ENSAI — École nationale de la statistique et de l'analyse de l'information

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

CONSTRUCTION D'UN SCORE DE COMORBIDITÉ EN CHIRURGIE HÉPATOBILIAIRE

Comment évaluer le risque de mortalité intra-hospitalière 90 jours après une hépatectomie pour un patient atteint d'une maladie du foie ?

PROJET DE STATISTIQUE 2024

AUTEURS
Léa KERHOUSSE
Fabien STEINMETZ
Hanan TOUZANI

TUTEURS
Professeure Sandrine KATSAHIAN
Docteur Stylianos TZEDAKIS

INTRODUCTION

Historiquement, plusieurs scores sont utilisés dans le domaine médical pour évaluer les risques des patients.
Les scores de Charlson, Quan et Bannay sont parmi les plus utilisés mais sont plutôt généralistes et sont pronostics d'une mortalité à un an.
Cela motive la construction d'un score spécifique à la chirurgie hépatobiliaire, pour évaluer le risque d'une hépatectomie pour un patient présentant des comorbidités dans les 90 jours.

OBJECTIFS

- Définir une méthodologie de construction d'un score de comorbidité spécifique à la chirurgie hépatobiliaire
- Comparer le score obtenu avec les scores de Charlson, Quan et Bannay
- Comparer avec d'autres modèles prédictifs tels que les Random Forest

01. DONNÉES

ENTRE LE 01.01.2015 ET LE 31.12.2019

39,241
PATIENTS OPÉRÉS DU FOIE

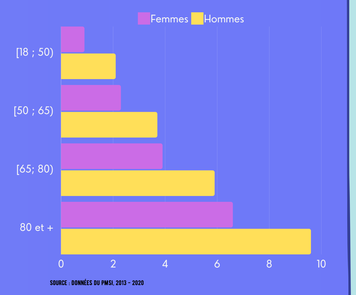
DONT **4 %** DE PATIENTS DÉCÉDÉS*

*DÉCÈS SURVENU À L'HÔPITAL DANS LES 90 JOURS SUIVANT L'HÉPATECTOMIE

ÂGE MÉDIAN
65 ans

PART DES FEMMES
45 %

PART DES PATIENTS DÉCÉDÉS PAR SEXE ET TRANCHE D'ÂGE



DONNÉES DU PROGRAMME DE MÉDICALISATION DES SYSTÈMES D'INFORMATION (PMSI)

02. MÉTHODOLOGIE DE LA MODÉLISATION

- Sélection des comorbidités parmi les 83 présentes selon l'expertise médicale du Dr Tzedakis
- Sélection des comorbidités selon l'étude de leur lien avec la variable de décès

48 COMORBIDITÉS + SEXE + ÂGE*

* EN 4 TRANCHES : MOINS DE 50 ANS, DE 50 À MOINS DE 65 ANS, DE 65 À MOINS DE 80 ANS, 80 ANS ET +

- Classes déséquilibrées car 4 % de patients décédés : nécessité de ré-échantillonner la base d'entraînement des modèles, soit en gonflant le nombre de patients décédés (sur-échantillonnage), soit en sélectionnant aléatoirement parmi les patients ayant survécu (sous-échantillonnage)

RÉ-ÉCHANTILLONNAGE DE LA BASE D'ENTRAÎNEMENT DES MODÈLES

- Méthodes de sélection des variables explicatives pas à pas: backward, forward et stepwise.
- Techniques de pénalisation du modèle de régression : Ridge, Lasso et Elastic Net
- Optimisation du seuil de classification pour améliorer la sensibilité et la spécificité
- Critères : Taux de vrais positifs (sensibilité) et de vrais négatifs (spécificité) supérieurs à 70 %, aire sous la courbe ROC (AUC) et précision

SCORE	PATIENTS VIVANTS OBSERVÉS	PATIENTS DÉCÉDÉS
PATIENTS VIVANTS PRÉDITS	7934	115
PATIENTS DÉCÉDÉS PRÉDITS	3368	354

MATRICE DE CONFUSION AVEC SEUIL DE CLASSIFICATION OPTIMAL

Nous obtenons donc un taux de vrais positifs de 75,48 % et un taux de vrais négatifs de 70,20 %. Ces deux taux dépassent l'objectif de 70 %.

MODÈLE RETENU POUR LE SCORE : ELASTIC NET AVEC SOUS-ÉCHANTILLONNAGE

03. RÉSULTATS : SCORE STYLIANOS

SPÉCIFIQUE À LA CHIRURGIE HÉPATOBILIAIRE ET DE MEILLEURE PERFORMANCE PRÉDICTIVE

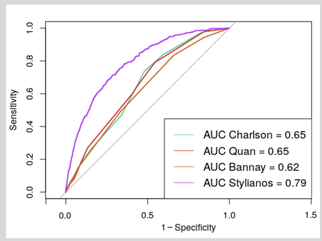
Issu du modèle pénalisé par Elastic Net construit avec une base d'apprentissage sous-échantillonnée.
Le score correspond à la somme des coefficients dans l'équation $\text{logit}(P(Y=1)) = X'\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_d X_d$

Méthode de normalisation : pour continuer avec une approche similaire à celle adoptée lors de la construction des scores de Charlson, nous simplifions le score en le divisant par son maximum, en le multipliant par 10 et en arrondissant à l'unité.
Le pronostic du score simplifié est quasiment identique au modèle de pronostic construit avec les coefficient initiaux de l'Elastic Net avec sous-échantillonnage.
Le score normalisé est donc la somme par individu des poids normalisés des comorbidités, des tranches d'âge et le sexe.

QUARTILES DU SCORE DE STYLIANOS ET PART DE PATIENTS DÉCÉDÉS DANS L'ÉCHANTILLON TEST

SCORE	[−12 ; −2]	[1 ; 2]	[3 ; 6]	[7 ; 38]
PART DE PATIENTS DÉCÉDÉS DANS LES 90 JOURS (%)	0.4	1.7	3.8	10.2

Comparaison avec les scores historiques : Les scores de Charlson, Quan et Bannay sont normalisés entre 0 et 1 en divisant par le maximum atteignable par un individu (respectivement par 27, 26 et 31), afin de les considérer comme des probabilités. La capacité de prédiction de la mortalité à 90 jours pour ces trois scores normalisés est ensuite évaluée grâce une courbe ROC et l'aire sous la courbe.



L'AUC du modèle du score Stylianos est de 0.79, bien au-dessus des AUC des scores de Charlson, Quan et Bannay. En effet, le score Stylianos a été calculé à partir de l'âge, le sexe et un ensemble plus important de comorbidités. De plus, il est davantage ajusté aux données ayant servi à entraîner le modèle (patients opérés d'une hépatectomie).

Score Stylianos : points pas comorbidité

10	SYNDROME HÉPATO-RÉNAL
7	ÂGE >= 80 ANS
5	MALADIE HÉPATIQUE MODÉRÉE À SÉVÈRE
4	INFARCTUS CÉRÉBRAL, COMPLICATION THROMBOEMBOLIQUE
3	ANÉMIE, ASCITE, VARICES GASTRO-ŒSOPHAGIENNES AVEC SAIGNEMENT, JAUNISSE OBSTRUCTIVE, MALNUTRITION, THROMBOSE DE LA VEINE PORTE, ÂGE DE 65 ANS À MOINS DE 80 ANS
2	EMBOLIE ET THROMBOSE ARTÉRIELLES, ARYTHMIE CARDIAQUE, MALADIE RÉNALE CHRONIQUE AVANCÉE, INSUFFISANCE CARDIAQUE CONGESTIVE, JAUNISSE, PARAPLÉGIE/HÉMIPLÉGIE, ULCÈRE GASTRODUODÉNAL
1	CANCER DES VOIES BILIAIRES, MALADIE CÉRÉBROVASCULAIRE, LYMPHOME, INFARCTUS DU MYOCARDE, CANCER PRIMITIF DU FOIE
-1	TROUBLES LIÉS À L'ALCOOL
-2	CIRRHOSE, ENCÉPHALOPATHIE HÉPATIQUE, MALADIE HÉPATIQUE LÉGÈRE, SEXE FÉMININ
-3	CHIMIOTHÉRAPIE POUR LE CANCER, TUMEUR BÉNIGNE DU FOIE
-4	LÉSION KYSTIQUE DU FOIE

04. AUTRE MODÈLE PRÉDICTIF : RANDOM FOREST

Le sous-échantillonnage de la base d'apprentissage permet au modèle de mieux détecter les vrais positifs (patients décédés correctement prédits), et donc la sensibilité.
Après optimisation des hyperparamètres du modèle Random Forest, ce dernier n'atteint pas les performances prédictives du modèle de régression pénalisé avec Elastic Net.

	AUC	SPÉCIFICITÉ	SENSIBILITÉ
MODÈLE ELASTIC NET	0.7872	0.7548	0.7020
MODÈLE RANDOM FOREST	0.7175	0.7335	0.7015

CONCLUSION

Le score Stylianos est un outil d'aide à la décision destiné aux chirurgiens du domaine hépatobiliaire plus performant que les scores historiques. Il pourrait être adapté à d'autres domaines médicaux.
Dans le futur, il faudrait adapter le score Stylianos en fonction de nouvelles maladies (comme la Covid-19) ou de l'évolution des traitements qui réduisent le niveau de gravité d'une comorbidité.
De plus, il est envisageable de tester d'autres méthodes de ré-échantillonnage.
Il serait judicieux d'explorer d'autres modèles prédictifs issus du machine learning tels que XGBoost.
Enfin, il serait possible de construire le score à partir d'un modèle de Cox, à l'instar des scores de Charlson, Quan et Bannay, si l'état d'un patient sorti de l'hôpital est connu.

IMPORTANT !
Un score de comorbidité ne se substitue pas à l'expertise médicale

Vito Toti
Nathan Saraiva
Maxime Tio

Bien-être des porcs en élevages alternatifs en France

Encadré par
Christelle Fablet

Introduction

Peu d'études ont été menées sur les élevages alternatifs en France, qui se développent mais qui restent confrontés à de nombreux défis

Le programme PIGAL, mené par l'ANSES, vise à compenser ce manque de connaissances. C'est dans ce cadre que nos travaux se placent

En particulier, notre étude vise à cerner **les facteurs de bien-être des porcs dans ces élevages alternatifs**

Élevages intensifs
- Élevage en bâtiment sur caillebotis
- Une haute rentabilité économique et une gestion sanitaire aisée

Élevages alternatifs
- Élevage en bâtiment sur litière ou avec accès extérieur
- Valorisation par des labels

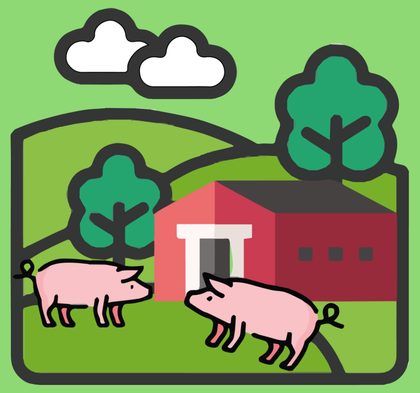
Données

Notre étude porte sur **112 élevages** de trois types :
1) **Naisseurs** (92 élevages), premières étapes du cycle de vie des porcs
2) **Engraisseurs** (100 élevages), phase d'engraissement avant l'abattoir
3) **Naisseurs-Engraisseurs** (80 élevages), englobe les deux précédents

Entre 400 et 800 variables explicatives dans des thématiques variées (alimentation, santé, hygiène, biosécurité, logement, pratiques et généralités)

Une variable à expliquer indiquant le niveau de bien-être animal dans l'élevage

➤ étape de pré-traitement afin de garder les variables les plus significatives pour réduire la multicolinéarité



Méthodes

Régression PLS multi-blocs (MBPLS)

A. Découper nos données en blocs distincts (santé, alimentation, hygiène...), puis construire un modèle pour prédire et expliquer le bien-être des porcs en caractérisant l'influence de chaque bloc et de chaque variable

B. Côté résultats, dans notre cas, si une variable présente un coefficient positif, cela indique qu'elle contribue à l'appartenance au meilleur groupe, et donc au bien-être

Algorithme des K-Means

A. Utiliser les résultats obtenus précédemment avec la méthode MBPLS pour regrouper les données en clusters afin d'identifier le groupe le plus favorable au bien-être

B. Analyser les surreprésentations ou les sous-représentations de certaines variables significatives

Régression Elastic Net

A. Construire un modèle de régression pénalisée avec les facteurs de bien-être étudiés précédemment pour prendre en compte la multicolinéarité et confirmer les analyses de la méthode MBPLS et des K-Means

B. L'interprétation des coefficients est similaire à celle de la méthode MBPLS

Conclusion

L'analyse des résultats obtenus indique une multitude de facteurs de bien-être dans une variété kaléidoscopique de domaines (éleveur engagé, logement décent, alimentation adaptée, environnement sain et propre...)

Ces conclusions pourront être utilisées par les éleveurs afin d'améliorer la qualité de vie des porcs dans leurs élevages

Néanmoins, la filière alternative reste confrontée à de nombreux défis (concurrence avec les élevages intensifs, inflation, gestion des maladies, coûts de production....)

Résultats

Quels sont les principaux facteurs déterminants du bien-être ?

Variables	Coefficient
Choix du système alternatif par passion	0.51
Stockage de la litière empêchant le contact direct avec la terre et les remontées humides	0.48
Localisation géographique de l'élevage en dehors de l'Ouest	0.39
Désinfection des salles de gestation par le personnel	0.38
Porcs non infectés par des parasites à 22 semaines de vie et +	0.32
Âge du bâtiment d'engraissement visité inférieur à 7 ans	0.31
Porcs non infectés par des parasites à 10-12 semaines de vie	0.22
Logement en plein air des truies gestantes	0.21

Régression MBPLS

La vocation des éleveurs

Le choix de l'élevage alternatif par passion est le principal facteur d'appartenance au groupe ayant le bien-être le plus élevé

Un environnement sain et propre

De bonnes pratiques en matière d'hygiène sont importantes pour le bien-être, de plus les maladies et les parasites sont moins fréquents

K-Means

Un logement favorable

Dans le groupe ayant le bien-être le plus élevé, la pratique du plein air est privilégié (90% contre 36% dans l'ensemble), et les bâtiments sont beaucoup plus récents

Une alimentation adaptée

La nourriture est plutôt produite localement (88% dans le meilleur groupe contre 66% au global) et les quantités sont ajustées aux besoins des porcs

Variables	Part de la modalité dans le meilleur groupe	Part de la modalité au global
Logement en plein air des truies gestantes	90%	36%
Fabrication des aliments pour les porcs en engraissement à la ferme	88%	60%
Quantité de nourriture des truies adaptée à la température ambiante	84%	59%
Bâtiment d'engraissement construit il y a moins de 7 ans	59%	24%
Faible nombre de porcs dans l'élevage	48%	28%
Au moins 3kg de nourriture est donnée lors de la mise à bas en été	42%	19%

Variables	Coefficient
Logement en plein air des truies gestantes	0.51
Stockage de la litière empêchant le contact direct avec la terre et des remontées humides	0.5
Présence de bovins à proximité de l'élevage	0.48
Localisation géographique de l'élevage en dehors de l'Ouest	0.39
Paille issue de l'agriculture biologique	0.35
Zone d'élevage délimités empêchant tout contact entre les suidés domestiques et sauvages	0.3
Choix de l'élevage alternatif par passion	0.23
Pratique de l'agriculture biologique	0.22
Bétaillère lavée après tout transport d'animaux	0.21

Régression Elastic Net

Confirmation des résultats

Les coefficients principaux sont positifs et confirment nos analyses précédentes (passion, environnement sain, logement favorable...)

Bonus : Pratiques biologiques

En plus de nos résultats principaux, on y voit la relative importance d'une agriculture certifiée biologique, par exemple via l'usage de paille biologique

BALAYAGE VISUEL DANS LE FOOTBALL

Facteurs influents et reconnaissance des scans

Tuteur : Anthony Sorel - Laboratoire M2S

Guilhem Figueres-Sarda
Corentin Hazeaux
Titouan Laurent
Baptiste Ortillion

Contexte

Le football est le sport le plus populaire au monde. Les **prises d'informations et de décisions** sont souvent clés dans la victoire tant au niveau offensif que défensif. Le **balayage visuel** est la meilleure solution pour prendre connaissance de ce qui entoure un joueur sur le terrain.

Pour cela, des données de matchs ont été collectées sur les joueurs de la génération 2008-2009 de **l'INF Clairefontaine**, impliquant des matchs de **"stop ball"** en cinq contre cinq, un jeu utilisé pour améliorer les compétences de jeu dans les petits espaces et les transitions rapides.

Objectif: Comparer les performances avant et après un entraînement spécifique aux scans visuels à l'aide de la réalité augmentée et développer un algorithme de reconnaissance afin d'identifier les scans parmi les données brutes.

Qu'est-ce qu'un scan ?

Il s'agit d'un **mouvement rapide** de la tête d'un joueur pour explorer activement son environnement de jeu. Ces mouvements de tête sont récupérés à l'aide de capteurs placés sur l'arrière du crane.

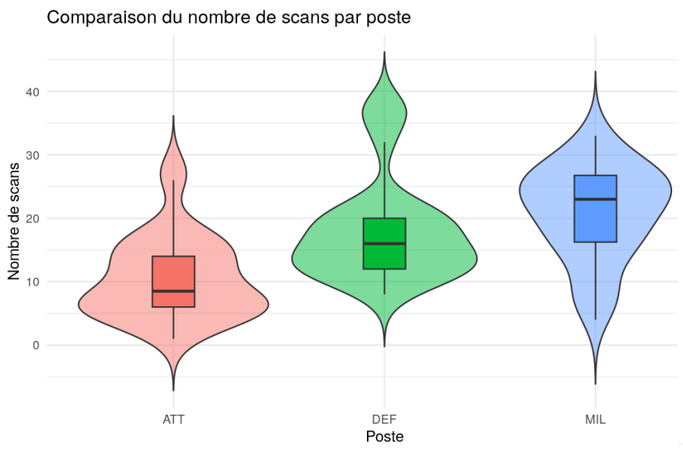
Les scans durant le jeu peuvent varier en intensité et en fréquence :

- **Scans simples :** mouvements de tête à gauche ou à droite
- **Scans doubles (triple, multiple...) :** plusieurs mouvements de tête successifs dans la même direction.
- **Scans complets :** balayage des deux côtés pour couvrir tout le terrain.

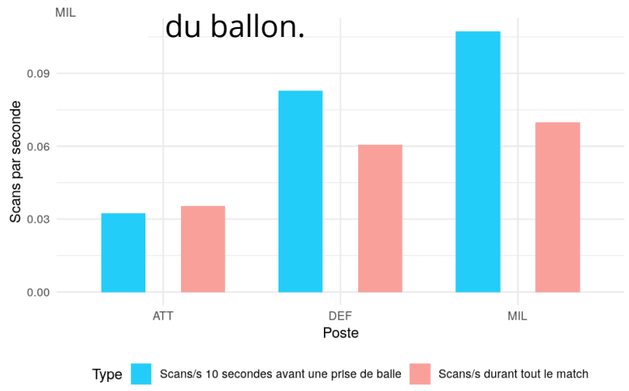
Les facteurs influençant le balayage visuel

Le poste du joueur : un facteur clé

Bien qu'une équipe bien organisée, communicative et en possession du ballon favorise une meilleure perception du jeu par les joueurs, il apparaît que le nombre de scan d'un joueur **n'est pas influencé** par son **appartenance une équipe** ni par le **nombre de passe** qu'il réalise, ce qui peut être contraire à une intuition initiale.



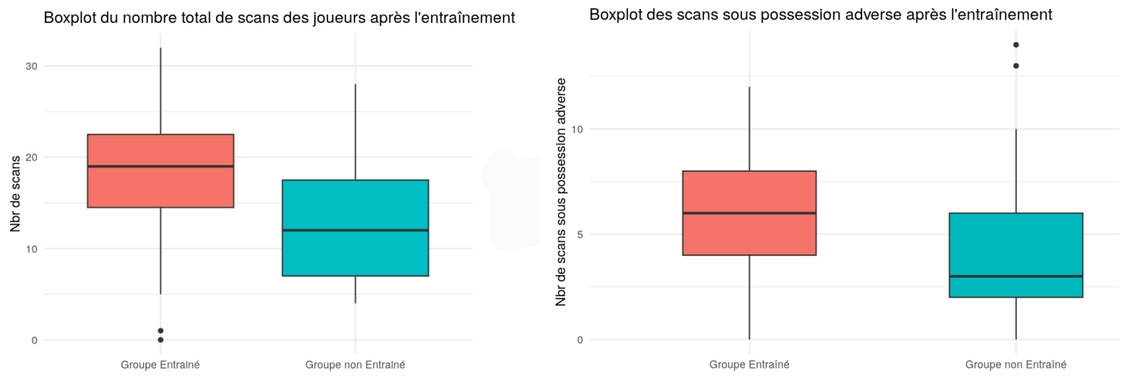
Cette analyse, complétée par des tests ANOVA, indique que le nombre de scans est lié au **poste** : comme on pourrait s'y attendre les **milieux**, suivis par les **défenseurs**, prennent plus souvent l'information que les attaquants. Cela peut s'expliquer par le fait que ces postes ont un rôle central dans la **distribution** du ballon.



De plus, **25%** des scans totaux se produisent dans les **10 secondes avant une prise de balle**, avec une fréquence plus élevée chez les milieux de terrain et les défenseurs.

La VR a amélioré les performances

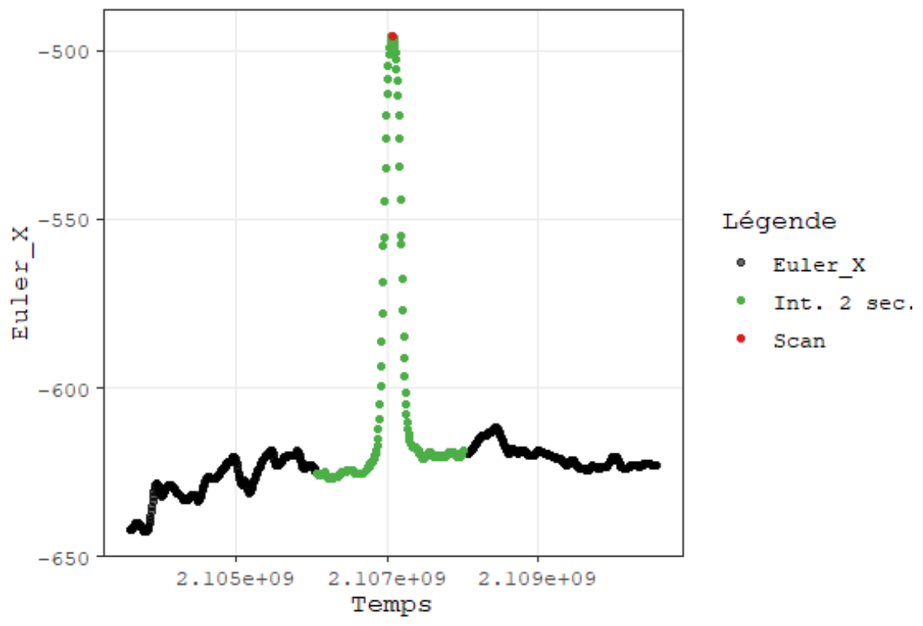
Parmi les 17 joueurs participant à l'expérience, **9** ont suivi l'entraînement en réalité virtuelle. Le test par permutation a révélé une **différence significative** entre les deux groupes, suggérant que l'**entraînement a été bénéfique** en augmentant la fréquence de balayage des joueurs, notamment lors des **phases défensives** où leur capacité de scanning s'est fortement **améliorée**.



Les performances des scans des joueurs sont impactées par divers facteurs tels que leur poste et la possession de l'équipe. L'entraînement spécifique en réalité virtuelle a augmenté la fréquence des scans, surtout en situation défensive.

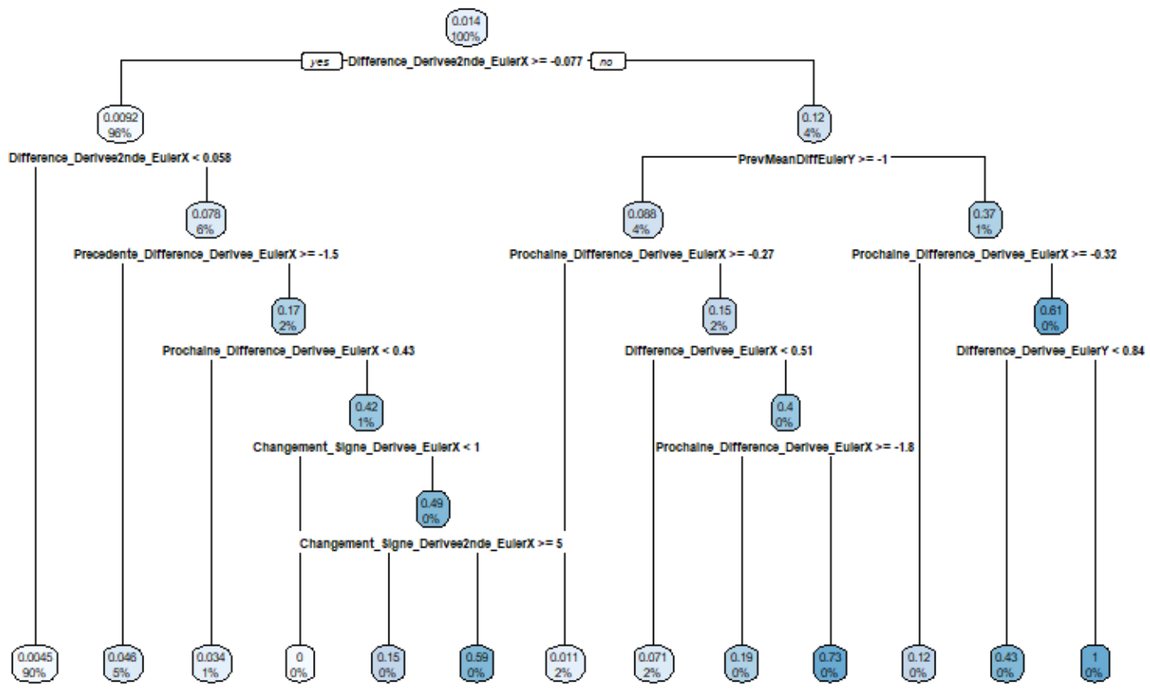
Reconnaissance graphique d'un scan

Représentation visuelle d'un scan



Les données de **Euler_X** décrivent la **rotation** de la tête du joueur selon l'axe X (de la gauche vers la droite). Chaque scan a une **durée** et un **amplitude** différentes. Ces deux paramètres sont importants pour la reconnaissance d'un scan.
Un scan dure entre un quart de seconde et une seconde et demie et a une amplitude allant du simple coup de tête (20°) jusqu'à un demi tour (180°). Bien que les amplitudes des **scans simples et doubles** tendent à être plus **extrêmes**, les moyennes entre les différents types de scans sont **similaires**, comme indiqué par le test d'égalité des moyennes.

Algorithme de reconnaissance des scans



En obtenant un nouveau jeu de données, nous pouvons examiner les **moyennes des valeurs des variables** ci-dessus à l'aide d'un **arbre CART** pour déterminer si un intervalle de valeurs tombe dans une case avec un **score** final **supérieur à 0.5**, ce qui indiquerait un **scan**, ou un score final **inférieur à 0.5**, ce qui indiquerait l'**absence** de scan sur cet intervalle.

La détection des scans dépend de la durée et de l'amplitude des mouvements. L'algorithme différencie les non-scans et atteint un taux de 62,5% de vrais scans dans les intervalles identifiés. Les variables clés sont la moyenne des dérivées seconde et première de Euler_X.

Comparaison appariée des méthodes de conservation d'échantillons de microbiote intestinal

ENSAI École nationale de la statistique et de l'analyse de l'information

Djuffo Kuété Sidoine
Kougoum M. Marilene
Marinoni Erwan

FACULTÉ DE PHARMACIE DE PARIS — UNIVERSITÉ PARIS DESCARTES — Université de Paris

Tuteur: Dr. Ionnanis Nicolis

Contexte

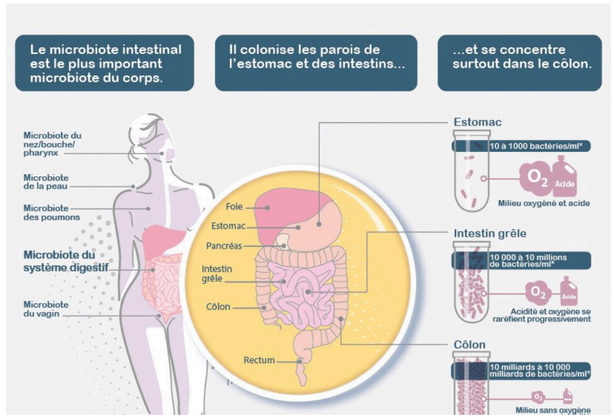
Après la naissance, les humains abritent un écosystème microbien complexe vital, notamment dans leur intestin, contribuant au développement immunitaire et à la santé globale. Sous la tutelle de la **Faculté de pharmacie de Paris**, cette étude examine l'impact des méthodes de conservation des échantillons de microbiote intestinal sur leur intégrité micro-biologique.

Objectif

L'objectif principal de cette étude **est d'effectuer une analyse comparative entre différentes méthodes de conservation alternatives d'échantillons de microbiotes intestinal.** La **méthode de référence** est la **congélation immédiate**.

Les concepts clés à définir :

Microbiote intestinal :



Source: Microbiote intestinal © PixScience pour l'Inserm

Le microbiote intestinal décrit l'ensemble des micro-organismes (bactéries, virus, champignons, etc.) présents dans les intestins. Situé dans l'intestin grêle et le côlon, c'est le microbiote le plus peuplé de l'organisme, abritant jusqu'à 10^14 microorganismes.

Genre bactérien :

Groupe d'organismes regroupant des espèces similaires sur la base de caractéristiques géné tiques et physiologiques.

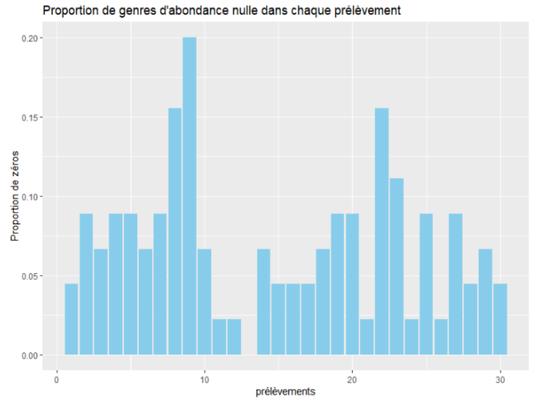
Compositionnalité des données microbiennes:

Elle renvoie au fait que les données microbiennes représentent des proportions relatives plutôt que des mesures absolues.

Données

- Nous disposons de **10 échantillons**, chacun **scindé en trois**, chaque partie étant soit **congelée** (CI), soit conservée à **température ambiante** (TA), soit conservée à l'aide d'une **solution de conservation** (OG).
- Les données disponibles sont des données de séquençage qui donnent les abondances de 43 genres microbiens dont 42 sont observés dans au moins 4 échantillons et ayant une abondance totale au moins égale à 1% du maximum des abondances totales .

Les données sont plus ou moins éparses: 90% des prélèvements,on a moins de 10% de variables ayant une abondance relative de 0.



Source: Travaux des auteurs sur R

- Les abondances microbiennes de nos données sont compositionnelles.

Méthode

La démarche méthodologique se décline en 3 principales étapes:

- Comparaison des indices de diversité intrinsèque (diversité alpha) : **Shannon, inverse de Simpson, Chao** ;
- Comparaison des diversités inter-échantillon (diversité béta): plusieurs approches ont été adoptées : **une ordination non contrainte PCoA, une CAP et ACP précédée d'une transformation CLR;**
- Mise en oeuvre d'une analyse d'abondance différentielle en utilisant l'algorithme MaAslin2.

Présentation des résultats

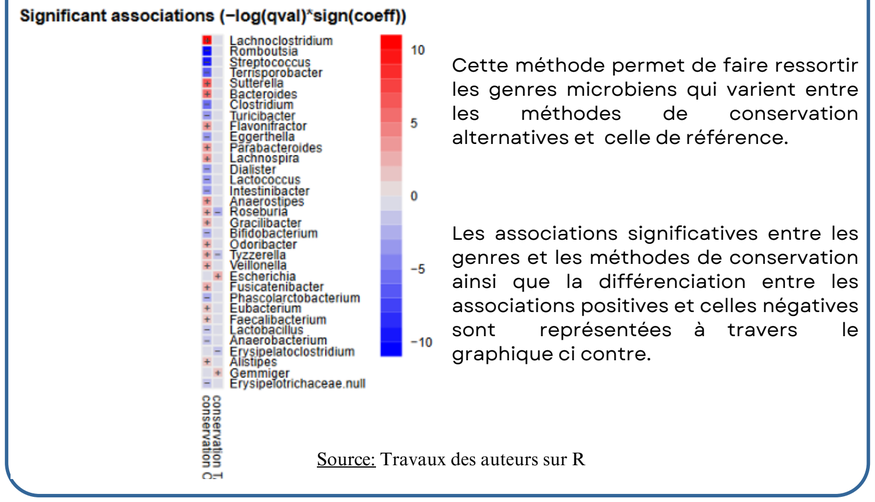
- **Stabilité de la diversité intrinsèque: Cas de l'indice inverse de Simpson: Cas de l'indice inverse de Simpson**

La diversité intrinsèque varie au sein des échantillons lorsqu'on passe de la conservation par congélation immédiate à celle par une solution de conservation. En revanche, on note une stabilité bactérienne lorsqu'on compare les échantillons conservés par congélation immédiate et ceux par température ambiante.

- **Comparaison des diversités inter-échantillon**

La composition des prélèvements conservés à l'aide d'une solution de conservation est significativement différente de ceux conservés par congélation immédiate.
En revanche, les méthodes de conservation congélation immédiate (CI) et température ambiante (TA) sont très similaires.

- **Analyse d'abondance différentielle**



Cette méthode permet de faire ressortir les genres microbiens qui varient entre les méthodes de conservation alternatives et celle de référence.

Les associations significatives entre les genres et les méthodes de conservation ainsi que la différenciation entre les associations positives et celles négatives sont représentées à travers le graphique ci contre.

Source: Travaux des auteurs sur R

Conclusion et discussion

La conservation à température ambiante représente une meilleure alternative comparativement à la conservation à l'aide d'une solution de conservation.

Le **jeu de données** étant de **taille restreinte**, il faut être **précautionneux quant à l'exploitation des résultats** de nos analyses. Au niveau des analyses d'abondance différentielle notamment, certaines différences d'abondance significatives entre les méthodes de conservation alternatives et celle de référence **reposent juste sur la présence de quelques individus atypiques**.

Comment bien grandir à Paris ?

Analyse d'une consultation citoyenne par NLP

CONTEXTE

La question "*Comment bien grandir à Paris ?*" a suscité la mobilisation de milliers de Parisiens qui souhaitent améliorer les conditions de vie de leurs enfants. Les participants ont partagé un total de **7 128 propositions**, regroupant leurs expériences, réflexions et idées. Ces propositions ont été recueillies sur la plateforme **Make.org** qui est une plateforme en ligne visant à mobiliser les citoyens en leur proposant de soumettre des propositions en réponse à une problématique donnée.

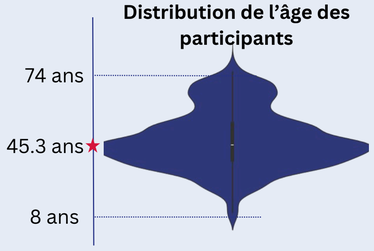
Un peu de statistiques descriptives...

114 911 participants
10 167 propositions, dont **7 128** validées
2,5 millions de votes sur l'ensemble des propositions

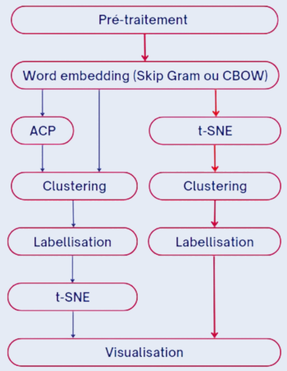
Objectif :
Mettre en place une **labellisation automatique** des propositions par **NLP**, qui rivalise avec la labellisation manuelle des modérateurs du site.

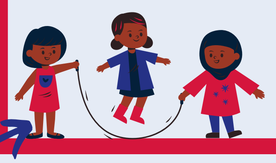
Nuage des mots les plus fréquents

Distribution de l'âge des participants



DÉROULÉ



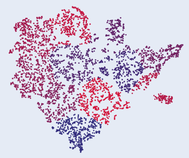


Étape 1 : T-SNE

Projection en 2D des vecteurs propositions

Le **t-SNE** (t-distributed stochastic neighbor embedding) est une méthode permettant de représenter des points d'un espace à grande dimension dans un espace de **2 dimensions**. On cherche une configuration selon un critère de **théorie de l'information** afin de conserver la proximité entre les points tout en réduisant la dimension.

Étape 2 : Clustering K-MEANS



Étape 3 : Labellisation des clusters

En examinant les mots les plus fréquents des propositions de chaque cluster, on arrive à en tirer un label par exemple ci-dessus : **Espace Verts**

PRÉPARATION DES DONNÉES

Étape 1 : Tokenisation
"il faut renforcer la sécurité le soir"
↓
"il", "faut", "renforcer", "la", "sécurité", "le" , "soir"

Étape 2 : Normalisation et suppression des stop words
"il", "faut", "renforcer", "la", "sécurité", "le" , "soir"
↓
"il", "faut", "renforcer", "sécurité", "soir"

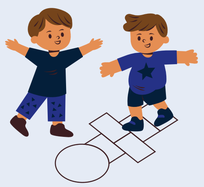
Étape 3 : Lemmatisation
"il", "faut", "renforcer", "sécurité", "soir"
↓
"il", "falloir", "renfort", "sécurité", "soir"

Word Embedding

Une fois les données pré-traitées, nous allons utiliser un modèle d'embedding **Word2Vec** pour convertir les mots en vecteurs numériques de grande dimension, qui capturent les relations sémantiques et contextuelles entre les mots.

$$\text{Vecteur Proposition} = \frac{\sum_{Proposition} \text{Vecteurs Mots}}{\text{Nb de mots}}$$

Le coin python
NLTK (Natural Language Toolkit) est une bibliothèque de NLP en Python. Elle fournit des outils pour l'analyse, la manipulation et l'apprentissage automatique de données textuelles.

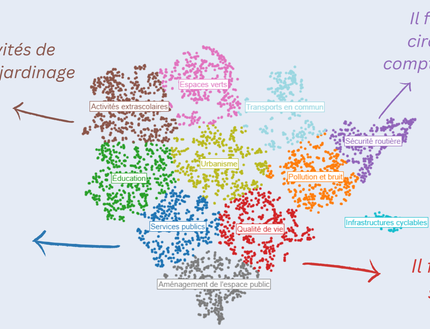


RÉSULTAT

Labellisation des propositions

Il faut favoriser la création d'activités de proximité pour les jeunes, comme le jardinage ou le gardiennage.

Il faut vider les poubelles pleines, y compris les dimanches et jours fériés



Il faut installer des feux de circulation qui prennent en compte la sécurité des piétons et des cyclistes.

Il faut renforcer la sécurité le soir

Conclusion :
Après clustering, on obtient 11 labels couvrant tous les thèmes pouvant répondre à la problématique initiale. Les techniques de NLP ont permis de rendre automatique la labellisation des propositions.

Limites de l'étude :
- Il y a une part de subjectivité pour trouver le modèle le plus puissant et le noms des clusters
- Certaines propositions sont atypiques et ne correspondent réellement à aucun cluster

Étudiants : BLYAU Chloé - COTTET Coralie - DELMARE Bastien
Tutrice : GAY Anne-Cécile
Coach : JUGDE Éric

Intégrer le risque climatique dans le risque de marché

Contexte

Notre objectif est d'élaborer une méthodologie évaluant l'impact du risque climatique sur les valeurs boursières. Étant donné la multiplicité des facteurs de risque, attribuer une variation boursière à une crise spécifique est complexe. Ainsi, nous visons d'abord à isoler et quantifier l'impact du risque climatique.

Risques bancaires

Il existe 4 types de risques bancaires

- **Risque de crédit** : risque qu'un emprunteur ne rembourse pas ses dettes selon les termes convenus
- **Risque de liquidité** : risque qu'une entité ne puisse pas liquider des actifs ou obtenir des fonds pour répondre à ses obligations financières
- **Risque opérationnel** : risque de perte résultant de défaillances internes
- **Risque de marché** : risque de perte dû à des variations des conditions de marché

Risques climatiques

Il existe 3 types de risques climatiques

- **Risque physique** : Risque de dommages directs aux biens, infrastructures ou ressources naturelles causés par des événements physiques tels que des catastrophes naturelles ou des accidents
- **Risque de transition** : Risque financier lié aux adaptations nécessaires pour répondre aux changements dans la politique, le marché, la technologie et les préférences des consommateurs visant à atténuer le changement climatique
- **Risque de réputation** : Risque de perte résultant d'une perception négative de l'entreprise par le public

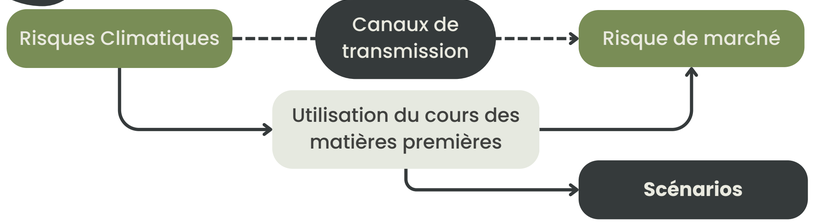
Matières premières

Les risques climatiques influent sur le risque de marché par divers canaux. Pour estimer le cours d'une action en tenant compte de ces risques, l'**analyse des marchés à terme des matières premières**, directement impactées par les changements climatiques, est pertinente. Le choix de ces matières premières est crucial pour obtenir des **prédictions** précises.

Notre stratégie

Les canaux de transmission des risques climatiques au risque de marché sont difficiles à estimer . Nous avons donc fondé notre étude sur les matières premières car ces dernières sont impactées par les évolutions climatiques et interviennent également dans les coûts de production d'entreprises. Ces matières premières vont constituer le socle de notre estimation de densité avec la construction de scénarios.

Notre approche



Méthode At-Risk

L'objectif de la méthode AtRisk est de déterminer la **densité de la distribution conditionnelle** aux variables significatives d'une action à un horizon temporel fixé. Elle se démarque par sa capacité à prédire plus qu'une espérance ou un comportement asymptotique.

1. Estimation des **quantiles empiriques**
2. Estimation des **quantiles théoriques** au temps **t+h** à partir du temps t
3. Méthode des moments : calcul des **paramètres de la loi**
4. Visualisation des métriques de risque et de la densité conditionnelle

Ces trois premières étapes constituent notre modèle de régression quantile. Nous sommes ensuite capables de prédire deux cas :

- un **scénario historique** pour lequel notre prédiction se base sur des **données réelles** au temps t
- un **scénario prédictif** pour lequel notre prédiction s'appuie sur de **nouvelles données plausibles** mais hypothétiques

Affichage

- Graphique 2D
- Graphique 3D
- Densité

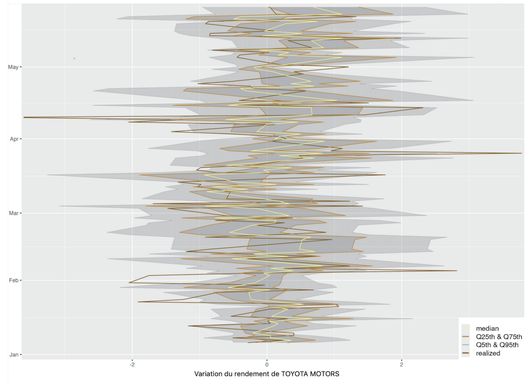
Cas pratique

Ce scénario historique rend compte d'un risque de transition que nous avons identifié sur le palladium (un matériau essentiel au secteur automobile). Le cours de ce métal pouvant encore fluctuer de nos jours, nous cherchons ici à estimer le risque sur le rendement de l'action de TOYOTA MOTORS avec une prévision à un horizon d'un mois.

Comparaison entre le rendement réel de l'action de Toyota et les différents quantiles estimés à horizon d'un mois

Les quantiles représentés en gris sur ce graphe sont ceux calculés à partir de données historiques dont la dernière temporalité a été observée un mois à l'avance.

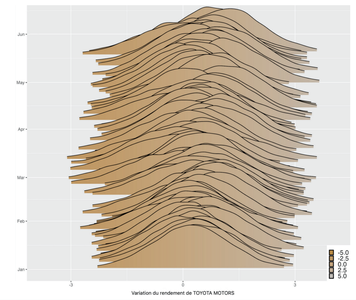
On observe ainsi que l'évolution historique du **rendement de l'action de Toyota** sur les cinq premiers mois de 2023 est pour l'essentiel **contenue entre nos 5èmes et 95èmes quantiles estimés** : ce graphe permet de valider la cohérence de notre estimation adossée aux valeurs historiques.



Évolution de la densité du rendement de l'action de Toyota conditionnelle aux matières premières significatives à horizon d'un mois

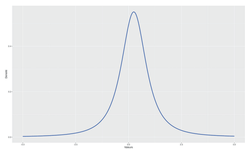
Cette représentation nous permet d'observer l'évolution de la variation de notre risque sur cinq mois (du 01/01/2023 au 30/05/2023).

L'évolution temporelle de ces densités révèle que celles observées en mai 2023 semblent légèrement plus aplaties que celles de janvier 2023, indiquant un **risque accru** vers la fin de la période étudiée. On constate un déplacement de ces pics vers la droite avec le temps, suggérant une **augmentation du rendement moyen** au fil du temps.



Densité de la distribution conditionnelle du rendement de l'action de Toyota à horizon d'un mois à partir des données actuelles

Paramètre	Densité à 1 mois
ξ	0.2458
ω	0.6358
α	-0.0583
ν	1.8521



La densité conditionnelle estimée du rendement de l'action de Toyota à horizon d'un mois suit une loi de Student, les paramètres sont :

- **ξ (localisation) :** représente la moyenne attendue du rendement de l'action.
- **ω (échelle) :** représente l'écart-type du rendement de l'action. Une valeur élevée de ω indique une forte volatilité de ce rendement et une plus grande incertitude quant au rendement réel.
- **α (asymétrie) :** une tendance à l'asymétrie négative indique que les rendements inférieurs à ξ sont légèrement plus probables par rapport aux rendements positifs de même ampleur.
- **ν (degrés de liberté) :** détermine la forme de la distribution.

La densité conditionnelle estimée du rendement de l'action de Toyota à horizon d'un mois suit une distribution de Student avec une moyenne attendue positive, une volatilité modérée, une légère asymétrie négative et une forme proche d'une distribution normale.

Conclusion

Cette méthodologie offre des perspectives de calcul intéressantes : nous avons pu estimer une densité conditionnelle aux valeurs significatives de notre régression quantile pour l'action de Toyota à 1 mois mais nous devons quand même discuter de sa qualité. Il aurait été intéressant d'inclure d'autres facteurs économiques et financiers à nos régressions (PIB Japon, cours des matières plastiques...). À travers cette approche, nous avons pu exploiter, visualiser et interpréter des modèles sur plusieurs cours d'actions pour mieux comprendre l'influence du climat sur ce secteur d'activité.

LE DANTEC Julien, POLLIER Fausto, VIELLARD Terence
Tuteur : BARTEAU Robin, DAYMIER Sarah

Estimation du risque de décès par cancer du poumon radio-induit dans une cohorte de mineurs d'uranium

École nationale de la statistique et de l'analyse de l'information

Élèves: Yanis COLLÉAUX , Marceau LE BOT , Jules Christian NIYO D

Tutrice: Sophie ANCELET

INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Contexte

L'exposition au **radon**, gaz cancérigène émis par la désintégration de l'uranium, est un problème de santé publique majeur. Les mineurs d'uranium sont particulièrement exposés à ce gaz pouvant entraîner de graves conséquences tel que le cancer du poumon. Comprendre la relation entre l'exposition au radon et le risque de décès par cancer du poumon est essentiel en termes de sécurité et de santé publique.

Objectif : Évaluer l'impact de l'exposition au radon sur le risque de décès par cancer du poumon chez les mineurs d'uranium tchèques.

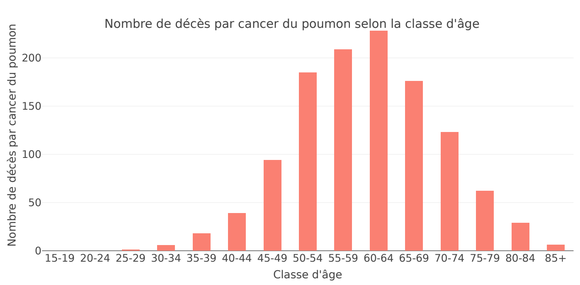
Données

Les données proviennent d'une cohorte tchèque de mineurs d'uranium de 9978 hommes répartis en deux sous-cohortes, et suivis de 1946 à 2012.

Deux bases d'informations sont à disposition dans cette étude :

- **Des informations agrégées**, stratifiées par période calendaire, âge, durée d'emploi et exposition cumulée au radon.
- **Des informations individuelles** sur l'âge, l'exposition cumulée au radon et le décès ou non par cancer du poumon.

Les analyses exploratoires montrent des corrélations entre l'âge, l'exposition au radon et le risque de décès par cancer du poumon.



Méthodologie

La méthodologie adoptée se base sur :

- L'utilisation de la **régression de Poisson** en excès de risque relatif ;
- L'examen de **modèles de survie** : un modèle de Cox et un modèle en excès de risque instantané (EdRI).

Des hypothèses sur le risque de base ont été formulées afin de renforcer la robustesse des résultats. Le modèle en excès de risque relatif s'écrit :

$$\begin{cases} Y_i \overset{\text{ind}}{\sim} \text{Poisson}(\text{PYR}_i \times \lambda_i) \\ \lambda_i = d_{i,0} \times (1 + \beta \times X_i^{cum}) \end{cases}$$

À travers 3 niveaux de stratification pour modéliser le risque de base $d_{i,0}$, les paramètres d'intérêt sont estimés en **fréquentiste et en bayésien.**

Résultats 1 : Poisson

Sur les **données agrégées**, la régression de Poisson aboutit aux résultats suivants :

- Sans hypothèse sur le risque de base, un décalage est notable entre l'inférence **fréquentiste** et **bayésienne**.
- Avec hypothèse sur le risque de base et selon le critère AIC en **fréquentiste** (DIC en **bayésien**), le meilleur modèle a un risque de base qui dépend de l'âge et de la période calendaire.

Interprétation : L'augmentation d'une unité de la moyenne des expositions au radon (WLM) entraîne une multiplication du risque de décès par cancer du poumon par environ 2.123 (2.146 en **bayésien**), toutes choses égales par ailleurs.

Résultats 2 : Modèles de survie

Sur les *données individuelles*, une estimation par **maximum de vraisemblance partielle** a donné :

Modèle	$\widehat{\beta}$	HR
Cox	0.382	1.465
EdRI	2.119	3.119

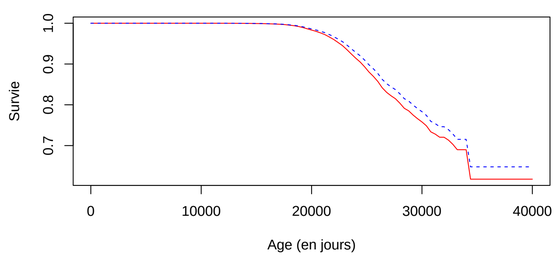


Figure 1: Fonctions de survie obtenues par l'estimateur de Breslow pour le modèle de Cox (*bleu*) et le modèle EdRI (*rouge*)

Le modèle EdRI est le plus pertinent et révèle un **Hazard Ratio (HR) de 3.119**. Les mineurs exposés à 1 WLM ont un risque de décès trois fois plus élevé que ceux non-exposés. Le risque de base a ensuite été considéré comme constant par morceaux. Cela donne l'estimation par **maximum de vraisemblance totale** suivante :

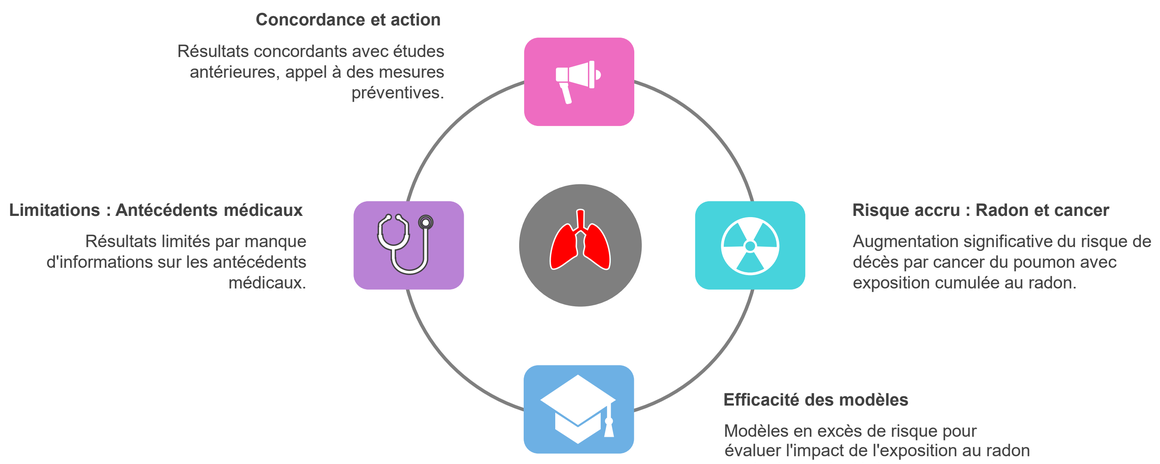
Modèle	$\widehat{\beta}$	HR
Cox	0.462	1.587
EdRI	1.372	2.372

L'estimation **bayésienne** montre à nouveau un modèle EdRI plus pertinent (DIC) et donne les résultats suivants :

Modèle	$\widehat{\beta}$	HR
Cox	0.396	1.49
EdRI	2.387	3.387

Conclusion

En conclusion, 4 points méritent d'être mis en avant :



Pour des résultats plus fins, l'obtention de données plus précises sur l'évaluation de l'exposition au radon serait possible en effectuant une **dosimétrie** de la quantité de radon réellement inhalée par les mineurs au niveau des poumons.