Time-penalised tree :
un algorithme temporel interprétable appliqué aux risques climatiques

Léonie Le Bastard – Finactys, Paris, France

Mathias VALLA – Chaire DIALog, Institut Louis Bachelier, Paris, France

José Garrido – Université Concordia, Montréal, Canada

Plan:

- **Introduction aux arbres de décision**
  - Fonctionnement (fonction de split)
  - Avantages/inconvénients
  - Extensions (bagging/boosting)
  - Utilisations en actuariat

- **Time-penalised trees**
  - Motivation
  - Données longitudinales
  - Fonctionnement
  - Exemple simple
  - Avantage/utilisations possibles

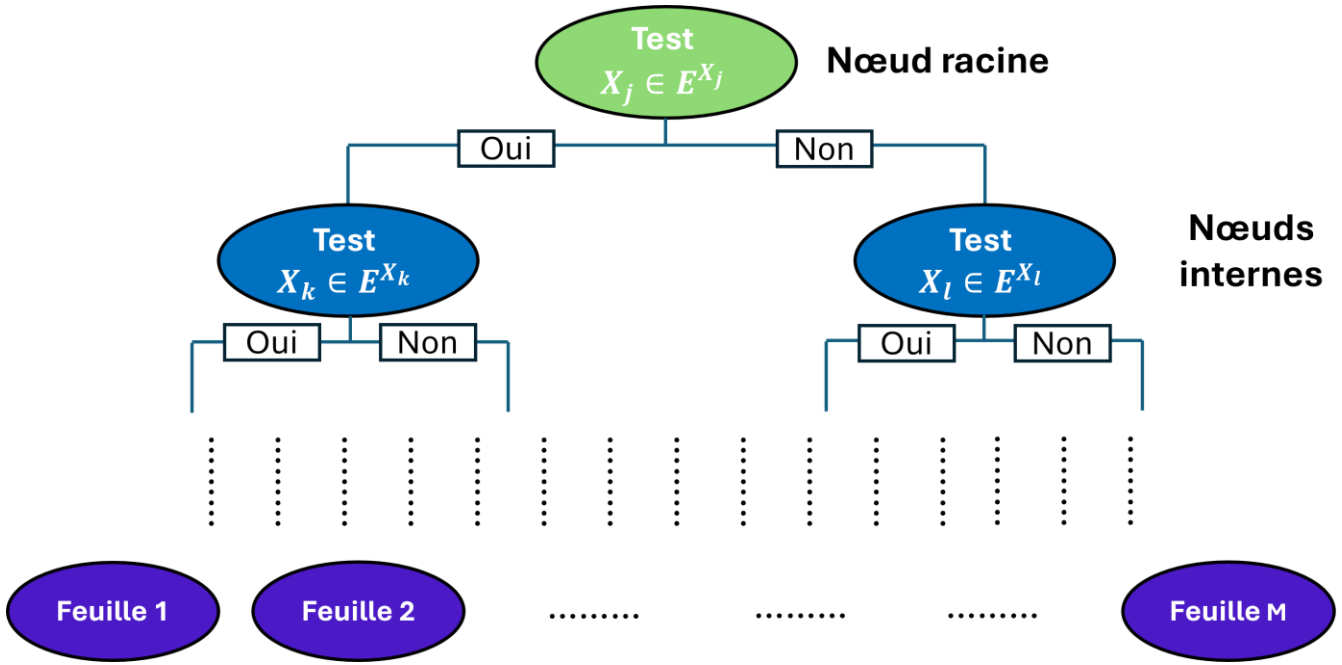
- **Application au risques climatiques**
  - Actuarial climate indices
  - Choix des variables et quantiles
  - Utilisation des TpT

- **Conclusion**

01 – Introduction aux arbres de décision

Principe

- Algorithme de classification
- **Objectif :** prédire une variable qualitative à partir de variables de tout type
- L'idée est de construire des classes « d'individus » les plus homogènes possibles



- Une fois l'arbre construit, classer un nouveau candidat se fait par une descente dans l'arbre, de la racine vers une des feuilles

Construction de l'arbre : choix des règles de split

Afin de construire l'arbre, l'algorithme doit décider quelle variable utiliser à chaque nœud afin de construire les classes **les plus homogènes possibles** dans ses nœuds fils

La règle choisie pour construire les nœuds de l'arbre s'appelle aussi la **fonction de split**

$$G(g_p; g_l, g_r) = I(g_p) - \left(\frac{N(g_l)}{N(g_p)} I(g_l) + \frac{N(g_r)}{N(g_p)} I(g_r) \right)$$

$N(g_l)$: nombre d'individus dans le nœud g_l
$I(.)$: fonction d'impureté

Il existe plusieurs critères permettant d'évaluer l'homogénéité/l'impureté d'un groupe :

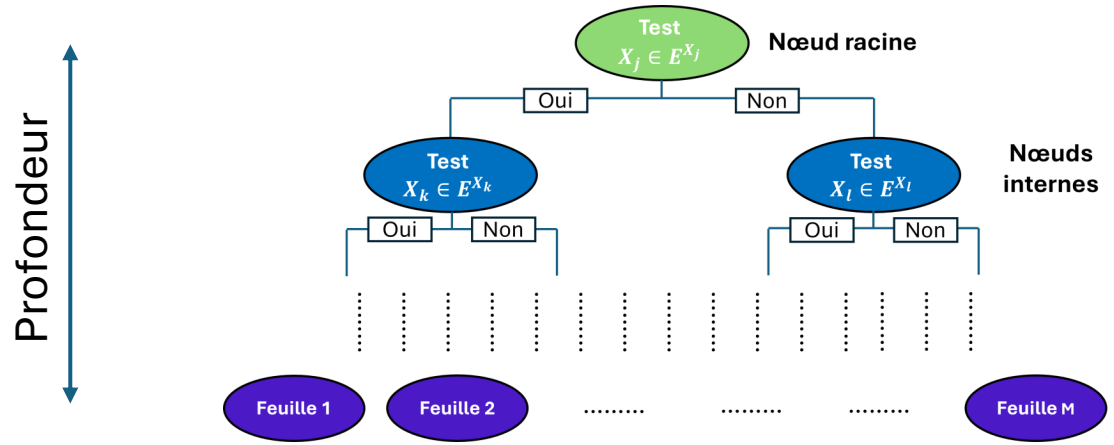
- Classification : indice de Gini ou l'entropie probabiliste
- Régression : Mean Square Error

Construction de l'arbre : critère d'arrêt

Peu d'intérêt de n'avoir qu'un individu par feuille : mise en place d'un critère d'arrêt nécessaire (surajustement)

Il existe pour cela plusieurs règles. Quelques exemples :

- Nombre minimal d'individus par feuille
- Profondeur maximale de l'arbre
- Ne pas subdiviser un nœud si la subdivision ne permet pas d'améliorer les performances de prédiction de l'algorithme



Avantages et inconvénients

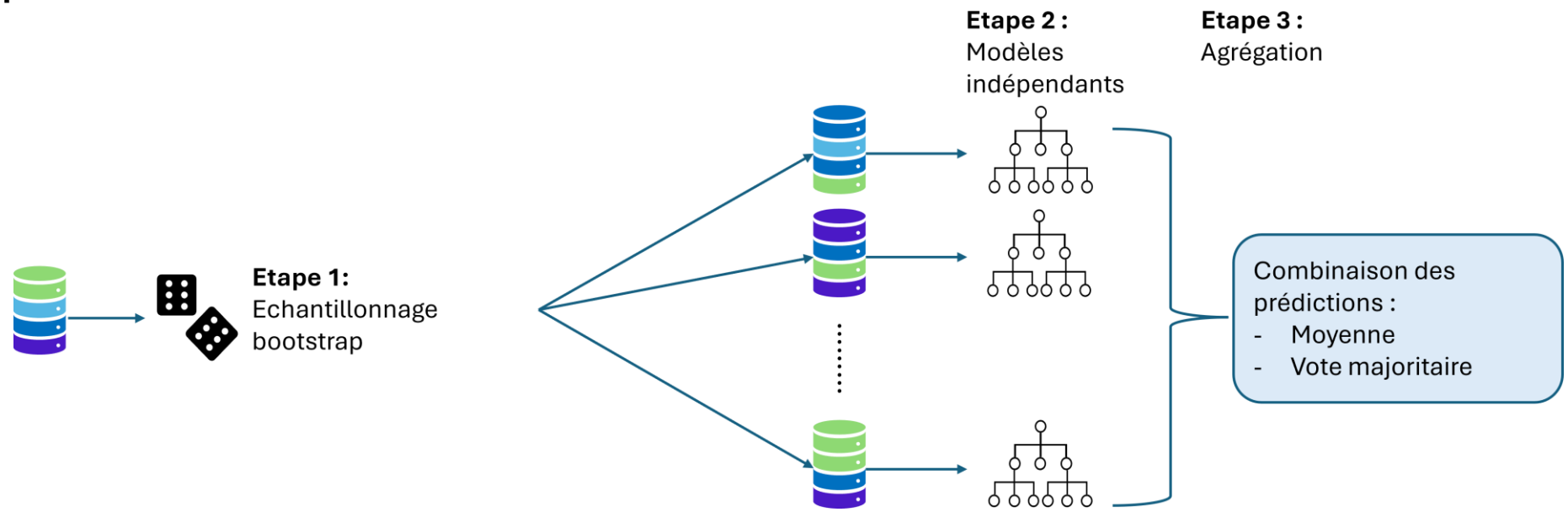
- Très visuel
- Règles de décision facilement compréhensibles
- Interprétable
- Possibilité de prendre en compte des interactions

- Risque de surapprentissage : Règle non généralisable à de nouvelles données.
- Algorithme assez instable, pouvant être sensible à des fluctuations de l'échantillon d'apprentissage
- Très simpliste : séparation par plan

Extensions

Bagging

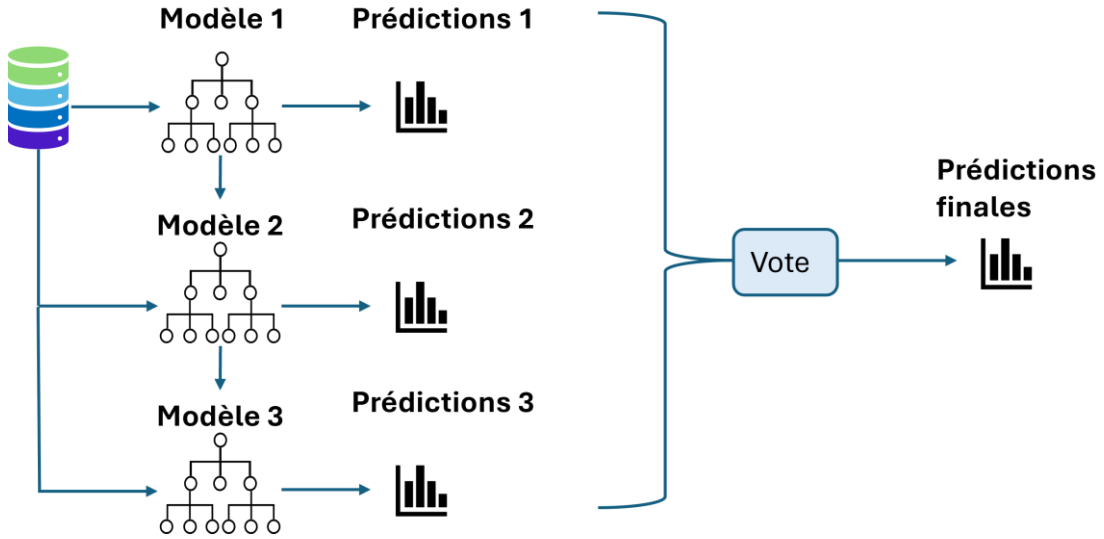
- Fusion de bootstrap et aggregating
- Technique qui améliore la précision et stabilité des modèles (réduction de la variance)
- **Étapes :**



Extensions

Boosting

- Technique d'ensemble qui combine plusieurs modèles faibles (ex. arbres de décision peu profonds) pour créer un modèle performant
- Enchaînement séquentiel : Chaque modèle est entraîné pour corriger les erreurs de ses prédécesseurs. **Contrairement au bagging, les modèles ne sont pas indépendants**



02 – Time-penalised trees

Limites

- Ne permet notamment pas de prendre en compte des données longitudinales

- Données historiques:
  - Deux observations similaires sont regroupées, même si elles sont historiquement très différentes
  - Comment différencier les observations temporellement ?

Questions de recherche

- Comment modifier les algorithmes d'arbres de décision pour qu'ils s'adaptent aux données variant dans le temps ?

- Quelles utilisations en actuariat ? Quels résultats ?

Données longitudinales

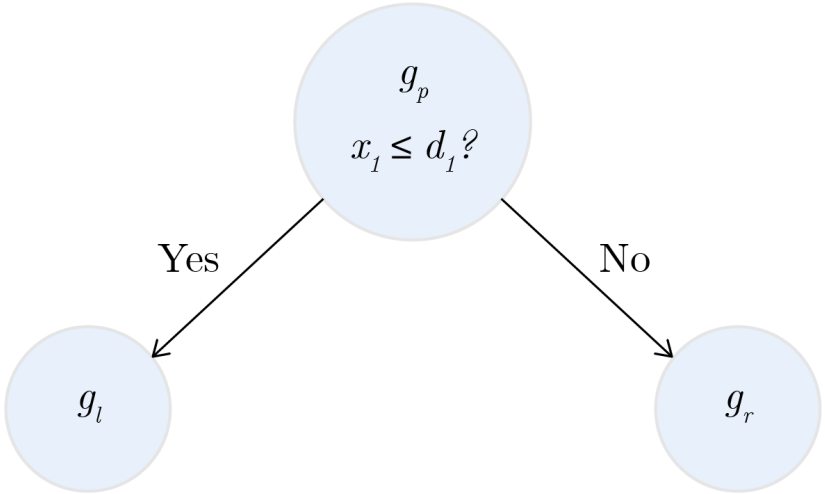
- Exemple:

ID	EVENT	START	END	PRODUCT	SEX	SENIORITY	F_i	CLAIM	CNTRCTS	AGE	YEAR
46784	0	0	1	3	2	0	8,38	0	1	66	2013
46784	0	1	2	3	2	1	8,40	0	1	67	2014
46784	0	2	3	3	2	2	8,57	0	1	68	2015
46784	0	3	4	3	2	3	11,90	0	1	69	2016
46784	0	4	5	3	2	4	12,10	0	1	70	2017
46784	0	5	6	3	2	5	12,28	0	1	71	2018
46784	1	6	7	3	2	7	15,06	-15,06	1	72	2019
7825	0	0	1	2	2	0	3,02	0	1	81	2016
7825	0	1	2	2	2	1	3,05	0	1	82	2017
7825	0	2	3	2	2	2	3,10	0	1	83	2018
7825	0	3	5	2	2	5	3,15	0	1	84	2019
264309	0	0	1	3	2	0	2,61	0	1	66	2016
264309	0	1	2	3	2	1	2,64	0	1	67	2017
264309	0	2	3	3	2	2	2,67	0	1	68	2018
264309	0	3	5	3	2	5	3,48	0	1	69	2019

- En actuariat:
  Historique de flux en assurance vie, données télématiques en auto, séries temporelles en finance, données climatiques en CATNAT

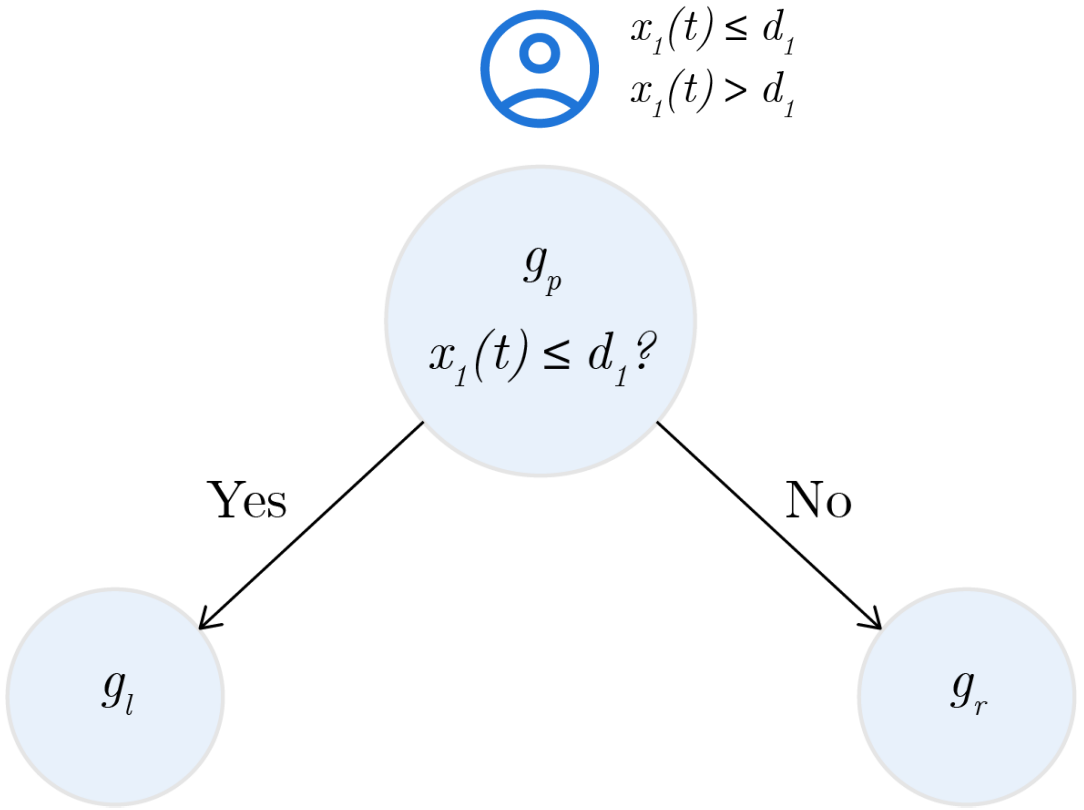
Rappels

Zoom sur la division d'un noeud



$$G(g_p; g_l, g_r) = I(g_p) - \left(\frac{\mathcal{N}(g_l)}{\mathcal{N}(g_p)} I(g_l) + \frac{\mathcal{N}(g_r)}{\mathcal{N}(g_p)} I(g_r) \right)$$

Problème:



Etat de l'art:

The multivariate methods

- Cart extension - Segal (1992) De ath' (2002) then extensions
- Designed for a longitudinal response but time-fixed covariates

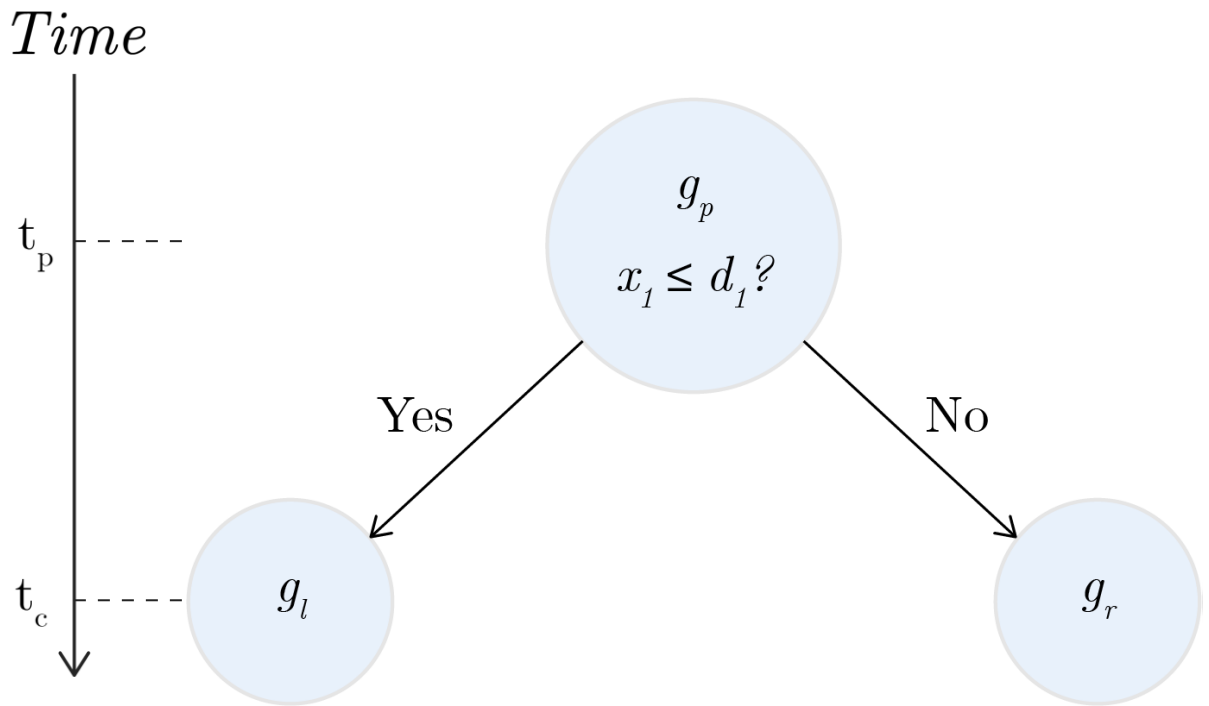
The « eventually not longitudinal » methods

- Unmodified models
- LongCART - Kundu (2019)
- Lagged responses - Ritschard and Oris (2005), Moradian (2021)
- Workaround the longitudinal nature of the data

The « state-of-the-art » methods

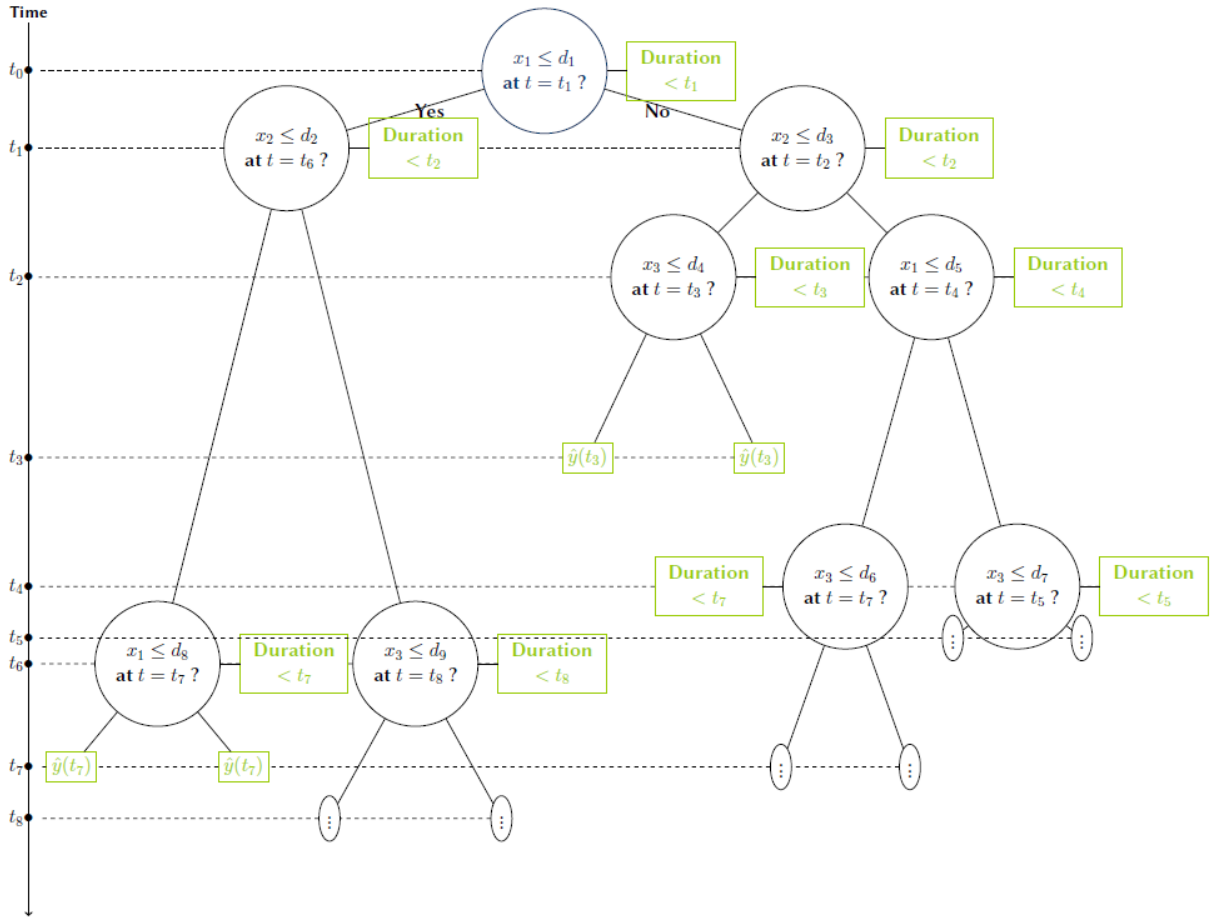
- Left-truncated and right censored trees and forests - Yao et al. (2020)
- Mixed-effect tree-based models - Hajjem et al. (2011 and 2014), Sela et al. (2012), Fu et al. (2015), Capitaine et al. (2021)
- Divide observations of a same subject into several pseudo subject observations.
- Spread a subject's observations across the tree leaves

Zoom sur un noeud:

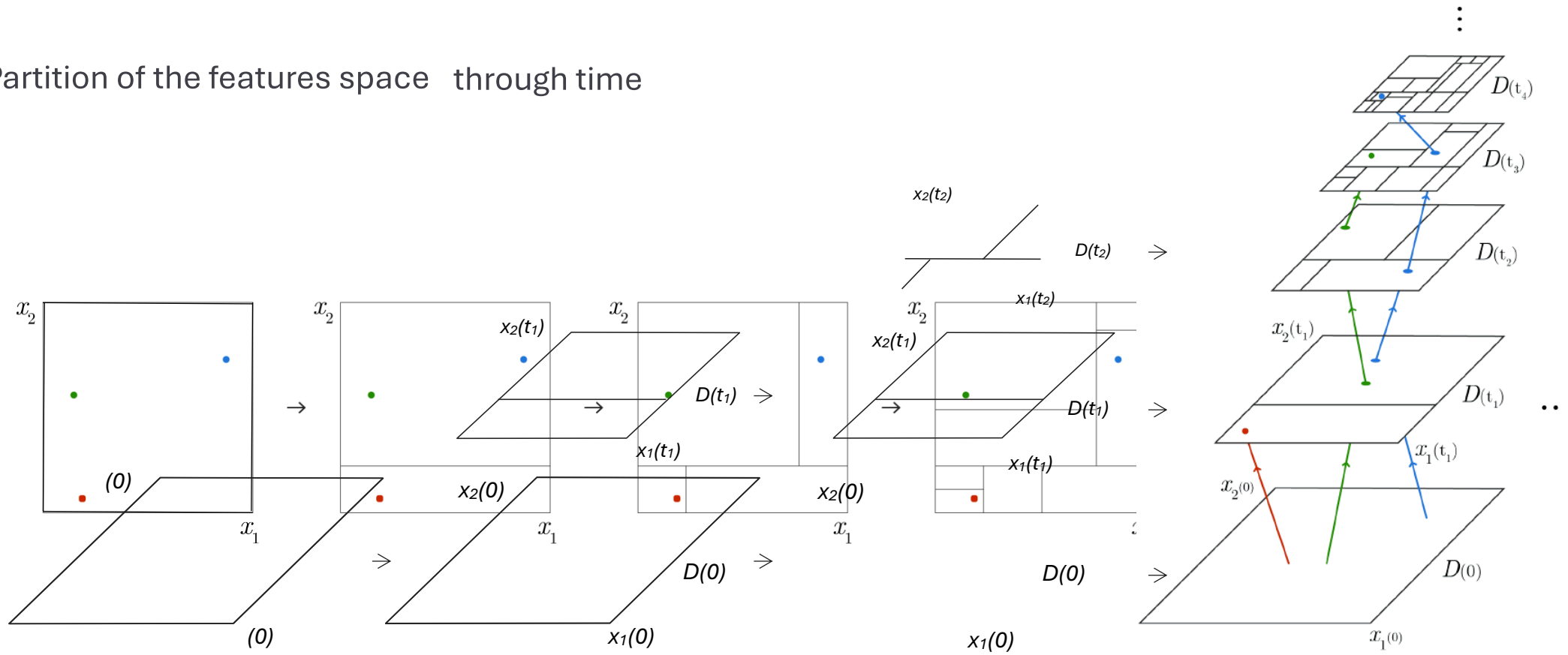


$$G_\gamma(g_p; g_l, g_r, g_t) = \left[I(g_p) - \left(\frac{\mathcal{N}(g_l)}{\mathcal{N}(g_p)} I(g_l) + \frac{\mathcal{N}(g_r)}{\mathcal{N}(g_p)} I(g_r) + \frac{\mathcal{N}(g_t)}{\mathcal{N}(g_p)} I(g_t) \right) \right] \cdot e^{-\gamma \cdot (t_c - t_p)}$$

Time-penalised tree:



Partition of the features space through time

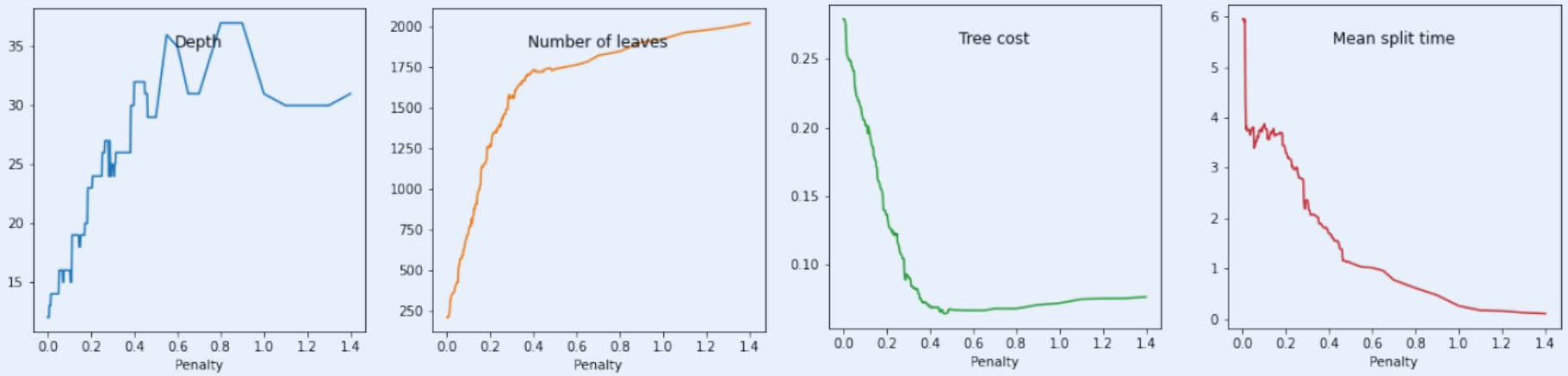


Proof of concept: rachats en assurance vie

- Exemple:

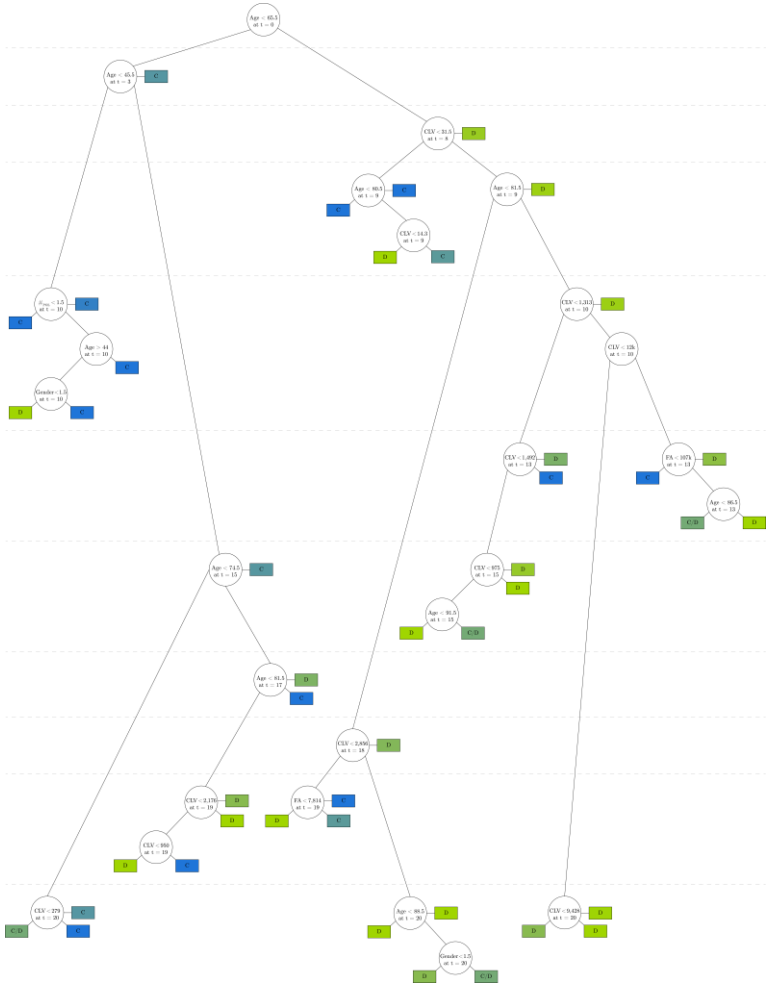
ID	EVENT	START	END	PRODUCT	SEX	SENIORITY	F_i	CLAIM	CNTRCTS	AGE	YEAR
46784	0	0	1	3	2	0	8,38	0	1	66	2013
46784	0	1	2	3	2	1	8,40	0	1	67	2014
46784	0	2	3	3	2	2	8,57	0	1	68	2015
46784	0	3	4	3	2	3	11,90	0	1	69	2016
46784	0	4	5	3	2	4	12,10	0	1	70	2017
46784	0	5	6	3	2	5	12,28	0	1	71	2018
46784	1	6	7	3	2	7	15,06	-15,06	1	72	2019
7825	0	0	1	2	2	0	3,02	0	1	81	2016
7825	0	1	2	2	2	1	3,05	0	1	82	2017
7825	0	2	3	2	2	2	3,10	0	1	83	2018
7825	0	3	5	2	2	5	3,15	0	1	84	2019
264309	0	0	1	3	2	0	2,61	0	1	66	2016
264309	0	1	2	3	2	1	2,64	0	1	67	2017
264309	0	2	3	3	2	2	2,67	0	1	68	2018
264309	0	3	5	3	2	5	3,48	0	1	69	2019

Pénalité optimale

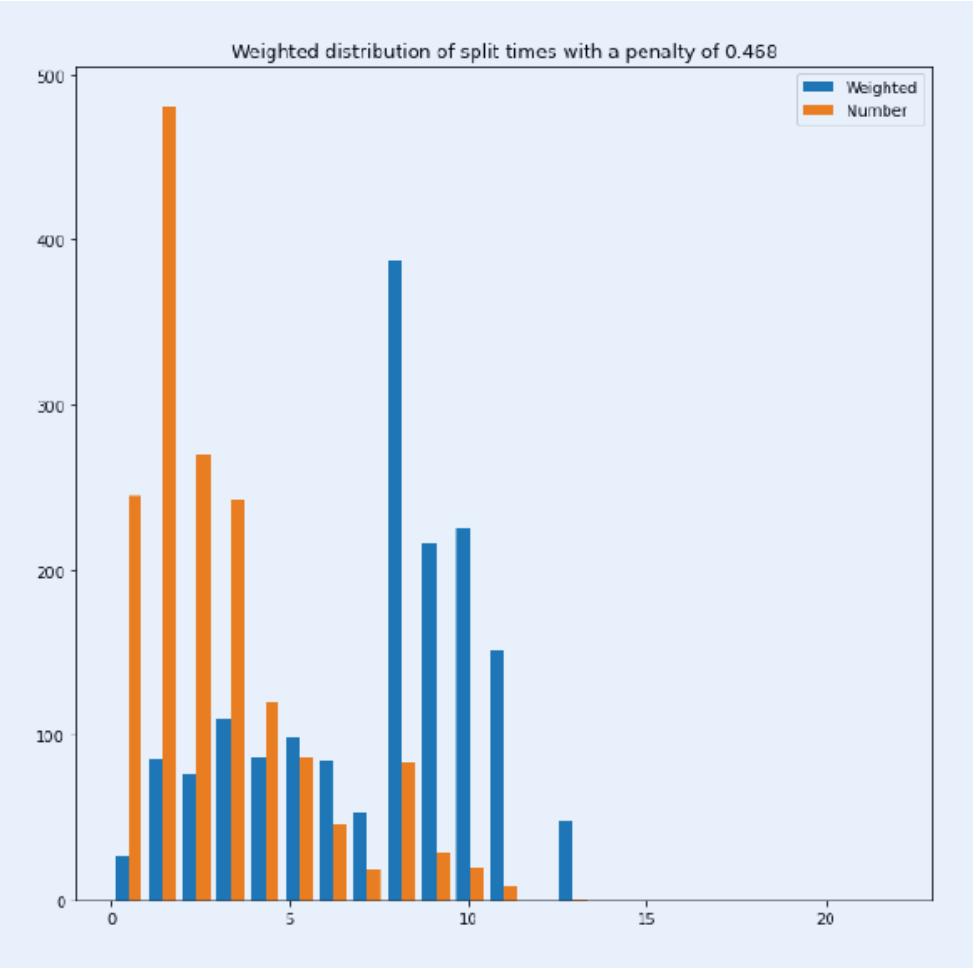


TpT optimal

$Time$



Distribution des splits temporels



Conclusion intermédiaire

- **Segmentation temporelle pertinente** : TpT identifie des périodes critiques, comme l'année 8, où les assurés peuvent résilier sans pénalité.
- **Covariables discriminantes** : Les variables telles que l'âge à la souscription, la valeur à vie du client (CLV), et le montant assuré se révèlent significatives dans la différenciation des trajectoires.
- **Amélioration de l'interprétabilité** : Contrairement aux méthodes longitudinales existantes, chaque assuré conserve une trajectoire unique dans l'arbre, rendant les résultats directement exploitables pour des analyses stratégiques.
- **Visualisation** : Les trajectoires individuelles sont visualisées dans un espace longitudinal, facilitant l'identification des groupes homogènes et des comportements à risque.

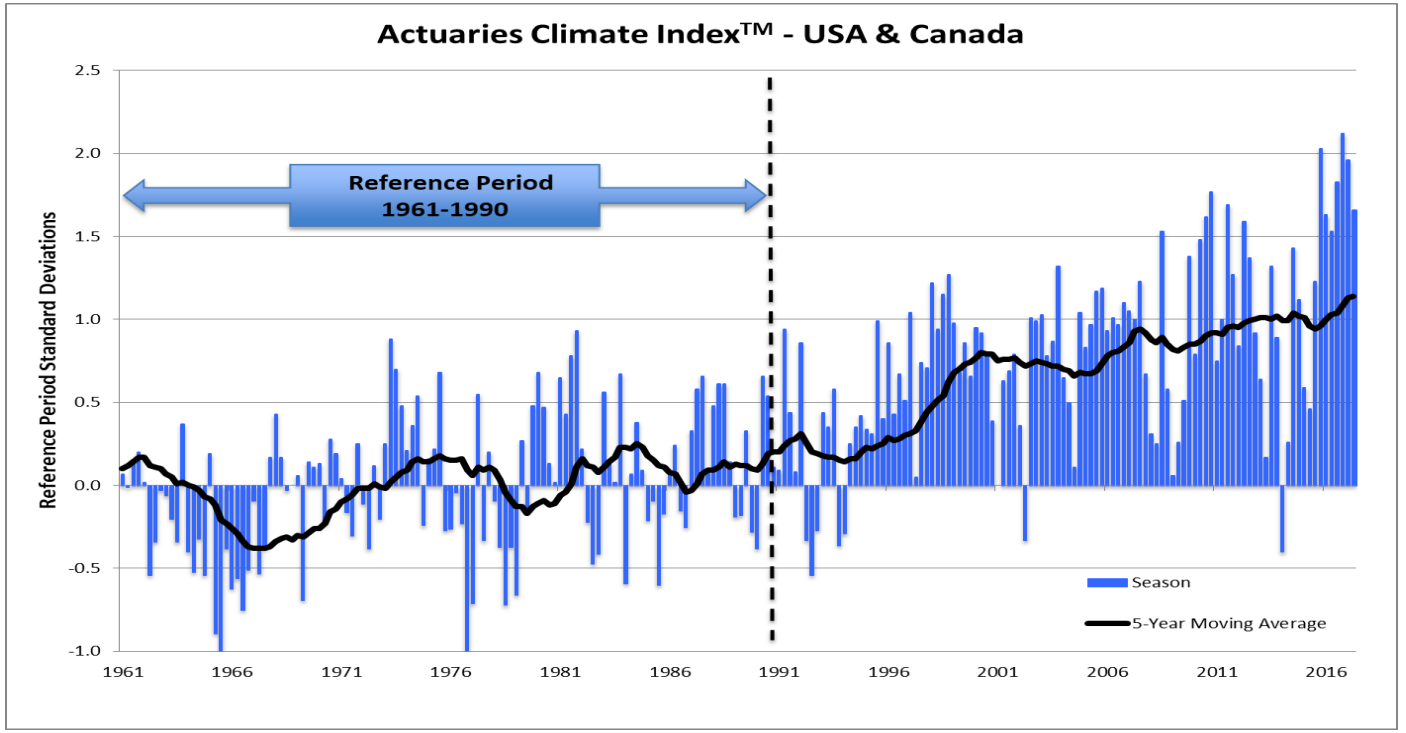
03 – Application au risque climatique

Indices climatiques actuariels

- L'Actuaries Climate Index™ (ACI) est un indice des risques climatiques extrêmes. Un peu comme l'Indice des Prix à la Consommation, qui suit les variations moyennes des prix dans le temps pour un panier de biens et services standard.
- Les actuaires mesurent et gèrent de nombreux types de risques : l'ACI mesure les risques climatiques sur la base d'un ensemble d'événements climatiques extrêmes et de changements du niveau de la mer.
- Une augmentation des valeurs de l'indice indique une hausse de la fréquence des événements climatiques extrêmes.
- Le climat est défini sur de longues périodes, de façon a distinguer les fluctuations à court et à long terme. Cependant, ces dernières années, les fluctuations a court terme ont été beaucoup plus souvent à la hausse qu'à la baisse, de façon anormale.

Actuarial climate indices

- La méthodologie de l'ACI utilise une période de référence de 30 ans, de 1961 à 1990. La moyenne est calibrée à zéro sur cette période.
  Les valeurs mensuelles ou saisonnières (barres de couleur) donnent une mesure quantitative des anomalies pour chacune des six composants de l'ACI.



Variables et quantiles

- 6 composantes:
  - Fréquence des températures extrêmement élevées (90ème percentile)
  - Fréquence des températures extrêmement basses (10ème percentile)
  - Fréquence des vents forts
  - Quantité maximale de précipitations importantes
  - Plus longue période de jours consécutifs sans précipitations (< 1 mm) au cours des 12 derniers mois
  - Changement du niveau de la mer

Le terme « extrême » se réfère aux valeurs en dessous ou au-dessus des quantiles à 10 et 90%.

Les valeurs sont mesurées comme des anomalies, ce qui, dans le contexte de l'ACI, correspond à la différence entre une valeur donnée pour un mois ou une saison et la valeur de cette même période au cours de la période de référence 1961-1990. Ces différences sont standardisées par l'écart-type sur la période de référence.

Calcul d'une composante

Le calcul se déroule en deux étapes :

Étape 1 : Calculer la valeur moyenne pour un mois ou une saison, puis soustraire la moyenne correspondante sur toute la période de référence de 30 ans (1961-1990) pour mesurer la variance.

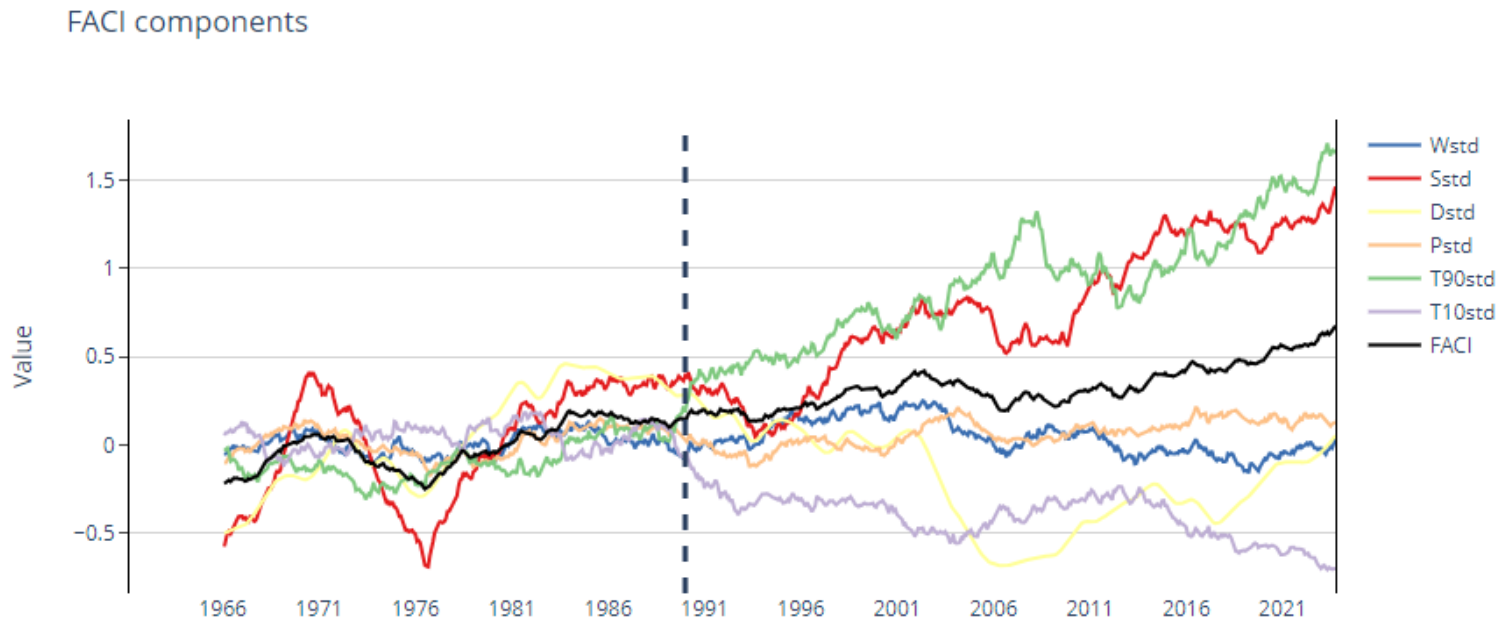
1. Les composantes (fréquence, jours, millimètres de pluie, etc.) étant exprimées dans des unités différentes, elles ne peuvent pas être additionnées directement.
2. Les écarts-types sont utilisés comme unité de standardisation.

Étape 2 : Diviser la différence obtenue par cet écart-type pour convertir les valeurs des composantes en valeurs standardisées.

1. Ces valeurs, appelées *anomalies standardisées*, permettent d'exprimer les composantes dans une unité commune (écart-type).
2. Formule : $(x-\mu_{ref})/\sigma_{ref}$, où x est la valeur mensuelle/saisonnière, μ_{ref} la moyenne de la période de référence et σ_{ref} l'écart-type correspondant.

Calcul de l'ACI - France

$$ACI = \frac{1}{6} \cdot (T90_{std} - T10_{std} + P_{std} + D_{std} + W_{std} + S_{std}),$$



Utilisation des TpT

- **But:** repenser le choix des quantiles de température retenues pour essayer d'expliquer le risque de surmortalité lié aux températures. Un sujet = un département de France métropolitaine.

Étape 1 - Génération d'un TpT entraîné:
  - avec les 6 composantes de l'ACI en variables d'entrée (dont T10 et T90)
  - les anomalies de mortalité mensuelles, par département français en variable cible

Étape 2 - Génération d'un TpT entraîné:
  - en ajoutant plus de quantiles de températures en variables d'entrée (T10, T20, T30, … , T90, T95)

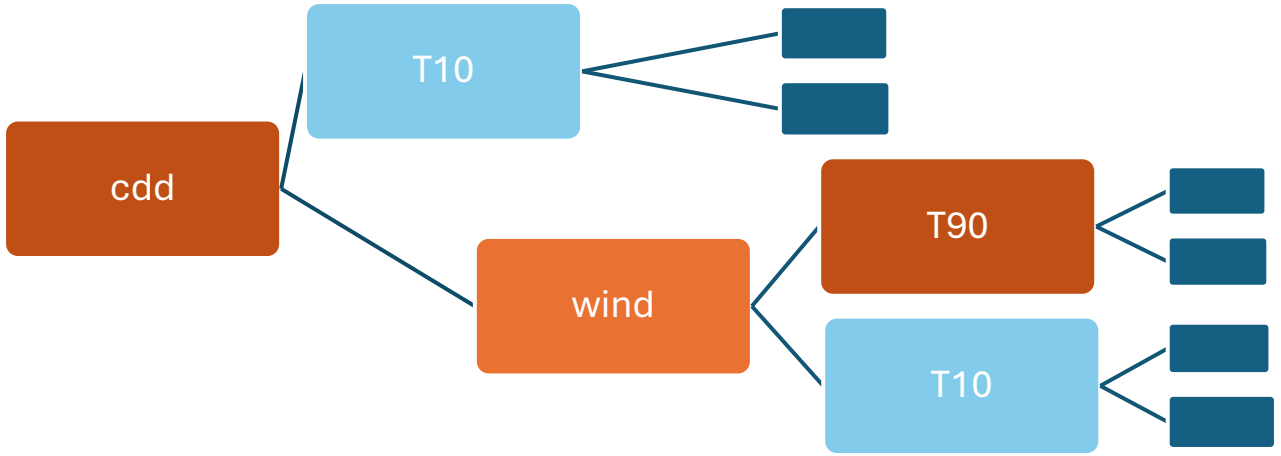
Étape 3 - Génération d'un TpT entraîné:
  - uniquement avec des quantiles plus extrêmes que la méthodologie originale (T95)

- **Question**: Est-ce que d'autres quantiles de température que T10/T90 pourraient grouper les départements à risque spatio-temporel homogène ?

Utilisation des TpT

Étape 1 - Génération d'un TpT entraîné:

- avec les 6 composantes de l'ACI en variables d'entrée (dont T10 et T90)
- les anomalies de mortalité mensuelles, par département français en variable cible

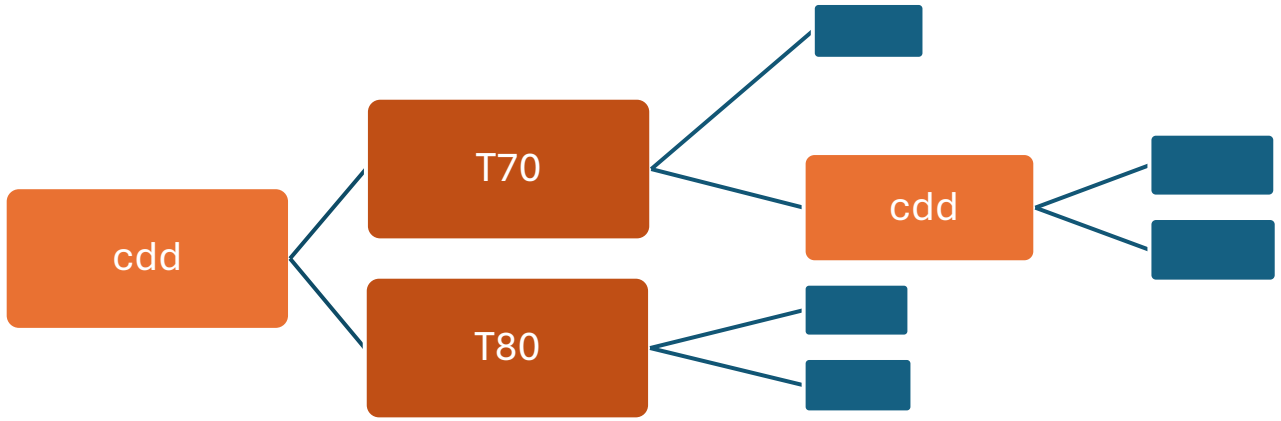


Variables d'importance: *cdd, T10, vent, T90*

Utilisation des TpT

Étape 2 - Génération d'un TpT entraîné:

- en ajoutant plus de quantiles de températures en variables d'entrée (T10, T20, T30, ..., T90, T95)

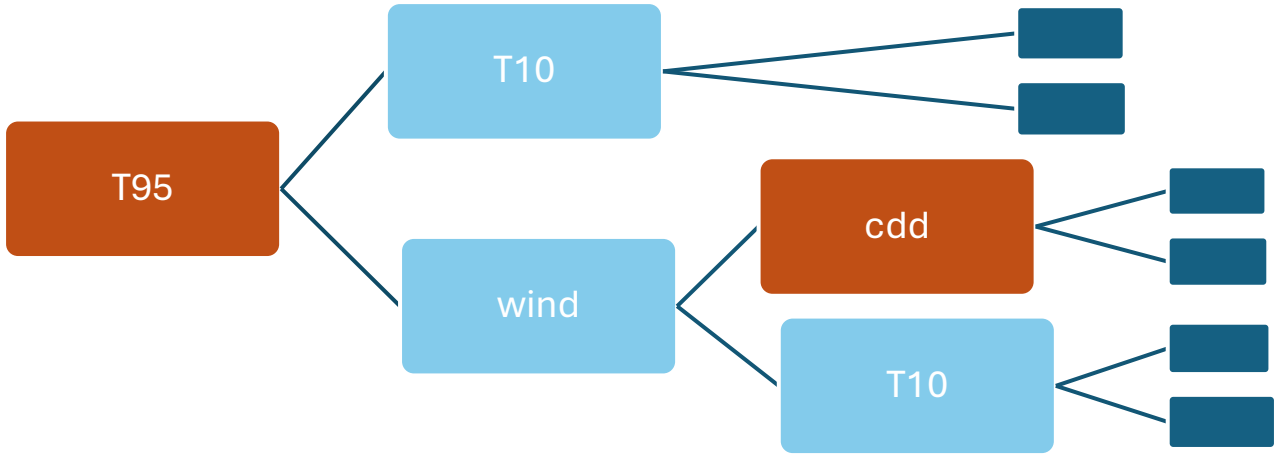


Variables d'importance: *sécheresse, T80, T70*
Gain d'impureté total: *+2% par rapport à la méthodologie classique*

Utilisation des TpT

Étape 3 - Génération d'un TpT entraîné:

- uniquement avec des quantiles plus extrêmes que la méthodologie originale (T95)



Variables d'importance: *T95, T10, vent*
Gain d'impureté total: *+5% par rapport à la méthodologie classique*

04 – Conclusions

Conclusions

Les modèles par arbres en actuariat

•Outils précieux pour la segmentation et l'interprétation des risques.
•Limites actuelles face aux données longitudinales, omniprésentes dans les applications actuarielles.

Time-penalised Tree (TpT) : une avancée innovante

•Conception adaptée à la prise en compte des variables temporelles dans des environnements dynamiques.
•Permet une partition conjointe de l'espace des covariables et du temps, améliorant ainsi la prévision et l'interprétabilité.

Conclusions

Propriétés théoriques prometteuses
•Efficacité démontrée dans des contextes complexes.
•Potentiel d'applications étendu aux autres domaines de actuariat et au-delà.

Limites actuelles
•Validation sur un ensemble de données limité.
•Comparaisons approfondies avec des techniques établies (bagging, boosting) encore en cours.

Perspectives de développement
•Création d'un package open-source pour une adoption plus large.
•Adaptation à des techniques avancées comme le bagging et le boosting.
•Exploration de nouvelles applications et domaines d'amélioration.

Bibliographie

TpT
•Valla, M. Time-penalised trees (TpT): introducing a new tree-based data mining algorithm for time-varying covariates. Ann Math Artif Intell (2024). https://doi.org/10.1007/s10472-024-09950-w

French ACI
•José Garrido, Xavier Milhaud, Anani Olympio. The definition of a French actuarial climate index; one more step towards a European index. 2023. ⟨hal-04491982⟩

•**Livre vert Chaire DIAlog**
Garrido, Jose & Milhaud, Xavier & Olympio, Anani & Popp, Max. (2024). Climate Risk and its Impact on Insurance. Link

05 – Appendix

Exemple d'utilisation des arbres en actuariat

- Tarification des primes d'assurance :
  - Assurance automobile : Identifier les facteurs influençant les risques (ex. âge, type de véhicule, etc.) pour ajuster les primes
  - Assurance santé : Segmentation des assurés en fonction de leur historique médical pour proposer des tarifs adaptés

- Analyse de la sinistralité :
  - Détecter les caractéristiques des clients associés à une probabilité plus élevée de sinistres, pour ajuster les politiques de souscription