



# **Inférence statistique inductive et apprentissage automatique interprétable avec le package « ale »**

**Chitu Okoli**

**Professor of Digitalization  
SKEMA Business School, Paris**

**State of the R**

**AgroParisTech, Palaiseau**

**17 février 2026**

# Résumé

- Introduction
- Introduction au package ale
- Statistiques ALE et l'inférence basée sur l'ALE
- Analyse d'un PETIT jeu de données avec ale
- Analyse d'un GRAND jeu de données avec ale
- Discussion

# Introduction



# Biographie professionnelle



- Nationalités : nigérienne et canadienne



- Avant carrière universitaire : analyste en informatique et administrateur des ressources humaines



- Ph. D. 2003 de Louisiana State University, Baton Rouge, États-Unis



- 2003 à 2017 : professeur (assistant et associé) dans John Molson School of Business, Concordia University, Montréal, Canada



- Depuis 2017 : professeur en digitalisation à SKEMA Business School, Paris



- Cours enseignés :



- Actuels : Méthodologies de recherches ; Intelligence des affaires ; Gestion des systèmes d'information

Sources d'images : [Wikimedia Commons](#)

# **Le principal défi de la découverte scientifique : découvrir les véritables effets et exclure ceux qui sont fallacieux**

- Pour une analyse scientifique, nous avons besoin à la fois de précision et d'interprétabilité.
- Pour identifier les effets réels, utilisez le modèle prédictif le plus précis disponible.
  - Pour comprendre la nature des effets identifiés, le modèle doit être interprétable.
- Pour éviter les faux positifs (relations fallacieuses)...  
Les approches déductives et inductives de la découverte scientifique diffèrent radicalement.

# Le besoin de mesures de taille d'effet indépendantes du modèle

Les modèles non linéaires peuvent mieux capturer les relations dans les données que les modèles linéaires classiques.

- Mais il est difficile de caractériser la taille d'effet globale d'une variable pour quantifier de telles relations.

# ***Accumulated Local Effects (ALE, effets locaux accumulés)***

## **pour l'interprétabilité visuelle**

Explications globales des algorithmes d'apprentissage automatique boîte noire (Apley et Zhu 2020).

Décomposition fonctionnelle d'un modèle sans distorsions lorsque les variables interagissent (Molnar, Casalicchio et Bischl 2020).

- Calcul relativement rapide (Molnar 2022).

# Introduction au package ale



# Installer version de développement du package ale

- Installer package pak pour pouvoir installer packages sur GitHub
  - `install.packages('pak')`
- Installer version de développement du package ale
  - `pak::pak('tripartio/ale')`
- Dirigez-vous vers le site web du package
  - <https://tripartio.github.io/ale/>

# Introduction to the ale package

- <https://tripartio.github.io/ale/articles/ale-intro.html>

# Statistiques ALE et l'inférence basée sur l'ALE



# La procédure classique de découverte utilisant la déduction par la vérification d'hypothèses

Proposez des hypothèses à l'avance, puis testez si les données ne rejettent pas l'hypothèse.

- Pour identifier les effets réels, utilisez une variante appropriée du modèle linéaire généralisé (GLM) :
- Relativement interprétable
  - Précision acceptable
- Pour éviter les faux positifs (relations parasites), formulez des hypothèses antérieures et utilisez des corrections de valeur p pour plusieurs hypothèses.
  - Évitez le p-hacking et le HARKing (hypothèses après la conscience des résultats).

# Limites de l'approche classique de test d'hypothèses déductives

- Proposez des hypothèses à l'avance, puis testez si les données ne rejettent pas l'hypothèse.
- Mais cela demande des analyses brillantes de salon.
  - Mais cette procédure va à l'encontre de la découverte scientifique exploratoire (et probablement à l'encontre de la quantité réelle de science).
- Pour identifier les effets réels, on utilise une variation appropriée du modèle linéaire généralisé (GLM).
  - Mais les coefficients seuls ne sont pas aussi interprétables qu'on le prétend souvent.
  - Mais la précision est souvent inférieure à celle des techniques plus modernes.
- Les visualisations peuvent améliorer l'interprétabilité.
  - Mais ils manquent de la rigueur de la vérification formelle d'hypothèses.
- Pour éviter les faux positifs (relations parasites), formulez des hypothèses antérieures et utilisez des corrections de valeur p pour plusieurs hypothèses.
  - Mais même s'il est beaucoup plus lent, le bootstrapping est très robuste contre le p-hacking.

# Mesures existantes et leurs limites

- Complexité de l'effet principal pour indiquer l'écart d'une variable par rapport à une tendance linéaire (Molnar, Casalicchio et Bischl 2020)
  - Aborde la non-linéarité mais pas d'autres aspects des effets.
- D de Cohen pour apprentissage automatique, en tenant compte des tendances centrales et des variations au sein et entre les groupes (Lötsch et Ultsch 2020)
  - Montre les effets prédictifs globaux mais manque de détails sur la nature des effets.
- $f^2$  de Cohen pour l'apprentissage automatique, basé sur l'erreur quadratique moyenne avec monotonie et directionnalité (Messner 2023)
  - Montre la direction des effets mais peut ne pas détailler suffisamment les relations complexes.
- Force d'interaction pour indiquer l'effet d'interaction d'un modèle (Molnar, Casalicchio et Bischl 2020)
  - Mais cela n'indique que les effets moyens sur toutes les variables, pas les effets individuels.
- H-statistique pour quantifier les interactions entre variables, avec des formes par paires et globales par caractéristique (Friedman et Popescu 2008)
  - Confirme l'existence des interactions mais ne décrit pas leur nature ou leur forme.

# ALE-based statistics for statistical inference and effect sizes

- <https://tripartio.github.io/ale/articles/ale-statistics.html>

# Analyse d'un PETIT jeu de données avec ale



# Analyzing a Small Rice Yield Dataset with ALE-Based Inference

- <https://tripartio.github.io/ale/articles/ale-gomez-rice.html>

# Analyse d'un GRAND jeu de données avec ale



# Analyzing a Large Corn Yield Dataset with ALE-Based Inference

- <https://tripartio.github.io/ale/articles/ale-acdc-corn.html>

# Discussion



# Travaux futurs

- Appliquer l'inférence basée sur l'ALE dans les études concrètes.
- Distributions statistiques paramétriques pour un calcul rapide des valeurs p.
- Développement détaillé de l'ALE pour les effets d'interaction.
- Implémentez des outils d'inférence basés sur l'ALE en Python et d'autres systèmes qui implémentent l'ALE.
- Comparaisons avec des mesures alternatives indépendantes du modèle

# SKEMA BUSINESS SCHOOL

**AT HOME WORLDWIDE\***

*\* Chez vous, partout dans le monde*

# References for this presentation

- Apley, Daniel W. and Jingyu Zhu. 2020. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. Journal of the Royal Statistical Society Series B: Statistical Methodology 82, 4, 1059–1086. <https://doi.org/10.1111/rssb.12377>
- Fanaee-T, Hadi. 2013. Bike Sharing. <https://archive.ics.uci.edu/dataset/275>
- Friedman, Jerome H. and Bogdan E. Popescu. 2008. Predictive Learning via Rule Ensembles. The Annals of Applied Statistics 2, 3, 916–954. <https://www.jstor.org/stable/30245114>
- Lötsch, Jörn and Alfred Ultsch. 2020. A non-parametric effect-size measure capturing changes in central tendency and data distribution shape. PLOS ONE 15, 9, e0239623. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0239623>
- Messner, Wolfgang. 2023. From black box to clear box: A hypothesis testing framework for scalar regression problems using deep artificial neural networks. Applied Soft Computing 146, 110729. <https://www.sciencedirect.com/science/article/pii/S1568494623007470>
- Molnar, Christoph. 2022. Interpretable Machine Learning. <https://christophm.github.io/interpretable-ml-book/>
- Molnar, Christoph, Giuseppe Casalicchio and Bernd Bischl. 2020. Quantifying Model Complexity via Functional Decomposition for Better Post-hoc Interpretability. In P. Cellier and K. Driessens (eds.), Machine Learning and Knowledge Discovery in Databases, Communications in Computer and Information Science, 193–204. Cham: Springer International Publishing.
- Okoli, Chitu. 2024. "Statistical inference using machine learning and classical techniques based on accumulated local effects (ALE)." arXiv preprint arXiv:2310.09877.