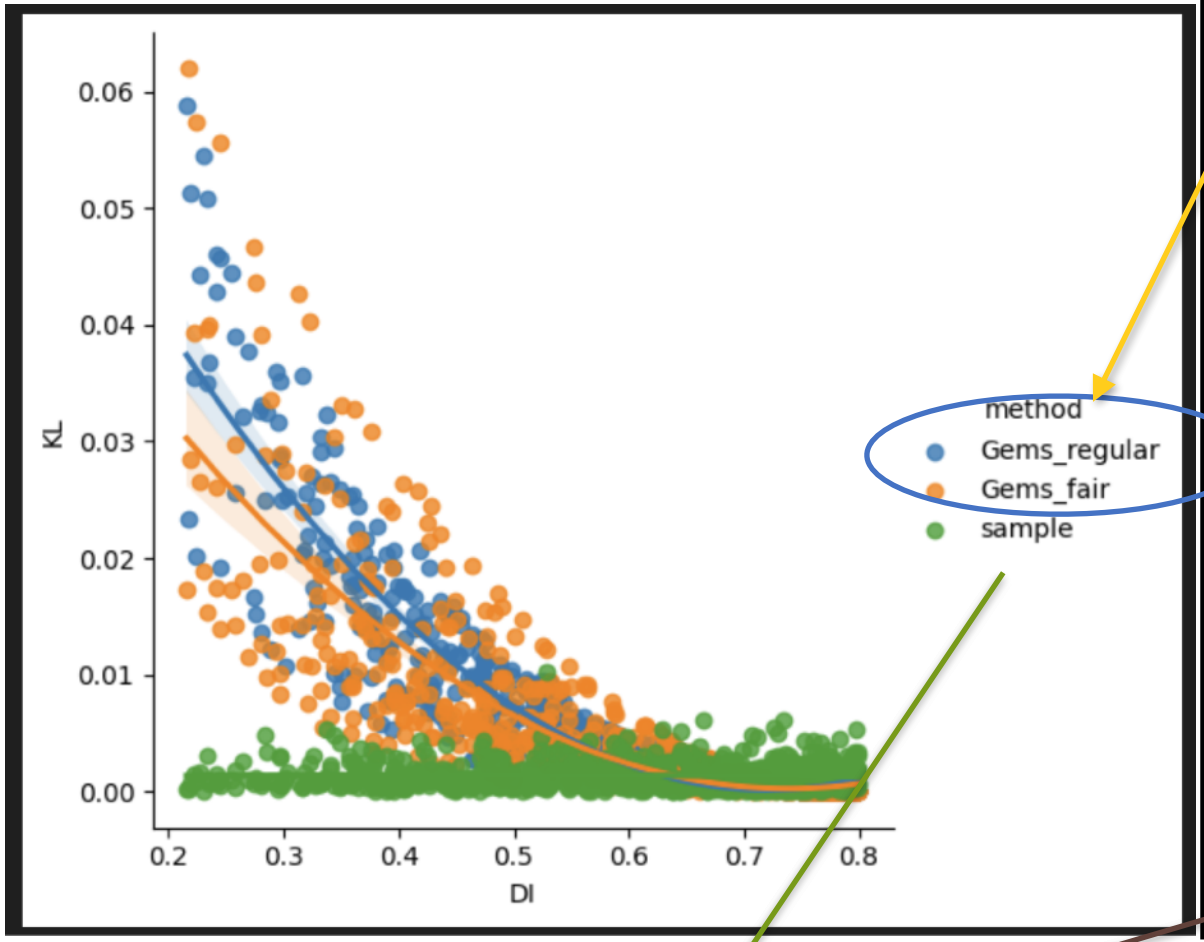


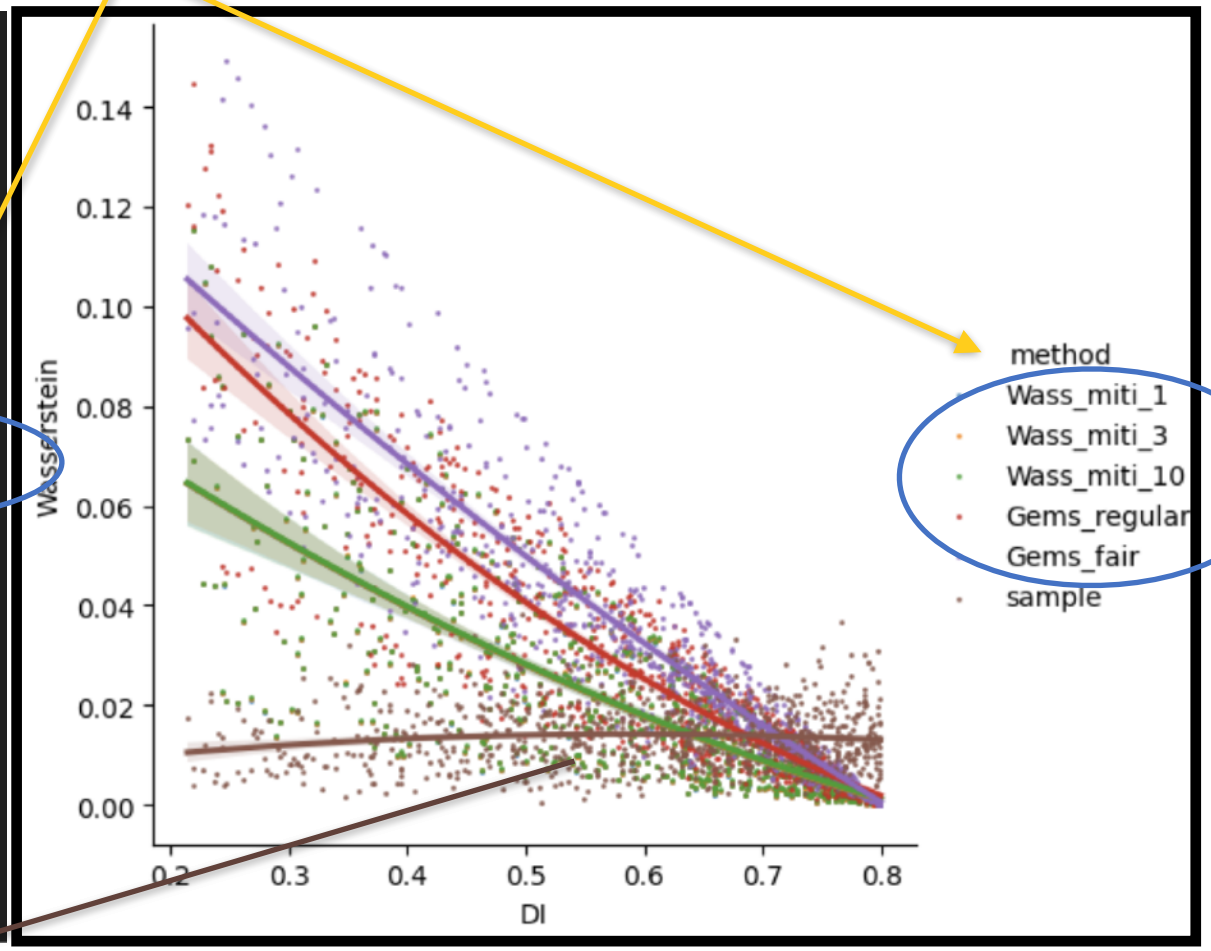
Sample dont on fait varier le DI

FAIR WASHING EN IMPOSANT DI =0.8

Globally Explaining models under Stress - GEMS-AI https://gems-ai.aniti.fr/

Samples « maquillés » pour vérifier DI = 0.8



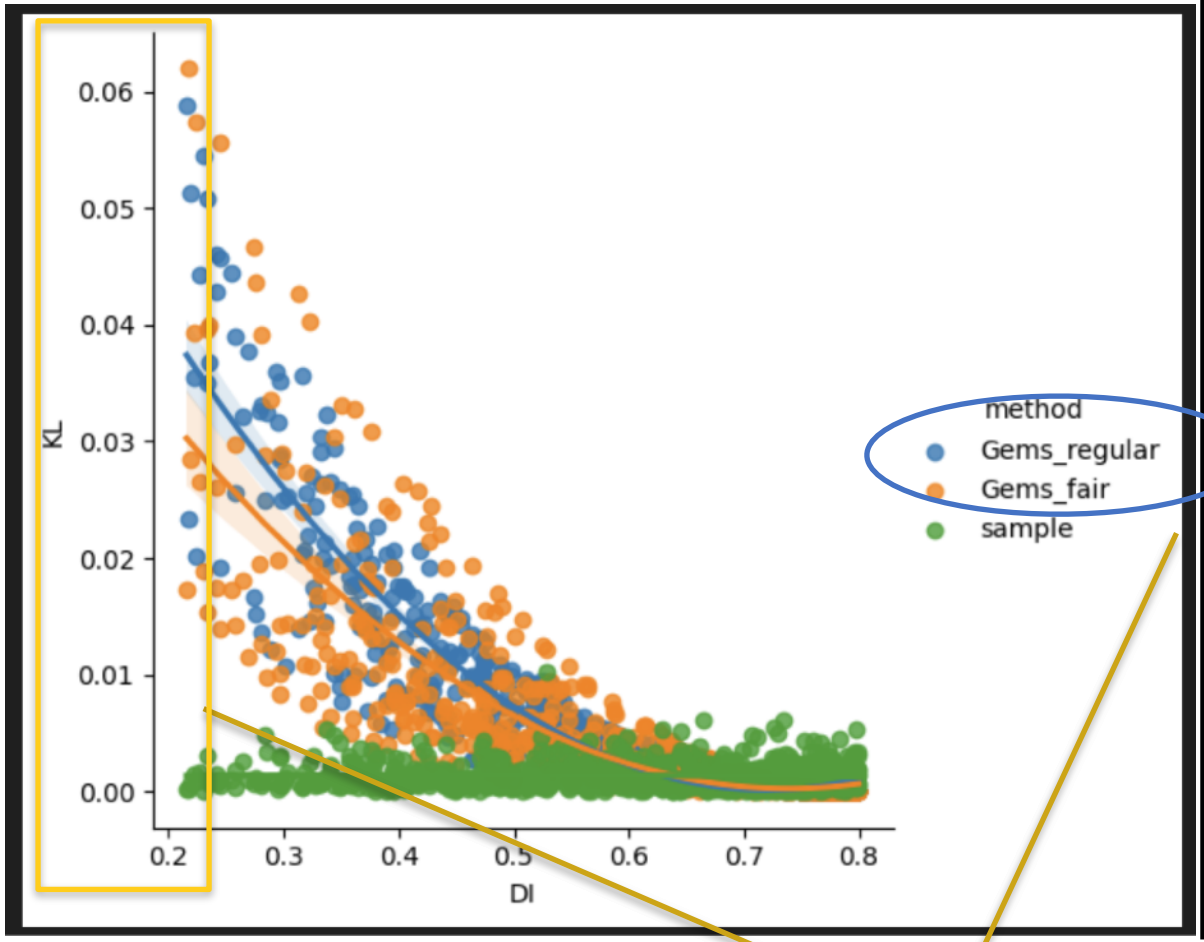


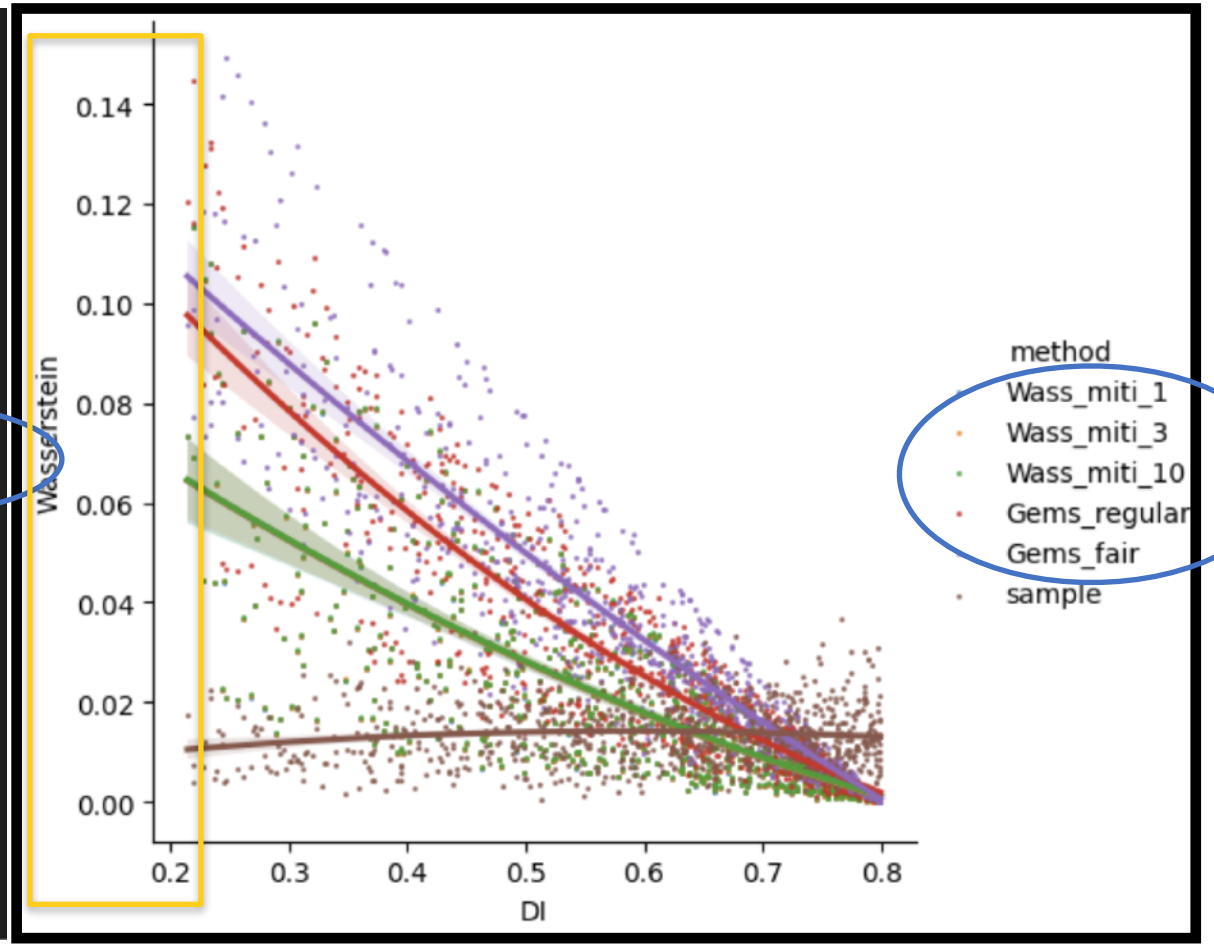
Sample dont on fait varier le DI

FAIR WASHING EN IMPOSANT DI =0.8

Globally Explaining models under Stress - GEMS-AI https://gems-ai.aniti.fr/

Samples « maquillés » pour vérifier DI = 0.8



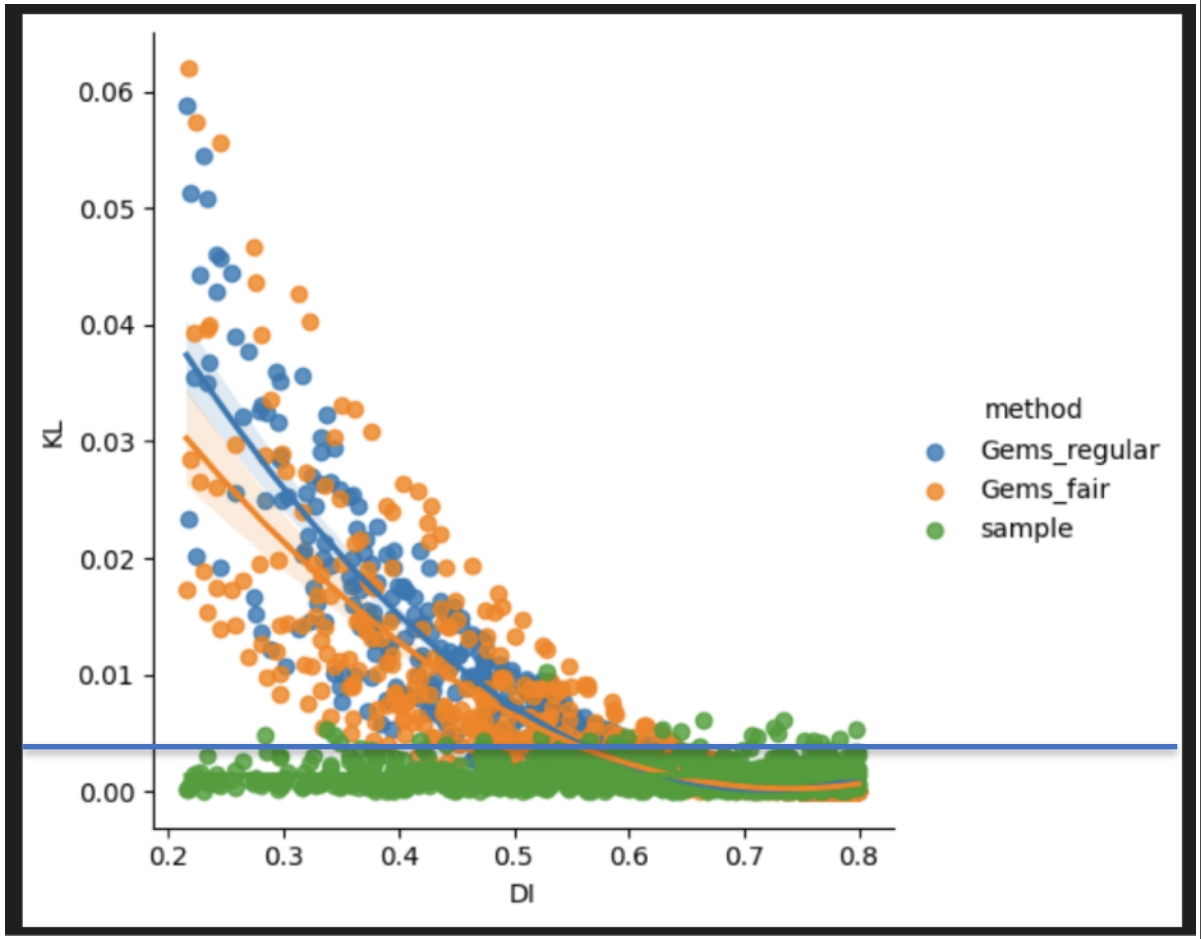


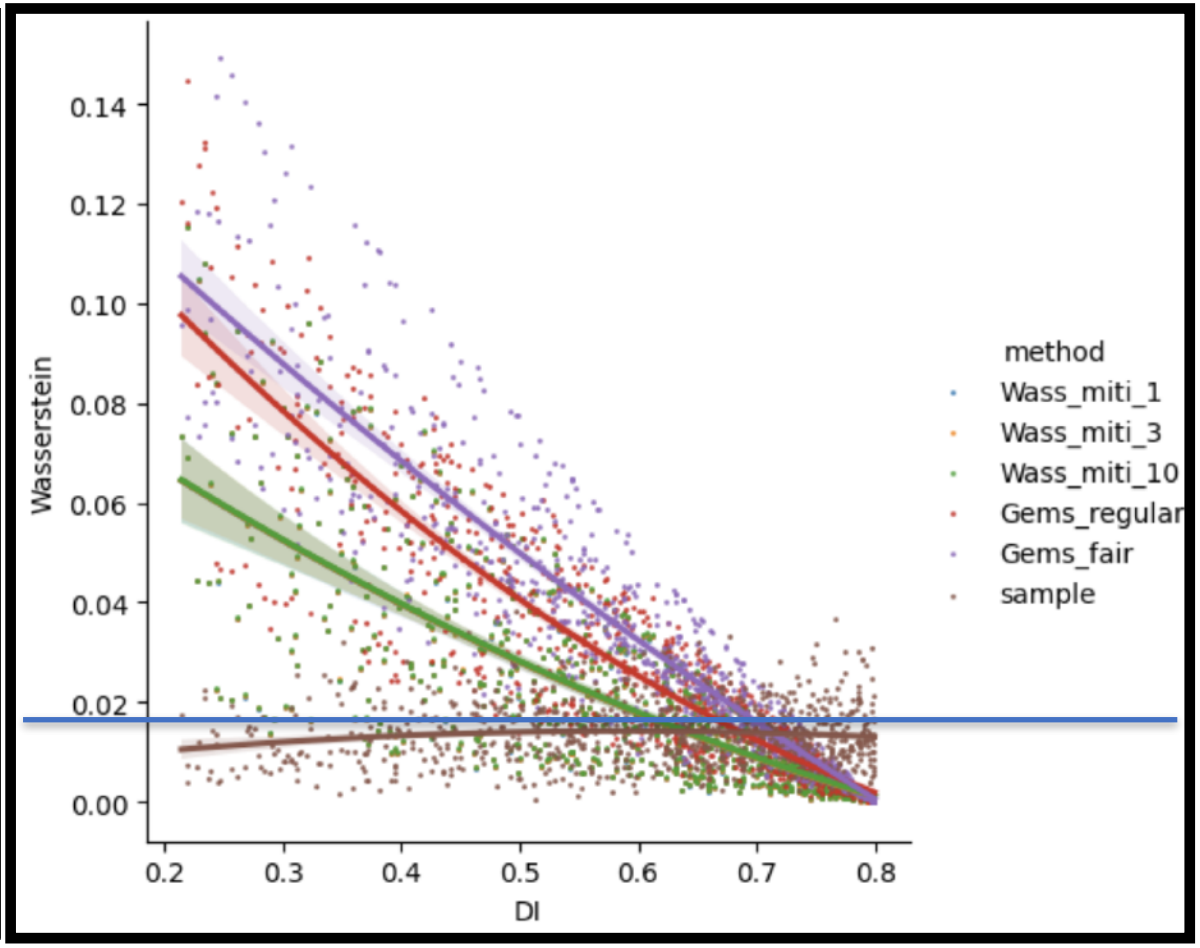
Distance entre sample modifié et sample originel

FAIR WASHING PAR PROJECTION ENTROPIQUE

Globally Explaining models under Stress - GEMS-AI https://gems-ai.aniti.fr/

Avantage distance « transport optimal »





Limite DI $\geq$ 0.52

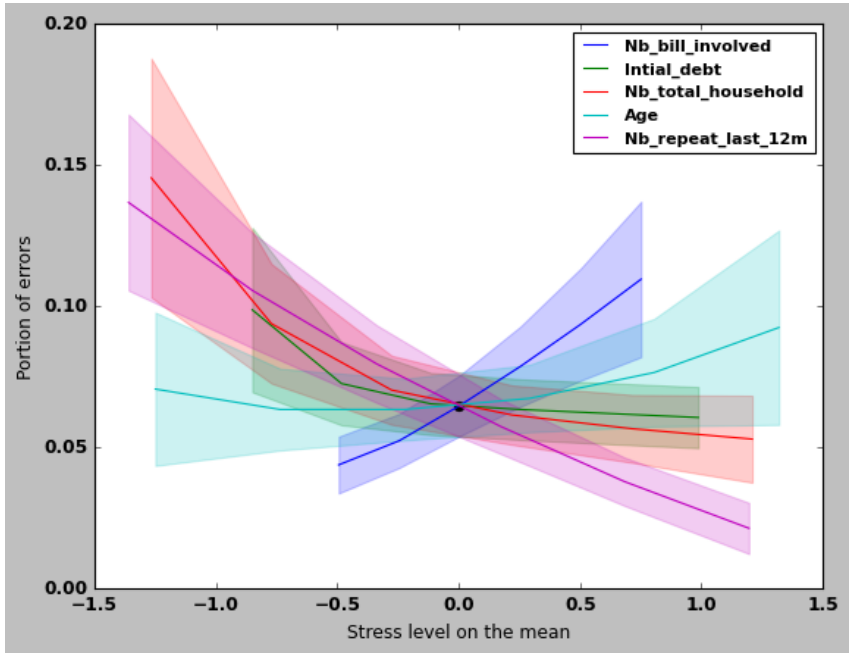
Limite DI $\geq$ 0.6

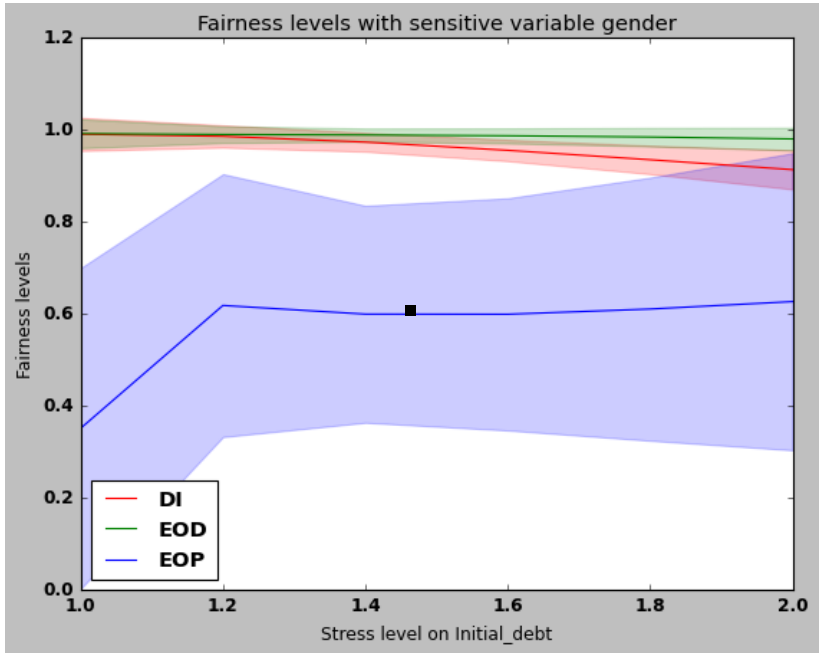
RECOMMENDATION 1/2 : SE PRÉPARER AUX PIRES CAS

Incertitude de la mesure des observations $Q \in \mathscr{V}_d(\varepsilon) = \{R \in \mathscr{P}(\mathbb{R}^p), d(R, P_n) \leq \varepsilon\}$

- **Des mesures robustes aux perturbations** (performance/biais/explicabilité) $\max_{Q \in \mathscr{V}(\epsilon)} DI_Q(f)$
- **Expliquer la variabilité des mesures de biais** sous **conditions de stress** sur les données

https://gems-amies-apps.apps.math.cnrs.fr/ Bachoc, Loubes, J. M., & Risser, L. (2023). **Explaining machine learning models using entropic variable projection,** Information & Inference.





RECOMMENDATION 2/2 : DES MESURES LOCALISÉES

Segmenter automatiquement les zones d'intérêt pour mieux détecter et mieux réparer

adapté à un critère de biais, de performance (ou d'explicabilité)

- Segmenter la distribution en **zones homogènes et à fort impact** (Rottembourg, Loubes et al. (2025))

$$Q = \sum_j \pi_j Q_j. \qquad \text{ou} \qquad Q = (1-\alpha)Q_\alpha + \alpha N$$

optimiser en fonction des distributions … notion de **Wasserstein gradient flow**

- Découper modèle en **concepts à apprendre** $f(\,.\,) = \sum_j p_j f_j(\,.\,)$ (Bolte, Loubes et al. 2025): nouvelles mesures intrinsèques du biais (cf « **SCOR Fondation workshop** » 15/05/2025)

Pour aller plus loin …

•M. de Vos, A. Dhasade, J. Garcia Bourrée, A-M. Kermarrec, E. Le Merrer, B. Rottembourg, G. Tredan : **Fairness Auditing with Multi-Agent Collaboration**, Volume 392: ECAI 2024

•J. Garcia Bourrée, H. Lautraite, S. Gambs, G. Tredan, E. Le Merrer, B. Rottembourg : **P2NIA: Privacy-Preserving Non-Iterative Auditing**, arXiv preprint arXiv:2504.00874

•J. Garcia Bourrée, E. Le Merrer, G. Tredan, B. Rottembourg : **On the relevance of APIs facing fairwashed audits**, arXiv preprint arXiv:2305.13883

•Besse, P., Castets-Renard, C., Garivier, A., & Loubes, J. M. (2018). **Can everyday AI be ethical**. *Machine Learning Algorithm Fairness (english version)*, *10*

•Gordaliza, P., Del Barrio, E., Fabrice, G., & Loubes, J. M. (2019, May). **Obtaining fairness using optimal transport theory**. In *International conference on machine learning* (pp. 2357-2365). PMLR.

•De Lara, L., González-Sanz, A., Asher, N., Risser, L., & Loubes, J. M. (2024). **Transport-based counterfactual models**. *Journal of Machine Learning Research*, *25*(136), 1-59

Bénesse, C., Gamboa, F., Loubes, J. M., & Boissin, T. (2024). **Fairness seen as global sensitivity analysis**. *Machine Learning*, *113*(5), 3205-3232.