Apparier la complexité ne suffit pas

Ce qu'une comparaison de représentations établit, et ce qu'elle ne peut pas encore établir

Résumé exécutif

Problème. Comparer deux représentations sans contrôler leur complexité confond structure et capacité : ce qu'on attribue au format n'est souvent que de la régularisation.

Méthode. Même famille d'estimation, même estimand marginal, complexité effective appariée par la trace de l'opérateur de lissage, processus générateurs semi-synthétiques calibrés sur des marges de covariables publiées, protocole préenregistré et horodaté avant exécution.

Résultat principal. L'appariement élimine une partie des différences entre représentations, mais pas toutes. À budget de complexité comparable, des transformations d'une même famille d'estimation conservent des performances d'estimation de l'estimand causal différentes malgré une complexité globale appariée ; l'analyse mécaniste suggère que l'expressivité de la classe fonctionnelle en explique une partie.

Généralisation. Sur une distribution de mondes explorée en post-enregistrement, l'avantage est probabiliste et non universel : P(axe continu meilleur que format large) = 0,73 ± 0,02. Dans plus d'un quart des mondes, une autre transformation fait mieux.

Résultat mécaniste. La concentration spectrale de la complexité ne fournit pas l'explication attendue. L'expressivité de la classe fonctionnelle est plus informative, sans être universellement prédictive : elle explique deux régimes sur trois et change de signe dans le troisième.

Borne d'identifiabilité. Aucune représentation ne compense une perte d'identifiabilité : sous confusion cachée non récupérable, toutes échouent.

Bornes. Une seule famille d'estimation. Des mécanismes dont le support structurel est défini par les auteurs. Aucune validation clinique : les données cliniques servent uniquement à calibrer des distributions marginales.

Conclusion. Lorsque l'objectif est d'attribuer une différence de performance à la représentation plutôt qu'à la capacité, l'appariement de complexité est une condition méthodologique nécessaire. Il ne suffit pas à expliquer les différences résiduelles.

Trois niveaux de revendication

Un article méthodologique qui superpose ses niveaux de preuve invite la réfutation triviale : il suffit d'attaquer le plus faible pour paraître avoir réfuté l'ensemble. Ce manuscrit les sépare avant d'exposer quoi que ce soit [25,27].

Niveau A : établi dans le cadre expérimental préspécifié. Des transformations représentationnelles distinctes produisent des biais différents malgré l'égalisation de leur complexité effective globale. Section 7.

Niveau B : soutenu, non isolé. Ces différences sont davantage compatibles avec une explication par l'expressivité de la classe fonctionnelle qu'avec une explication par la capacité. Le run est cohérent avec cette lecture ; il ne l'isole pas de toutes les alternatives. Sections 8 et 9.

Niveau C : testé, non retenu. Le mécanisme de cette adéquation serait l'alignement entre le signal et les directions auxquelles l'opérateur de lissage attribue sa complexité. Cette hypothèse a été testée et n'est pas soutenue sous la forme proposée. Section 9.

Le lecteur pressé peut s'arrêter à A. Le lecteur méthodologique lira la section 11, qui cartographie ce qui est contrôlé et ce qui ne l'est pas.

Résultat clé

> **df_eff(R_1) = df_eff(R_2)** n'implique ni **F_R_1 = F_R_2**, ni **Biais(R_1) = Biais(R_2)**.
>
> *Deux pipelines de même complexité effective globale peuvent opérer sur des classes fonctionnelles différentes, et conserver des performances d'estimation de l'estimand causal différentes.*

1. Question de recherche

La question naïve oppose le temps porté comme axe au temps éclaté en colonnes. Elle est mal posée. La version défendable est plus étroite :

> *« À partir des mêmes covariables sources, comment différentes transformations représentationnelles, associées à une même famille d'estimation, modifient-elles la classe fonctionnelle accessible à complexité effective comparable ? »*

C'est-à-dire que ranger les mêmes données différemment change ce qu'un modèle peut comprendre, même quand on lui donne exactement la même marge de manœuvre.

Le changement n'est pas terminologique. La formulation antérieure, « à information d'entrée identique », était **fausse au sens statistique**. Nos quatre transformations ne préservent pas la même information :

- La grille ordonnée réalise une compression plusieurs-vers-un,
- Le format large détruit une partie de la métrique continue,
- L'axe continu la conserve et ajoute une base quadratique,
- La forme hiérarchique injecte une structure absente des covariables sources.

Quatre opérations sont donc confondues sous le mot « disposition » : agencement, encodage, ingénierie de variables, restriction de classe fonctionnelle. **Nous ne les démêlons pas.** Ce que nous établissons n'est donc pas que « la représentation compte », mais que la construction de l'espace de caractéristiques reste déterminante après égalisation d'un budget global de complexité. C'est moins spectaculaire et beaucoup plus difficile à réfuter.

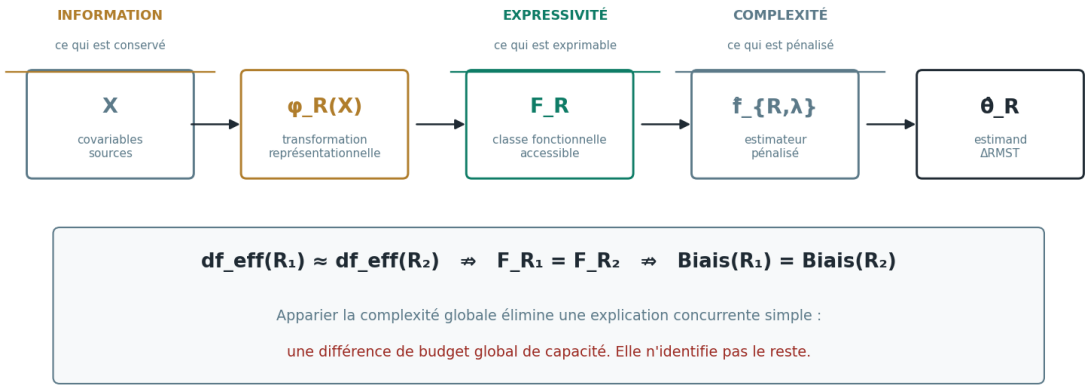

Figure 1. *La chaîne représentationnelle et ses trois niveaux. Égaliser les degrés de liberté effectifs n'égalise ni les classes fonctionnelles accessibles ni les biais.*

2. Trois niveaux qu'un seul mot recouvre

Le mot *symétrie* recouvre trois choses distinctes, et leur confusion est la principale faiblesse conceptuelle de la littérature sur le biais inductif.

- Le premier niveau est la **structure de l'espace d'entrée** (ordre, métrique, voisinage, hiérarchie) que la transformation fixe.
- Le deuxième est l'**hypothèse fonctionnelle** (lissage, localité, additivité) que le couple transformation-modèle rend peu coûteuse ou non.
- Le troisième seulement est la **symétrie au sens strict** : invariance sous un groupe explicitement défini, $f(gx) = f(x)$ pour $g \in G$.

En écrivant le modèle comme composition $f = h \circ \varphi$, l'invariance effective dépend conjointement des deux termes. La transformation, qui fixe φ, ne la détermine pas seule.

Nos expériences testent le deuxième niveau, pas le troisième. Nous parlerons donc d'*adéquation structurelle*, et réserverons *symétrie* au cadre théorique.

3. État de l'art

Six lignées se croisent ici, dont aucune, seule, ne nomme son objet et est donc auto-suffisante.

1. La première est celle du **biais inductif et de l'invariance**, dont la formulation la plus aboutie est géométrique : une architecture encode une hypothèse d'invariance sous un groupe, dont découle le partage de paramètres [1,2,3,4,5]. Ce vocabulaire a été forgé pour les images et les graphes.
2. La deuxième est celle des **modèles de survie profonds**, jusqu'au cadre exponentiel par morceaux qui fonde la famille employée ici [6,7,8,9,10]. Tous font varier l'architecture ; aucun ne fait varier la transformation à architecture fixée.
3. La troisième est celle de l'**apprentissage tabulaire**, qui demande comment encoder une variable, non quelle structure sa transformation déclare [11,12,13,14].
4. La quatrième est celle de la **comparaison indirecte ajustée sur la population**. La reconstruction de données individuelles depuis des courbes publiées [15] alimente la comparaison appariée [16,17,18] ; cette lignée consomme la donnée reconstruite, quand notre banc en est un producteur.
5. La cinquième est celle de l'**analyse fonctionnelle et de la régularisation temporelle** : expansions de base, lissages par produit tensoriel [19], modèles à coefficients variables [20], régularisations imposant la parcimonie des différences [21,22]. Le compromis que notre banc met en évidence — paramètres séparés par intervalle contre fonction lisse partagée — y est un problème ancien et bien caractérisé. Notre apport n'est pas de le découvrir mais de le relire en termes de transformation.
6. La sixième est celle des **dimensions effectives et de l'alignement spectral** : méthodes à noyaux, dimensions effectives, a priori bayésiens, où la question n'est pas combien de degrés de liberté un estimateur consomme mais dans quelles directions il les dépense [29,30,31,32,33].

4. Données

4.1 Ce que la reconstruction produit

L'inversion des équations de Kaplan-Meier, selon la méthode de Guyot et collègues [15] admise par le NICE et la HAS, produit des jeux individuels temps-événement-censure validés contre les rapports de risque publiés.

Nous avons utilisé cette méthode pour recréer des « patients sources » à partir des courbes de survie présentées dans les articles sur les essais cliniques ZUMA-7 et TRANSFORM pour obtenir de la matière première pour notre banc d'essai.

Source	Critère	HR publié	HR reconstruit	Écart
TRANSFORM	survie sans événement (EFS)	0,375	0,359	−0,016
TRANSFORM	survie globale (OS)	0,757	0,751	−0,006
ZUMA-7	survie sans événement (EFS)	0,398	0,398	0,000
ZUMA-7	survie globale (OS)	0,730	0,729	−0,001

Elle ne restitue pas les covariables : l'inversion porte sur une courbe agrégée.

4.2 Aucune conclusion clinique ne suit de cette étude

Nous employons une **simulation semi-synthétique calibrée sur des marges de covariables publiées** [28]. La présence de noms de cohortes cliniques dans un article méthodologique produit spontanément une impression de validation empirique. Il faut donc être explicite.

> *Les données cliniques servent uniquement à calibrer certaines distributions marginales de covariables.* ***Aucune conclusion sur le lymphome, sur l'efficacité des CAR-T, sur l'hétérogénéité des traitements ou sur la décision clinique ne suit de ce travail****.*

Ce qui vient des cohortes TRANSFORM et ALYCANTE : les distributions marginales, rien d'autre.

Ce qui n'en vient pas : les corrélations, le mécanisme causal, la forme de l'hétérogénéité temporelle, les interactions, le régime de censure réel.

Les covariables binaires sont tirées par seuillage d'une copule gaussienne aux marges publiées, les continues par transformée inverse d'une log-normale.

C'est-à-dire qu'on fabrique des patients fictifs dont les caractéristiques ont les mêmes proportions que celles publiées, et restent liées entre elles de façon plausible : les oui/non par un seuil calé sur la proportion connue, les valeurs chiffrées par une distribution qui reproduit la forme habituelle des grandeurs biologiques.

5. Méthodes

5.1 Les quatre transformations

- Le **format large** discrétise chaque covariable continue en tranches quantiles encodées en indicatrices.
- La **grille ordonnée** agrège les covariables en un score unique, discrétisé en niveaux.
- L'**axe continu** conserve les covariables comme dimensions continues, augmentées de leurs termes quadratiques.
- La **forme hiérarchique** croise les covariables avec une strate de sévérité binaire dont l'encodage est déclaré et fixe.

5.2 Une seule famille d'estimation, et ce que cela implique

Modèle de survie à temps discret ajusté par régression logistique pénalisée sur format personne-période. Ce choix réduit les biais inductifs concurrents de celui qu'on étudie, mais cette famille n'est pas sans opinion : elle impose une forme additive, une fonction de lien, une pénalisation quadratique, une approximation par intervalles.

Il s'ensuit que l'étude porte sur une tranche de l'espace des questions :

$$\mathbf{R \times DGP \mid M = M_0}$$

où

- R désigne la représentation,
- DGP le mécanisme,
- M la famille d'estimation.

Nous faisons varier les deux premiers termes et fixons le troisième. Savoir si les différences observées sont intrinsèques aux espaces fonctionnels ou tiennent à leur interaction avec cette pénalisation particulière demeure ouvert, et constitue le prolongement naturel de ce travail (section 12).

5.3 Estimand

Différence marginale de survie moyenne restreinte à horizon fixe, par g-formule :

$$\Delta\mathrm{RMST}(\tau) = \int_0^\tau S^1(t)\,dt - \int_0^\tau S^0(t)\,dt$$

Un coefficient conditionnel dépend des covariables incluses, si bien que deux transformations comparées sur leur propre coefficient ne visent pas la même cible. La RMST marginale est invariante à la spécification.

On mesure le temps de vie moyen gagné sur une fenêtre fixe, une quantité qui reste la même quelle que soit la façon de construire le modèle, plutôt qu'un « effet toutes choses égales par ailleurs », qui changerait de sens à chaque mise en forme et rendrait la comparaison impossible.

5.4 Appariement de complexité

La trace de l'opérateur de lissage est la mesure standard des degrés de liberté effectifs d'un estimateur linéaire pénalisé [29,30] :

$$\mathrm{df}_{\mathrm{eff}} = \mathrm{tr}[(X^{\top}WX + \lambda P)^{-1}X^{\top}WX]$$

La cible est celle atteinte par l'**oracle**, ajustement sur le vrai confondant latent ; le paramètre de pénalité de chaque transformation est résolu par recherche dichotomique pour l'atteindre avant toute lecture du biais.

Deux réserves :

1. L'appariement égalise une **quantité**, non une géométrie : deux opérateurs de même trace peuvent avoir des spectres très différents (section 9).
2. Et l'oracle est une **information privilégiée** ; cette réserve est levée en section 8.5, où une cible sélectionnée hors échantillon reproduit le classement.

5.5 Récupérabilité, plan et préinscription

La récupérabilité est jugée au niveau de la condition, sur le biais oracle moyen, non réplication par réplication, sans quoi le bruit d'échantillonnage de l'oracle en effectif fini serait confondu avec une non-identifiabilité structurelle.

On vérifie qu'un monde fabriqué (synthétique) est sain en regardant si le modèle témoin retrouve la bonne réponse en moyenne, pas à chaque tirage : sinon on éliminerait des mondes valides pour un simple coup de malchance, en croyant y voir un défaut de fond.

Plan partiellement factoriel : cœur croisant deux jeux de marges, trois forces de confusion, trois régimes et trois niveaux de régularité spectrale (54 conditions) ; couronne de robustesse en variation d'un facteur à la fois (6 conditions). Deux mille réplications par condition, nombre dérivé d'une contrainte d'incertitude déclarée [26].

Le code, le protocole et le plan d'analyse ont été déposés sur OSF et horodatés avant exécution ; le run vérifie au lancement l'empreinte du code et refuse de s'exécuter si elle diffère.

6. Campagne exploratoire préalable (non confirmatoire)

Aucun résultat de cette section n'est retenu comme probant. Elle explique pourquoi le plan confirmatoire a la forme qu'il a.

Une première expérience suggérait un avantage de la base partagée ; un plan à degrés de liberté appariés a montré que cet avantage n'était pas attribuable à la continuité. Une expérience annexe donnait l'axe continu gagnant en confusion multiple et temporelle, le format large en régime non linéaire, jusqu'à ce qu'un test d'objection renverse la lecture : un axe continu enrichi d'un terme quadratique repassait devant partout. Le format large ne l'emportait que parce qu'on le comparait à un continu mal spécifié.

C'est ce test qui a produit l'hypothèse préinscrite : si l'avantage tient à l'adéquation entre forme exprimable et forme réelle, une transformation doit gagner exactement dans les régimes dont sa classe fonctionnelle épouse la structure.

7. Résultats confirmatoires

Soixante conditions, deux mille réplications, deux jeux de marges. Quatre conditions écartées de l'analyse primaire, toutes à confusion cachée. Aucune condition du cœur écartée.

7.1 Résultat primaire

Régime de confusion	n	Format large	Grille	Axe continu	Hiérarch.
Multiple	12	0,2318	0,2313	**0,2147**	0,2195
Temporelle	12	1,0171	1,1129	**0,9265**	0,9720
Non linéaire	12	0,0241	0,0339	**0,0223**	0,0850

Valeur absolue du biais sur ΔRMST, à complexité effective appariée. Erreur type de Monte-Carlo ≈ 0,004.

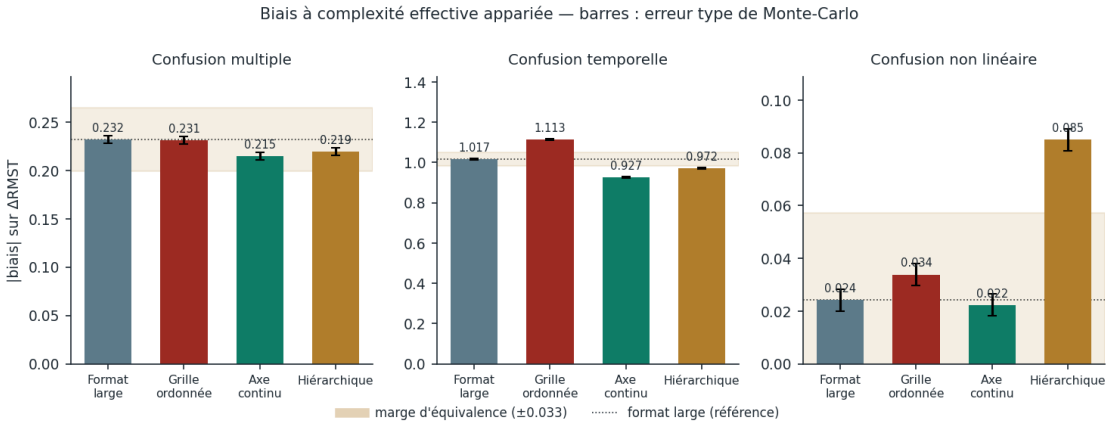

Figure 2. *Biais par régime et transformation. Barres : erreur type de Monte-Carlo. Bande dorée : marge d'équivalence autour du format large.*

7.2 Décision par marge d'équivalence

Régime	**Comparaison**	**Écart**	**> marge (0,033) ?**
multiple	continu vs large	0,017	non
multiple	hiérarchique vs large	0,012	non
temporelle	continu vs large	0,091	**oui**
temporelle	hiérarchique vs large	0,045	**oui**
temporelle	grille, *pire* que large	0,096	**oui**
non linéaire	continu vs large	0,002	non
non linéaire	hiérarchique, *pire* que large	0,061	**oui**

À deux mille réplications, la correction de multiplicité rejette l'égalité dans douze conditions sur douze pour l'axe continu en régime multiple. L'écart correspondant vaut la moitié de la marge. Nous concluons sur la marge, non sur le rejet.

L'origine de la marge doit être dite sans embellissement : elle provient de la différence minimale détectable observée en campagne exploratoire. Ni grandeur cliniquement fondée, ni fraction déclarée de la RMST. Sa robustesse est examinée en section 8.1.

7.3 Stabilité de Monte-Carlo

Une exécution à cent vingt réplications sur machine distincte donne le même ordre dans les trois régimes, à moins de trois millièmes près. L'erreur type passe de 0,0160 à 0,0040, facteur 4,04 pour une racine théorique de 4,08.

Cette précision porte sur l'erreur conditionnelle à chaque monde. Elle ne dit rien de la performance moyenne sur une distribution de mondes (voir section 8.3).

7.4 Échec structurel de la grille ordonnée

Le format large, l'axe continu et la forme hiérarchique atteignent la trace cible dans la totalité des conditions. **La grille ne l'atteint dans aucune.** Ayant réduit les covariables à un score unique, sa classe fonctionnelle plafonne à une complexité inférieure à celle de l'oracle, quelle que soit la pénalité.

La comparaison n'est donc pas « quatre transformations à complexité appariée » : c'est trois transformations appariées, et une quatrième structurellement sous-paramétrée dont l'incapacité est elle-même un résultat pré-spécifié.

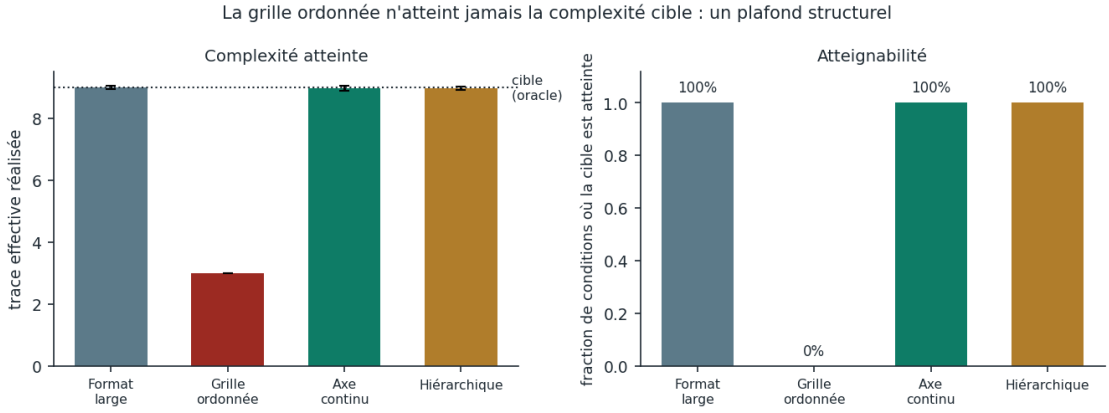

Figure 3. *Complexité réalisée et fraction de conditions où la cible est atteinte.*

8. Robustesse post-enregistrement et tests de résistance

Les analyses de cette section sont **post-hoc et hors préinscription**. Elles testent la robustesse du résultat de la section 7 ; aucune n'y ajoute de revendication confirmatoire.

8.1 Sensibilité à la marge

Sur $\delta \in [0{,}020 ; 0{,}050]$, le verdict est réfuté en tout point, sur les deux jeux de marges. La frontière de décision, par dichotomie, se situe à $\delta \approx 0{,}140$: il faudrait une marge **4,2 fois supérieure** à celle retenue pour conserver H1, soit 26,8 erreurs types de Monte-Carlo.

δ	multiple : mieux / pire	temporel : mieux / pire	non linéaire : mieux / pire
0,020	8 / 0	22 / 12	0 / 15
0,033	**0 / 0**	18 / 12	0 / 14
0,050	0 / 0	12 / 12	0 / 7

Le régime multiple perd la totalité de ses dépassements entre 0,020 et 0,033 ; le temporel en conserve dix-huit. La marge se situe juste au-dessus du bruit du premier régime et très en dessous du signal du second (adéquation constatée, non construite).

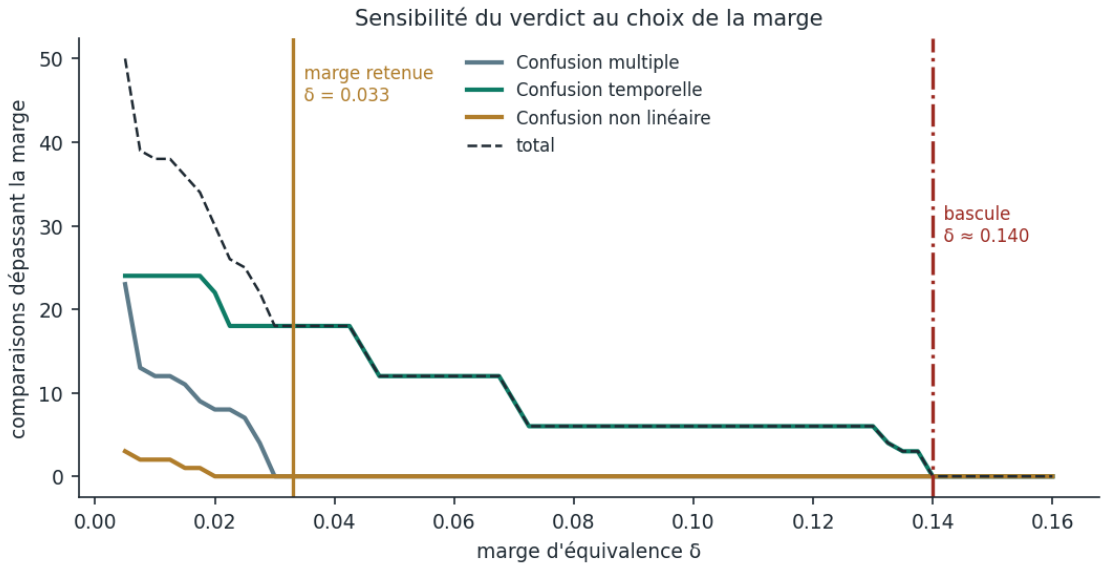

Figure 4. *Comparaisons dépassant la marge selon δ. Ligne dorée : marge retenue. Ligne rouge : frontière de bascule.*

8.2 Décomposition biais-variance

L'appariement égalise aussi la variance : étendue relative des écarts types entre transformations de 1,6 % à 3,6 %. Rien ne le garantissait.

Mais le classement bascule en régime non linéaire : par biais, l'axe continu devance le format large ; par RMSE, l'ordre s'inverse. **Toute affirmation de dominance doit nommer son critère.**

Régime	**Part de variance dans le MSE**	**Dominé par**
Confusion multiple	52,8 %	variance
Confusion temporelle	**5,6 %**	**biais**
Confusion non linéaire	93,0 %	variance

Nos conclusions les plus fortes portent sur le régime où le choix d'estimand est le mieux justifié ; les plus faibles sur ceux où il l'est le moins.

8.3 Variabilité entre mondes

Quatre cents mondes tirés d'une distribution déclarée, vingt-cinq réplications chacun ; sept quantités que le plan traitait comme connues y deviennent aléatoires.

Transformation	**Var. entre mondes**	**Var. intra-monde**	**Part entre**
Format large	0,2935	0,0363	89,0 %
Grille ordonnée	0,3362	0,0370	90,1 %
Axe continu	0,2441	0,0365	87,0 %
Hiérarchique	0,2836	0,0368	88,5 %

88,6 % en moyenne. L'écart type de la performance entre mondes vaut environ 0,54, cent trente fois l'erreur type de Monte-Carlo du run confirmatoire.

> *Deux mille réplications ne corrigent pas soixante mondes.*

L'estimand adapté est une probabilité de supériorité, $\Pi(R_1, R_2) = P\_DGP(|biais(R_1)| < |biais(R_2)|)$:

Comparaison	Π	Écart type
Axe continu bat format large	**0,730**	0,022
Axe continu bat hiérarchique	**0,800**	0,020
Hiérarchique bat format large	0,657	0,024

Le classement moyen est identique à celui du run confirmatoire. Mais l'axe continu n'est le meilleur que dans **63,0 %** des mondes : la dominance est probabiliste, non universelle.

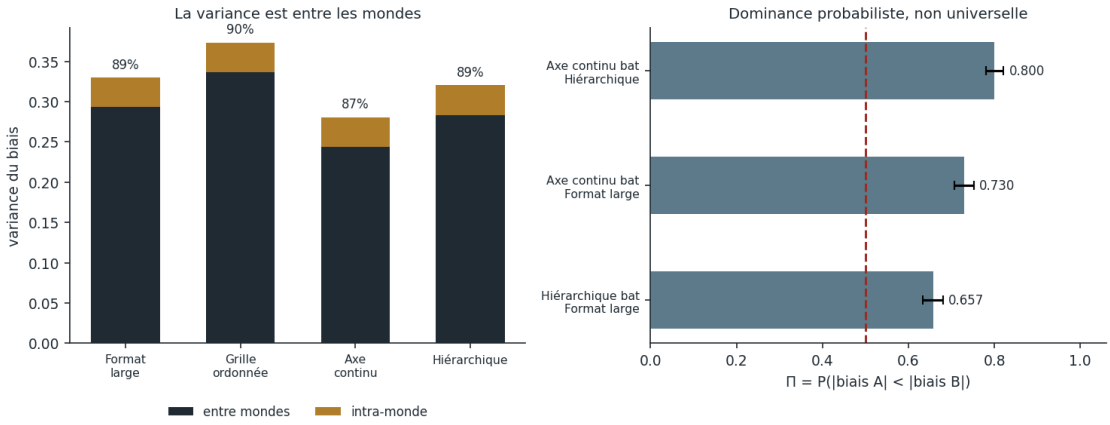

Figure 5. *À gauche : décomposition de la variance. À droite : probabilités de supériorité sur la distribution de mondes.*

8.4 Sensibilité à l'effectif

Six effectifs de 200 à 5 000, cent vingt mondes, comparaison appariée. Π reste plat : 0,783 à 200, 0,800 à 5 000 ; pente contre log n de +0,002. Les deux biais absolus décroissent bien avec l'effectif (les estimateurs s'améliorent) mais la probabilité que l'un batte l'autre ne bouge pas.

L'agrégat masque trois directions opposées : l'avantage croît en confusion multiple (0,700 à 0,933), plafonne en temporelle, et décroît vers le hasard en non linéaire (0,643 à 0,500).

8.5 Sensibilité à la cible de complexité

Sur des multiples de 0,5 à 1,5 fois la trace oracle, Π varie de 0,770 à 0,790, pour une erreur type de 0,024 : **le classement ne dépend pas du niveau de complexité**, sur un facteur trois.

Une cible sélectionnée par validation croisée sur la déviance hors échantillon, sans information privilégiée, donne Π = 0,760 ± 0,025 contre 0,770 ± 0,024 pour l'oracle, soit un écart de −0,010 ± 0,035.

> *L'appariement ne nécessite donc pas nécessairement une cible oracle : dans ce banc, une cible de complexité sélectionnée hors échantillon reproduit le classement obtenu avec la cible oracle.*

Cette cible croît avec l'effectif (2,66, puis 4,95, puis 7,98) là où la cible oracle saturait (8,74 à 8,98). Le canal que la théorie du biais inductif invoque est donc ouvert, et Π ne décroît pas pour autant : 0,800, 0,720, 0,760. La persistance de l'avantage observée en 8.4 n'est pas un artefact du bridage de la complexité.

8.6 Couverture des intervalles

La base gelée ne produit pas d'intervalles ; nous les construisons par méthode delta sur la g-formule, variance bayésienne de Wood pour l'estimation pénalisée.

La variance est correctement estimée : rapport se/sd de 1,01 à 1,03. Mais la couverture suit le rapport biais sur dispersion.

Régime	**biais / sd**	**Couverture**
Non linéaire	0,84	**94,8 %**
Multiple	1,47	70,8 %
Temporelle	5,72	**7,6 %**

La sous-couverture ne provient pas d'une mauvaise estimation de la variance ; elle résulte du biais de l'estimateur. Elle constitue néanmoins une **défaillance inférentielle** si ces intervalles sont utilisés pour couvrir l'estimand causal.

Une tension doit être déclarée : le régime temporel porte la conclusion de la section 7, et c'est celui où la couverture est la plus dégradée. Le biais étant *différentiel* entre transformations, la comparaison des biais entre transformations reste interprétable ; mais dans ce régime, aucune des quatre ne fournirait un intervalle utilisable en pratique.

8.7 Dépendance non gaussienne

Six familles de copules déclarées avant exécution (gaussienne en témoin, Clayton, Gumbel, Student, mélange de sous-populations, dépendance au bras) sur six cents mondes. Ces familles ont été définies indépendamment de toute question de représentation.

Famille	**Π(continu bat large)**	**±**	**mondes**
Gaussienne (témoin)	0,696	0,046	102
Clayton (queue inférieure)	**0,589**	0,044	124
Gumbel (queue supérieure)	0,760	0,043	100
Student (queues lourdes)	0,771	0,043	96
Mélange	0,802	0,042	91
Dépendance au bras	0,701	0,049	87

Écart entre non gaussiennes et témoin : **+0,021 ± 0,050**, nul dans le bruit. L'axe continu reste le meilleur dans les six familles.

Deux observations. Le témoin gaussien se situe **en dessous** de trois des cinq familles irrégulières, ce qui rend moins plausible l'hypothèse simple selon laquelle la régularité gaussienne du banc aurait été choisie de manière à avantager l'axe continu. Le support des mécanismes reste néanmoins défini par les auteurs (section 10.3). Et Clayton est la seule famille sous 0,60 : une dépendance de queue inférieure est la structure que la classe fonctionnelle de l'axe continu exprime le moins bien, ce qui indique où chercher un contre-exemple.

8.8 Contrôle négatif : la confusion cachée

Les quatre conditions à confusion cachée sont écartées de l'analyse primaire, l'oracle n'y récupérant pas la vérité (biais moyens de −0,72 à −2,14). Dans ces conditions, **aucune transformation ne fait mieux qu'une autre de manière utile** : toutes échouent.

> *Une transformation peut améliorer l'approximation fonctionnelle. Elle ne fabrique pas l'identifiabilité causale.*

Ce contrôle est essentiel à la lecture de tout l'article. Il empêche une extrapolation abusive du résultat vers les représentations, les données synthétiques ou les jumeaux numériques, où l'on serait tenté de croire qu'une bonne représentation récupère le bon mécanisme. Réserve opérationnelle : dans le monde réel, l'analyste ignore de quel côté de cette frontière il se trouve.

9. Investigation mécaniste

9.1 La prémisse spectrale est établie

Parmi les trois transformations réellement appariées :

Transformation	**Trace**	**Directions effectives**	**Gini**	**|biais|**
Format large	8,989	15,5	0,181	0,445
Axe continu	8,972	10,9	**0,457**	**0,408**
Hiérarchique	8,941	15,8	0,371	0,448

Traces appariées à 0,53 % près, indice de concentration variant de 82 %. Deux pipelines de complexité identique répartissent cette complexité très différemment. Le degré de liberté effectif ne suffit donc pas à caractériser un estimateur.

Trois façons de ranger les données reçoivent exactement le même budget de souplesse, mais le dépensent très différemment, la mesure standard de complexité dit combien un modèle peut s'ajuster, jamais dans quelle direction.

9.2 Le mécanisme proposé n'est pas soutenu

La corrélation entre concentration spectrale et biais est nulle : −0,07, −0,06, +0,05 selon le régime. Une statistique scalaire de concentration ne remplace pas la trace.

Un rattrapage était possible : ce n'est peut-être pas le *nombre* de directions qui compte mais leur *alignement* avec le signal. Nous avons donc projeté le vrai effet confondant sur la base propre de l'opérateur, en distinguant l'**expressivité** (part du vrai effet tombant dans l'espace du design) et la **rétention** (part de l'exprimable survivant au rétrécissement).

On a vérifié si ce qui compte est d'éclairer au bon endroit plutôt que largement en séparant deux questions : le modèle peut-il seulement représenter la forme recherchée, et la conserve-t-il une fois qu'il l'a trouvée ? C'est la première qui décide.

Transformation	**Expressivité**	**Rétention**	**Alignement**	**|biais|**
Format large	0,273	0,717	0,190	0,503
Axe continu	0,270	**0,865**	0,230	**0,467**
Hiérarchique	**0,363**	0,603	0,250	0,501

Corrélations intra-régime avec le biais :

Régime	**Expressivité**	**Rétention**
Confusion multiple	**−0,670**	+0,171
Confusion temporelle	**−0,556**	+0,154
Confusion non linéaire	+0,515	**−0,620**

Les résultats sont **davantage compatibles avec une explication par l'expressivité de la classe fonctionnelle qu'avec une explication par la concentration spectrale, sans qu'un descripteur unique explique les trois régimes.** L'expressivité prédit deux régimes sur trois et change de signe dans le troisième (celui que la section 8.2 montre dominé à 93 % par la variance).

Cette décomposition rejoint une littérature établie : l'erreur de généralisation en régression à noyau dépend conjointement du spectre et de l'alignement de la cible avec les fonctions propres [31,32]. Un descripteur mieux fondé existe et n'a pas été employé, la distribution cumulée de puissance de Canatar et collègues [32]. Un test authentique de la conjecture devrait passer par elle, et la préinscrire.

9.3 Un résultat transversal : l'agrégation masque le mécanisme

Deux fois dans ce travail, une moyenne sur des familles hétérogènes a produit une conclusion que le détail contredit. En section 9.2, la corrélation globale entre alignement et biais est positive alors que deux régimes sur trois donnent une corrélation négative. En section 8.4, Π est plat sur l'effectif alors que les trois régimes évoluent dans des directions opposées.

Ce n'est pas une anecdote méthodologique mais un résultat transversal :

> *Les performances agrégées sur des familles hétérogènes de mécanismes peuvent masquer précisément l'interaction que l'étude cherche à caractériser.*

Formellement, E_DGP[L(R, DGP)] ne suffit pas à caractériser L(R, DGP) lorsque l'interaction R × DGP est précisément l'objet d'intérêt. Un benchmark global peut alors classer correctement les méthodes tout en expliquant incorrectement pourquoi elles l'emportent.

> *Une moyenne de performance sur un ensemble de situations dit quelle méthode l'emporte le plus souvent. Elle ne dit pas dans quelles situations, ni pourquoi. Quand c'est justement l'interaction entre la méthode et la situation qu'on étudie, un classement global peut être exact et son explication fausse.*

Cette observation dépasse notre banc et concerne toute pratique de comparaison sur des collections hétérogènes de jeux de données.

10. Discussion

10.1 Ce que l'appariement fait, et ne fait pas

L'apport méthodologique tient en un geste, peu coûteux et qui change les conclusions. Sa portée doit être énoncée exactement.

> *Apparier la complexité globale élimine une explication concurrente simple : une différence de budget global de capacité. Cela n'identifie pas l'explication restante.*

C'est ce que montre l'ensemble du travail : les différences subsistent après appariement, et notre investigation mécaniste (section 9) échoue à les réduire à un descripteur unique.

Deux questions distinctes doivent être séparées ici, et leur confusion explique une partie des désaccords sur les comparaisons de représentations. *Quel pipeline fonctionne le mieux ?* appelle une comparaison à configuration optimale, où imposer une complexité égale serait artificiel. *Quelle part de la différence vient de la représentation ?* exige au contraire l'appariement. Ce travail traite la seconde, et ses conclusions ne se transposent pas à la première.

10.2 Face aux méthodes statistiques classiques

Le compromis mis en évidence (paramètres séparés par intervalle contre fonction lisse partagée) est un problème ancien de la régularisation fonctionnelle. Nous l'accordons sans réserve. Notre apport est de montrer qu'il gouverne des écarts que la littérature attribue couramment à la « représentation » ou au « format », et de fournir un protocole qui sépare les deux.

10.3 Biais de simulation : incertitude paramétrique et incertitude structurelle

L'objection la plus sérieuse est que les auteurs écrivent à la fois le mécanisme et les représentations destinées à y réussir. Quatre dispositifs y répondent partiellement :

1. Les marges des covariables proviennent de publications tierces,
2. La forme du confondant est échantillonnée sur un axe de régularité incluant des formes qu'aucune transformation ne peut exprimer,
3. La conception est horodatée avant exécution,

4. La structure de dépendance est échantillonnée dans six familles définies indépendamment de toute question de représentation (8.7).

Ce qui demeure doit être formulé précisément, car la reconnaissance vague, « le mécanisme reste écrit par nous », sous-estime le problème :

> *La distribution de mondes est aléatoire* ***conditionnellement à une famille de mécanismes dont le support structurel reste défini par les auteurs.***

Nous avons randomisé les paramètres, les mondes, les structures de covariance, les effectifs et les budgets de complexité. Nous n'avons pas randomisé l'espace des mécanismes admissibles. L'incertitude paramétrique est large ; l'incertitude structurelle reste limitée.

11. Ce qui est contrôlé, ce qui ne l'est pas

Dimension	**Contrôle ou test de résistance**	**Limite résiduelle**
Complexité globale	appariée par la trace, balayée sur un facteur trois (8.5)	spectre complet des opérateurs non égalisé
Incertitude de Monte-Carlo	contrôlée, 2 000 réplications	
Marges des covariables	calibrées sur publications tierces	dépendances réellement observées inconnues
Effectif	testé de 200 à 5 000 (8.4)	régime asymptotique général non atteint
Structure de dépendance	six familles, dont trois à queue (8.7)	espace universel des dépendances non couvert
Famille d'estimation	fixée à M_0	aucune autre famille testée
Mécanisme causal	trois régimes, distribution de mondes (8.3)	support structurel défini par les auteurs
Identifiabilité	oracle et contrôle négatif (8.8)	confusion cachée non détectable en pratique
Inférence par intervalles	couverture mesurée (8.6)	intervalles inutilisables sous biais dominant

« Contrôle ou test de résistance » ne signifie pas neutralisation : plusieurs dimensions sont explorées, non maîtrisées. La colonne de droite dit ce qui reste ouvert dans chaque cas.

Cette carte remplace une lecture linéaire des précautions dispersées dans le texte. Elle est le complément honnête du résumé exécutif.

Une mesure qui ne mesure pas ce que son nom annonce. Le protocole prévoyait de quantifier une incertitude générative par rééchantillonnage du générateur. L'implémentation rééchantillonne les réplications déjà produites, reproduisant la variance de Monte-Carlo : le rapport entre les deux quantités ressort à 1,01. Nous le déclarons comme limite, non comme résultat.

12. Portée de l'inférence et falsification suivante

Résultat établi dans le cadre expérimental préspécifié. Dans cette famille pénalisée de modèles de survie à temps discret, à complexité globale appariée, des transformations représentationnelles distinctes produisent des biais différents malgré une complexité globale appariée, et les dépassements de marge surviennent de manière structurée selon les régimes. Leur interprétation par l'expressivité de la classe fonctionnelle relève du niveau B, non du niveau A. Sur une distribution de mondes, l'avantage est probabiliste : l'axe continu produit un biais inférieur au format large dans 73 % des mondes, et il est le meilleur des trois transformations appariées dans 63 % d'entre eux. Cet avantage ne dépend ni du niveau de complexité retenu, invariant sur un facteur trois, ni de l'accès à une information privilégiée, ni de la régularité de la structure de dépendance des covariables. Il ne se résorbe pas lorsque l'effectif croît de 200 à 3 000 et que la cible de complexité, sélectionnée par validation croisée, augmente avec lui.

Conjecture méthodologique. Le principe devrait s'étendre aux familles pour lesquelles une notion pertinente de complexité effective peut être définie. Conceptuellement plausible, empiriquement non démontré : dans un réseau profond, la trace de l'opérateur de lissage n'a pas d'analogue immédiat.

Le test indépendant prioritaire n'est pas davantage de simulations. Le rendement marginal d'un banc dont les auteurs définissent le support structurel est décroissant, et dix mille mondes supplémentaires n'y changeraient rien. L'expérience qui trancherait est d'une autre nature : un tiers spécifie les mécanismes, les paramètres, les distributions, les interactions et la dynamique temporelle, sans connaître les résultats obtenus par les représentations ; le pipeline est ensuite exécuté en aveugle. Aucune expérience isolée ne sera « décisive » au sens strict, mais celle-ci porterait sur la seule dimension que nos analyses ne peuvent pas atteindre. C'est le prolongement que nous identifions comme prioritaire, avec l'extension à d'autres familles d'estimation, la question R × M × DGP dont ce travail n'étudie qu'une tranche.

Positionnement. Ce travail ne montre pas que certaines représentations sont intrinsèquement meilleures. Il contribue à la méthodologie des comparaisons attributionnelles de représentations. La survie est ici le banc expérimental, non l'objet de la doctrine.

Disponibilité du code

Élément	Référence
Préinscription	OSF DOI 10.17605/OSF.IO/TPG5S
Base gelée	`plasmode.py`, SHA-256 `9ca0aa6137c3a289…792592a`
Run confirmatoire	`plasmode_e2.py`, `run_parallel.py`
Analyses post-hoc	`sensibilite_marge.py`, `decomposition_variance.py`, `analyse_spectrale.py`, `alignement_spectral.py`, `variabilite_mondes.py`, `effectif_complexite.py`, `cible_complexite.py`, `couverture.py`, `dgp_adversarial.py`
Environnement	Python 3.12.3, numpy 2.4.4, scipy 1.17.1, pandas 2.3.3

Le run confirmatoire vérifie l'empreinte de la base gelée au lancement et refuse de s'exécuter si elle diffère. **Aucune donnée patient n'est employée à aucune étape**.

Références

1. Mitchell TM. The need for biases in learning generalizations. Rutgers University Technical Report CBM-TR-117; 1980.
2. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798-1828.
3. Cohen T, Welling M. Group equivariant convolutional networks. In: Proceedings of ICML; 2016.
4. Battaglia PW, Hamrick JB, Bapst V, et al. Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261; 2018.
5. Bronstein MM, Bruna J, Cohen T, Veličković P. Geometric deep learning: grids, groups, graphs, geodesics, and gauges. arXiv:2104.13478; 2021.
6. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv. BMC Med Res Methodol. 2018;18:24.
7. Lee C, Zame WR, Yoon J, van der Schaar M. DeepHit. In: Proceedings of AAAI; 2018.
8. Lee C, Yoon J, van der Schaar M. Dynamic-DeepHit. IEEE Trans Biomed Eng. 2020;67(1):122-133.
9. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. J Mach Learn Res. 2019;20(129):1-30.
10. Bender A, Rügamer D, Scheipl F, Bischl B. A general machine learning framework for survival analysis. In: Proceedings of ECML PKDD. Springer; 2021.
11. Huang X, Khetan A, Cvitkovic M, Karnin Z. TabTransformer. arXiv:2012.06678; 2020.
12. Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. Revisiting deep learning models for tabular data. Adv Neural Inf Process Syst. 2021;34:18932-18943.
13. Somepalli G, Goldblum M, Schwarzschild A, Bruss CB, Goldstein T. SAINT. arXiv:2106.01342; 2021.
14. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? Adv Neural Inf Process Syst; 2022.
15. Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data. BMC Med Res Methodol. 2012;12:9.

16. Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials. Pharmacoeconomics. 2010;28(10):935-945.
17. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. NICE DSU Technical Support Document 18. Sheffield: NICE Decision Support Unit; 2016.
18. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. Med Decis Making. 2018;38(2):200-211.
19. Wood SN. Generalized additive models: an introduction with R. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2017.
20. Hastie T, Tibshirani R. Varying-coefficient models. J R Stat Soc Series B. 1993;55(4):757-796.
21. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. J R Stat Soc Series B. 2005;67(1):91-108.
22. Kim SJ, Koh K, Boyd S, Gorinevsky D. l1 trend filtering. SIAM Rev. 2009;51(2):339-360.
23. Cox DR. Regression models and life-tables. J R Stat Soc Series B. 1972;34(2):187-220.
24. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457-481.
25. Pawel S, Kook L, Reeve K. Pitfalls and potentials in simulation studies. Biom J. 2024;66(1):e2200091.
26. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074-2102.
27. Boulesteix AL, Lauer S, Eugster MJ. A plea for neutral comparison studies in computational sciences. PLoS One. 2013;8(4):e61562.
28. Schreck N, Slynko A, Saadati M, Benner A. Statistical plasmode simulations. Stat Med. 2024;43(9):1804-1825.
29. Buja A, Hastie T, Tibshirani R. Linear smoothers and additive models. Ann Stat. 1989;17(2):453-510.
30. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York: Springer; 2009. §7.6.
31. Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola J. On kernel-target alignment. Adv Neural Inf Process Syst. 2001;14.
32. Canatar A, Bordelon B, Pehlevan C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. Nat Commun. 2021;12:2914.
33. Kaufman S, Rosset S. When does more regularization imply fewer degrees of freedom? Biometrika. 2014;101(4):771-784.