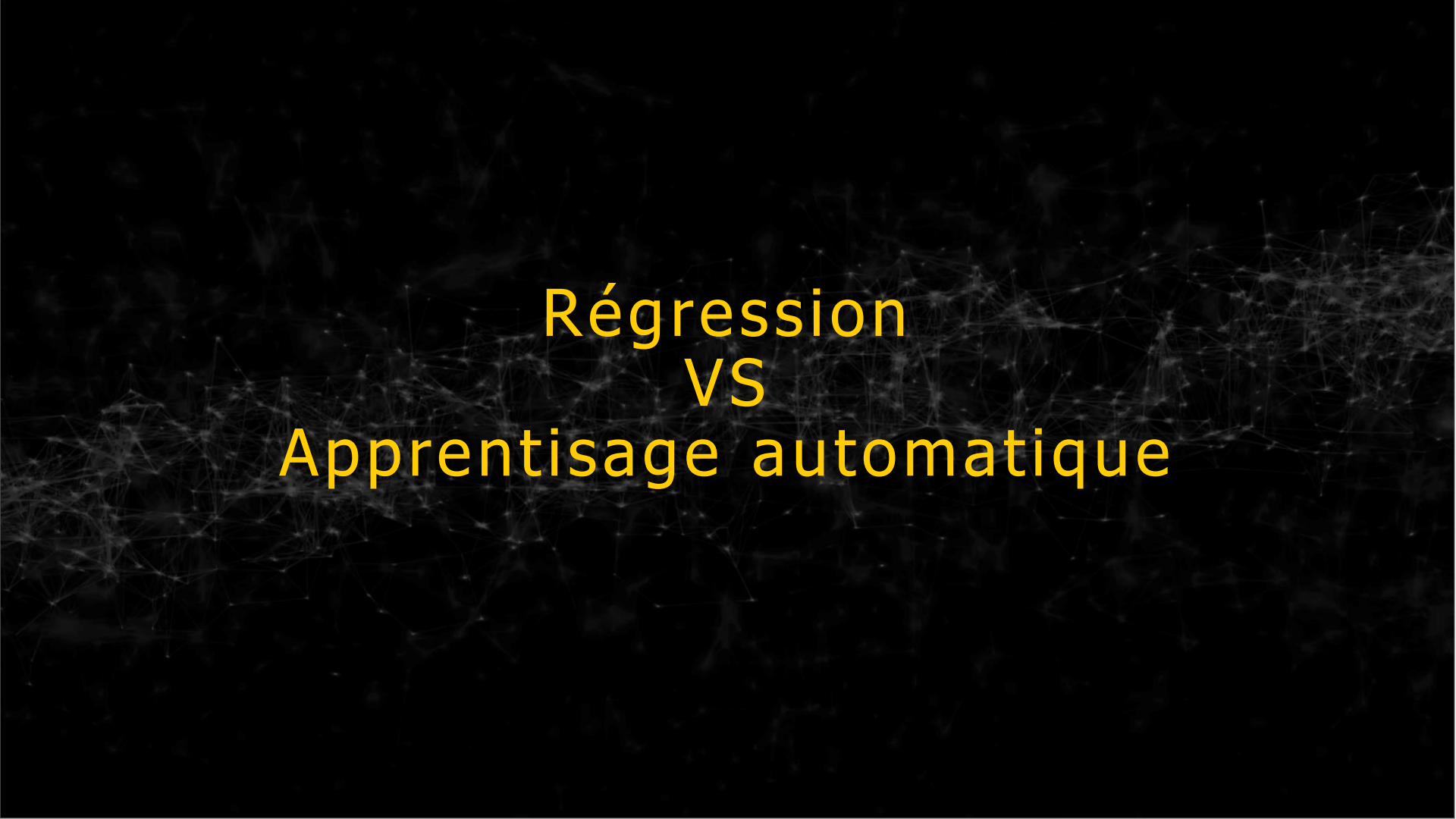


Machine learning VS Régression

Qui est le grand gagnant?

Introduction

- _ Comparaison entre régression et apprentissage automatique
- _ Mise en situation
- _ Vous donnez les clés pour savoir quel utiliser dans quel contexte



Régression VS Apprentissage automatique

L'approche

- _ On ne compare pas à une méthode d'apprentissage automatique particulière, on va plus garder à l'esprit un ensemble de méthode
- _ Le but est de vous amener à avoir l'esprit critique des deux

Différences à haut niveau

Régression

- _ Branche des mathématiques qui étudie les données
- _ Inférer la relation entre les variables

Tout est à propos de la relation entre les variables et leur significativité

Apprentissage automatique

- Sous ensemble de méthodes d'intelligence artificielle
- Prédire les résultats de manière précise
- **Tout est à propos des résultats**

Type de données

Régression

- _ Numérique
- _ Catégorique
- _ Une variable réponse et des variables explicatives

Apprentissage automatique

- Tout type de données: numérique, catégorique mais aussi plus large; image, son, vidéo...etc

Mise en œuvre

Régression

- _ Très rapide
- _ Moindre carré ou maximum de vraisemblance facile à mettre en place

Apprentissage automatique

- Peut être très long dépendamment de la quantité de données ou complexité du modèle

Ajout de variables

Régression

- Ca augmente le nombre de paramètres à estimer: difficulté dans la convergence des méthodes d'estimation
- Ajout de variable explicative inutiles dans l'explication de Y mène à de grande erreur standard dans les estimations des paramètres et donc du sur-ajustement

Apprentissage automatique

- Les modèles d'apprentissage automatiques permettent d'avoir un grand nombre de variables explicatives.

Ajout de variables

Régression

— Ajout d'interaction difficile en régression. Si on a p variables, alors on a $p*(p-1)/2$ interactions possibles

Apprentissage automatique

- Les interactions se font naturellement en apprentissage automatique: par exemple dans un arbre.

Variable confondantes

Régression

- Toutes les variables confondantes (reliées à la fois Y et à au moins une variable explicative dans X) doivent être incluses dans le modèle afin que les prévisions soient sans biais.

Apprentissage automatique

- En apprentissage automatique appliqué, la connaissance des variables confondantes amène des choix spécifique sur la préparation des données, le choix d'algorithme, initialisation de l'algorithme...etc

Sélection de variables

Régression

- Très utilité en régression pour diminuer le nombre de paramètres.
- Plusieurs méthodes sont possible: traditionnelle ou encore par pénalisation (ex: lasso)

Apprentissage automatique

- Se fait naturellement, par exemple dans une forêt aléatoire
- Le concept d'importance des variables fait une selection des variables naturelles

Multicolinéarité

Régression

- S'il y a de l'information redondante dans les variables explicatives cela cause problème pour l'estimation. On peut simplement enlever la ou les colonnes superflues.
- Si l'information est "presque" redondante, alors c'est plus difficile à détecter, les variances des estimations sont artificiellement gonflées.
- La régression régularisée de type Ridge avec un pénalité sur la norme L_2 permet de limiter les problèmes liés à la multicolinéarité.

Apprentissage automatique

- Ce n'est pas un problème en apprentissage automatique! En effet le but n'est pas de modéliser le lien entre les variables
- <https://towardsdatascience.com/why-multicollinearity-isnt-an-issue-in-machine-learning-5c9aa2f1a83a>

Hypothèses

Régression

- _ Linéarité
- _ Homoscédasticité
- _ Non corrélation
- _ Normalité

- _ Beaucoup d'hypothèses à vérifier

Apprentissage automatique

- Peu d'hypothèse

Linéarité

Régression

- La linéarité du modèle est une hypothèse primordiale en régression
- Parfois, ceci peut être vrai localement: pas besoin d'avoir la linéarité sur tout le domaine

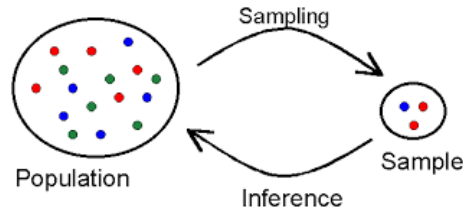
Apprentissage automatique

- Les méthodes d'apprentissage automatique ont été construites pour des données hautement non linéaire!

Inférence

Régression

- Principale force de la régression
- Test statistique sur les paramètres -> on peut inférer sur le mode de génération des données



Apprentissage automatique

- Ce n'est pas l'objectif

Simulation

Régression

- On estime la distribution des erreurs, on peut donc simuler de nouvelles observations
- Par contre pour s'assurer que les données simulées représente bien la réalité, on doit faire des hypothèses assez strictes (normalité, linéarité...etc)

Apprentissage automatique

- Ce n'est pas l'objectif

Erreur

Régression

- _ On **modélise** l'erreur
- _ On veut des erreurs assez petite mais surtout bien modéliser

Apprentissage automatique

- On **minimise** l'erreur
- On veut des erreurs le plus petit possible pour prédire au mieux

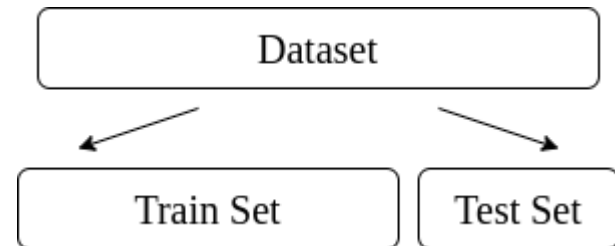
Prédiction

Régression

- On peut prédire des observations en dehors du jeu de données initial: mais ces prédictions peuvent être peu précises ou avec un intervalle de prédiction très large (multi colinéarité)

Apprentissage automatique

- Bâtir le meilleur modèle de prédiction: apprentissage sur les données





Les deux ont aussi des choses
en commun

Points communs

- _ Les deux exploitent des observations multivariés
- _ Les deux prédisent ou estiment les résultats basé sur les entrées
- _ En régression maximiser le vraisemblance est équivalent à minimiser l'entropie en apprentissage automatique.
- _ Les deux transforment les données en des informatiques tangibles



Qui est le grand gagnant?



Aucun
ou
Les deux!

Forces et faiblesses de chacun

- _ Basé sur l'analyse précédente, on va mettre en lumière les forces et faiblesse de chacun.
- _ Le but est de vous amener à être autonome dans le choix des deux méthodes dépendamment du contexte.

Forces de la régression

- **Interprétabilité** : Les modèles de régression linéaire sont relativement simples et faciles à interpréter, ce qui les rend utiles pour comprendre la relation entre les variables. Cela peut être très utile et même « rassurant » dans certains domaines.
- **Facilité de mise en œuvre** : Les modèles de régression linéaire sont faciles à mettre en œuvre, même avec des ensembles de données volumineux.
- **Vitesse de calcul** : Les modèles de régression linéaire sont généralement plus rapides à estimer que les modèles d'apprentissage automatique plus complexes, ce qui peut être un avantage pour les ensembles de données plus petits.

Faiblesses de la régression

- **Limitations de la linéarité** : Les modèles de régression linéaire ne sont efficaces que si la relation entre les variables est linéaire. Si la relation est complexe ou non linéaire, le modèle peut ne pas être approprié.
- **Sensibilité aux données aberrantes** : Les modèles de régression linéaire sont sensibles aux données aberrantes, qui peuvent considérablement affecter les résultats.
- **Sous-spécification** : Si le modèle de régression linéaire ne prend pas en compte toutes les variables importantes, il peut être sous-spécifié et les résultats peuvent être biaisés.

Forces de l'apprentissage automatique

- **Performance élevée** : Les modèles d'apprentissage automatique peuvent être très performants dans la résolution de problèmes complexes de classification, de prédiction, de reconnaissance d'image et de traitement de langage naturel.
- **Capacité à apprendre des motifs complexes** : Les modèles d'apprentissage automatique peuvent identifier des motifs complexes dans les données, même si la relation entre les variables est non linéaire ou difficile à décrire.
- **Adaptabilité** : Les modèles d'apprentissage automatique peuvent s'adapter à différents types de données et de problèmes, ce qui les rend flexibles et utiles dans une variété de domaines.

Faiblesses de l'apprentissage automatique

- **Complexité** : Les modèles d'apprentissage automatique peuvent être très complexes et nécessiter des ensembles de données massifs pour l'entraînement, ce qui peut rendre leur utilisation plus difficile et leur mise en œuvre plus coûteuse.
- **Difficulté d'interprétation** : Les modèles d'apprentissage automatique sont souvent difficiles à interpréter, ce qui peut rendre difficile de comprendre comment le modèle arrive à ses prédictions.
- **Surapprentissage** : Les modèles d'apprentissage automatique peuvent être sujets au surapprentissage, ce qui signifie qu'ils peuvent apprendre à mémoriser les données d'entraînement au lieu de généraliser les modèles pour les nouvelles données. Cela peut entraîner une perte de précision et de performance lors de la prédiction sur de nouvelles données.

Lequel utiliser?

- _ Chacun à ses forces et ses faiblesses, cela dépend de ce que l'on veut faire.
- _ Ca dépend du contexte, des données, de l'objectif....etc. Il faut se poser des questions au début du travail pour s'assurer d'utiliser le bon modèle.
- _ Justement faisons quelques mises en situation

Les questions à se poser

- _ Quelles type de données je possède?
- _ Quel sont les objectifs de l'étude?
- _ Quels genre de résultats dois-je donner?
- _ Qu'est ce qui as été fais à ce sujet?

Mise en situation

- _ On va voir quatre mise en situation de domaine très varié
- _ Pour chacun on va essayer de se poser les bonnes questions pour voir lequel est le plus approprié entre régression et apprentissage automatique
- _ Encore une fois il n'y a pas de bonne réponse, il faut simplement bien répondre aux objectifs.

Mise en situation – durée de conservation d'un vaccin

Biopharmaceutique

Lorsqu'une entreprise biopharmaceutique produit un nouveau vaccin, il doit soumettre aux autorités de santé du pays dans lequel il souhaite commercialisé le vaccin, une estimation de durée de conservation.

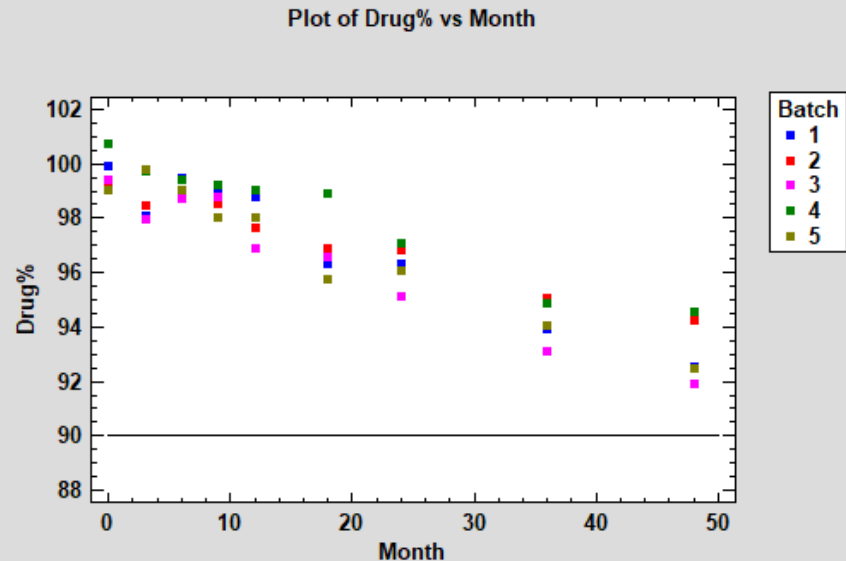
L'entreprise produit des lots de vaccin pour analyser dans le temps – étude de stabilité. A des temps bien précis, on analyse les vaccins pour s'assurer de leur efficacité.

Mise en situation – durée de conservation d'un vaccin

On suit dans le temps 5 lots (batch) de vaccins

On mesure dans le temps le % de concentré actif restant.

La limite basse du pourcentage de concentré actif pour que le vaccin soit encore efficace est de 90% (ligne noire)



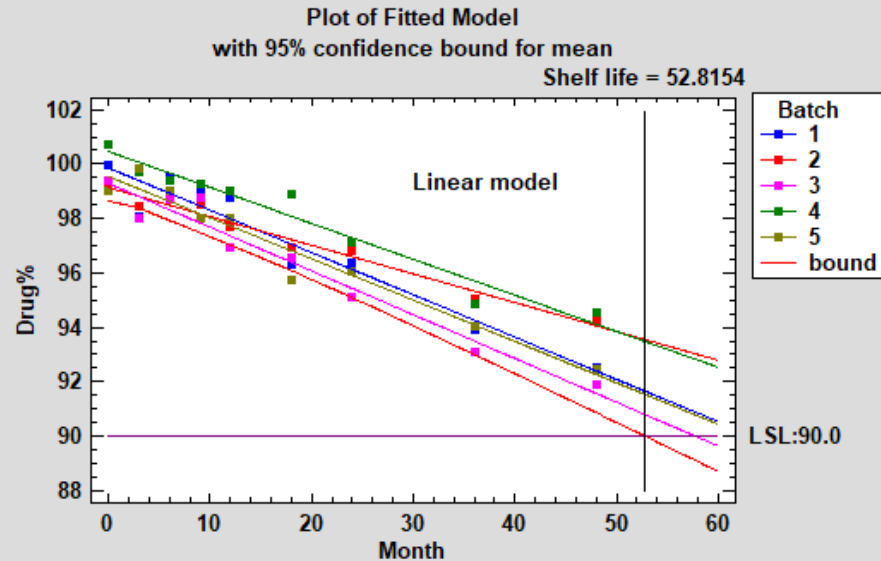
Mise en situation – durée de conservation d'un vaccin

- U.S. Department of Health and Human Services, Food and Drug Administration, (2003). *Guidance for Industry, Q1A(R2) Stability Testing of New Drug Substances and Products*.

An approach for analyzing the data on a quantitative attribute that is expected to change with time is to determine the time at which the 95 percent, one-sided confidence limit for the mean curve intersects the acceptance criterion. If analysis shows that the batch-to-batch variability is small, it is advantageous to combine the data into one overall estimate. This can be done by first applying appropriate statistical tests (e.g., p values for level of significance of rejection of more than 0.25) to the slopes of the regression lines and zero time intercepts for the individual batches. If it is inappropriate to combine data from several batches, the overall retest period should be based on the minimum time a batch can be expected to remain within acceptance criteria.

Mise en situation – durée de conservation d'un vaccin

On prédit une durée de conservation du vaccin soumis à 52.81 mois



Mise en situation – durée de conservation d'un vaccin

- Même s'il s'agit ici d'une prédiction que l'on veut faire: c'est-à-dire la durée de conservation d'un vaccin, on n'utilise pas de modèle d'apprentissage automatique:
 - Très peu de données pour faire un modèle d'apprentissage automatique
 - Les autorités (FDA) qui va réviser le rapport d'analyse des données donne dans sa guidances (Q1A) qu'il faut utiliser un modèle de régression linéaire pour prédire.
 - Dans le domaine pharmaceutique (vaccin) les modèles d'apprentissage automatique sont peut utiliser car sont considérés comme une boîte noire pour les autorités de santé (FDA ou Santé Canada)

Régression

Mise en situation – Estimation du cout de reconstruction

Assurances

Objectif: développer un modèle pour extraire les caractéristiques de la maison et évaluer les coûts de reconstruction uniquement en fonction d'images!

Utilise dans la tarification des polices d'assurances ou encore dans les inspections (fraudes)

Mise en situation – Estimation du cout de reconstruction

Obtenir une estimation du coût de reconstruction d'une maison très rapidement et à moindre coût.

Utiliser dans la priorisation des inspections des maisons



Mise en situation – Estimation du cout de reconstruction

- _ Données d'image hautement non linéaire et très complexe
- _ Pas de modèle linéaire en fonction d'image
- _ Prédiction précise car décision importante relié à cela
 - _ Priorisation inspection des maisons assurées

**Apprentissage
automatique**

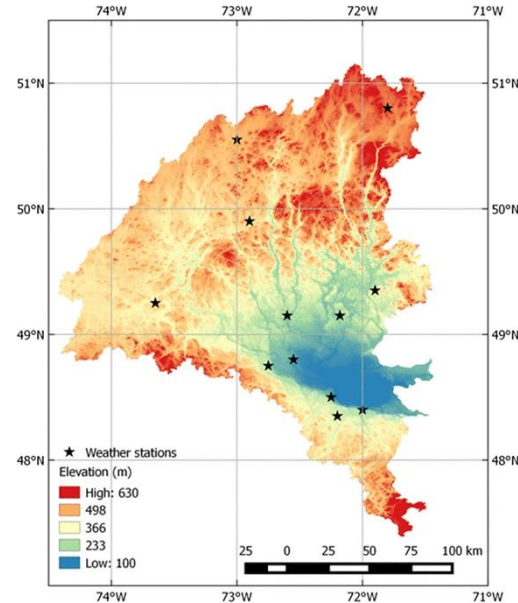
Mise en situation – prévision du débit pour l'optimisation des réservoirs

- La prévision de débit d'eau est utilisée pour comprendre les futurs débits entrants dans les réservoirs hydroélectriques.
- On souhaite modéliser et simuler des débits.
- Méthode testée sur un bassin (Lac St-Jean) versant canadien utilisé par Rio Tinto – division aluminium pour produire de l'hydroélectricité pour leurs usines de fusion d'aluminium.

Mise en situation – prévision du débit pour l'optimisation des réservoirs

Méthode

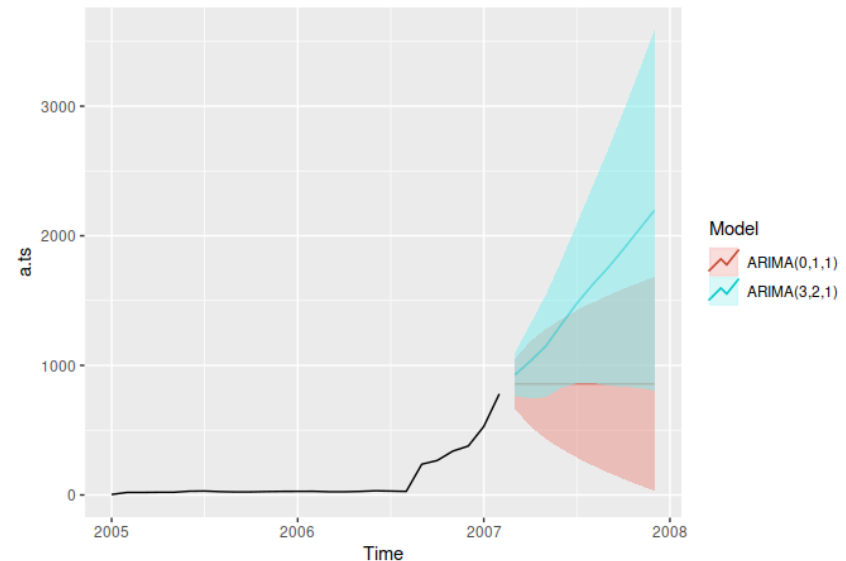
- Modélisation statistiques (semblable à la régression)
- L'erreur à été modélisé avec les données historiques
- On utiliser l'erreur pour simuler et prédire les débits



Régression

Mise en situation – séries temporelles

- Une série temporelle: une variable numérique mesuré dans le temps
- On peut modéliser une série temporelle avec un modèle de série chronologique (ARIMA, Autoregressive integrated moving average par ex.) qui modélise la série temporelle un peu comme un modèle de regression
- On souhaite prédire les prochaines observations (forecasting)



<https://otexts.com/fpp2/seasonal-arima.html>

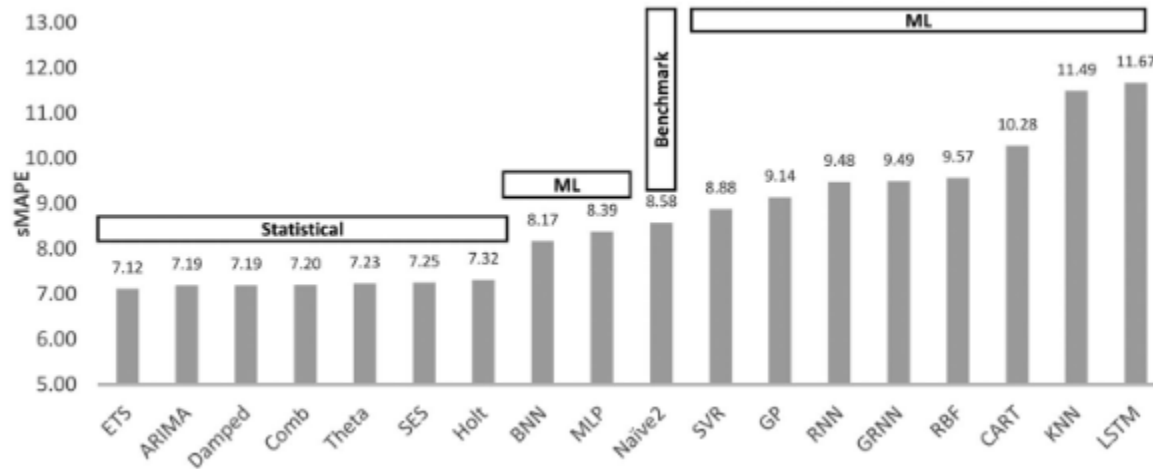
Mise en situation – séries temporelles

Les données

- Une série temporelle: une variable numérique mesuré dans le temps
- M3 – 3003 séries temporelles utilisé dans les comparaison de modèle statistique en série temporelles

Time interval between successive observations	TYPES OF TIME SERIES DATA						
	Micro	Industry	Macro	Finance	Demographic	Other	TOTAL
Yearly	146	102	83	58	245	11	645
Quarterly	204	83	336	76	57		756
Monthly	474	334	312	145	111	52	1428
Other	4			29		141	174
TOTAL	828	519	731	308	413	204	3003

Mise en situation – séries temporelles



Makridakis et al. *Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward* (August 2022)

Mise en situation – séries temporelles

- _ Les deux méthodes : séries temporelles et machine learning s'appliquent en séries temporelles
- _ L'objectif peut être de comparer les méthodes sur un jeu de données particulier

Makridakis et al. *Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward* (August 2022)

**Apprentissage
automatique**

Régression

The background of the slide is a solid black field. Overlaid on this is a complex, abstract network of thin, white lines. These lines connect numerous small, white dots or nodes, which are scattered across the entire frame. The network appears to be a representation of a data structure, a neural network, or a complex system. The lines and dots are more densely packed in some areas, particularly towards the right side of the image, and more sparse in others.

Conclusion

Unification des méthodes

- _ On a vu que les méthodes de régression sont différentes des méthodes d'apprentissage automatique
- _ L'utilisation de l'une ou de l'autre dépend de l'objectif de l'étude
- _ Les avancées en interprétabilité par exemple démontre un souhait de rassemblement.
- _ Régression et Apprentissage automatique comme méthode pour travailler avec des données