Témoignage d'un doctorant en première année

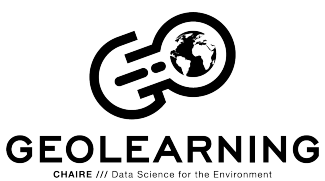
Journée Geolearning

Antoine Doizé

Advisors : D. Allard, P. Naveau, O. Wintenberger

Sorbonne-Université - Fondation des Mines de Paris

November 17, 2023



Presentation du profil

Une formation entre ingénierie et recherche 1/2

Ecole polytechnique : Département Data Science et IA (promo X17)

1. Stage Nephelaï : Algorithmes de détection d'anomalies dans des transactions financières (Classification, Modèles statistiques, Séries temporelles).
2. Stage Aqemia : Algorithmes génératifs de molécules à potentiel thérapeutique (Modèles génératifs et prédictifs, optimisation en grande dimension).

ENS Paris-Saclay : Master Mathématiques Vision Apprentissage

1. Master de recherche en IA : Deep Learning, Computational Statistics, Object Recognition, Convex Optimisation, Time Series, Bayesian Machine Learning, Kernel Methods...
2. Stage BCG Gamma : Modélisation bayésienne de l'impact des campagnes marketing. (Regression linéaire bayésienne, Chaînes de Markov, Optimisation).

Une formation entre ingénierie et recherche 2/2

Sorbonne-Université : Master Probabilité Modèles Aléatoires

1. Master de recherche en probabilités : Calcul Stochastique, Théorèmes limites, Processus de Markov, Statistiques de l'apprentissage, Géométrie Aléatoire, Bayesien Non-Paramétrique...

Stages de recherche

1. Stage de recherche Sorbonne-Université : Extreme Value Theory. Etude des propriétés du bootstrap pour les estimateurs de shape de queue de distribution.
2. Stage Inria : Extreme Value Theory pour l'étude des pluies. Modélisation du comportement des régimes extrêmes de pluie (sécheresse et pluies intenses) mesurés en une station météo.

Modèles statistiques spatiotemporels pour simuler les périodes de sécheresses et de pluies intenses sur des périodes longues et à l'échelle régionale (1/2)

Notions et difficultés clés de l'étude

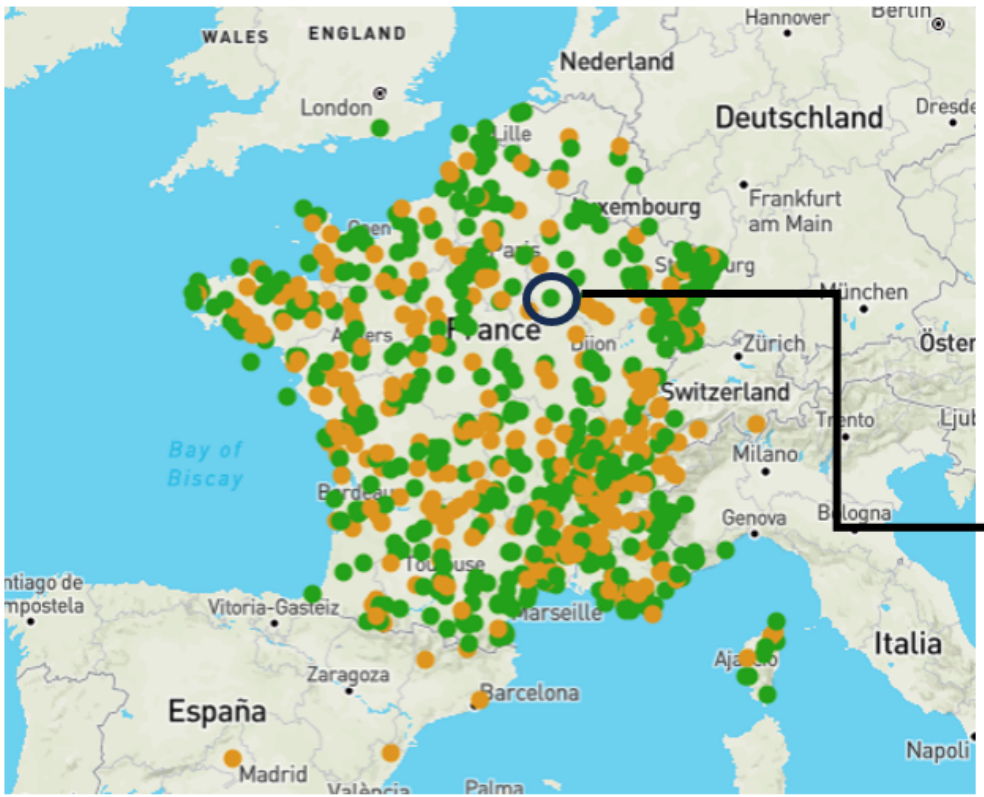
1. Etude de séries temporelles longues et sparse : il ne "pleut pas" beaucoup plus souvent qu'il ne pleut.
2. Comportements extrêmes des précipitations : On cherche particulièrement à bien modéliser les régimes extrêmes (inondations, longues sécheresses) : alors qu'on en observe peu (ils sont par définition rares).
3. Système non Markovien : il y a des phénomènes de persistance dans les sécheresses.
4. Système non stationnaire : le changement climatique impacte le comportement des précipitations (et en particulier des extrêmes).

Modèles statistiques spatiotemporels pour simuler les périodes de sécheresses et de pluies intenses sur des périodes longues et à l'échelle régionale (2/2)

Enjeux de l'étude

1. Enjeu risque climatique : mieux comprendre l'évolution des régimes des catastrophes naturelles (pluies intenses / longues sécheresses)
2. Enjeu d'aménagement du territoire :
   - Construction et phénomène de rétractation des argiles
   - Choix des cultures à implanter / des essences d'arbres les plus adaptées à un climat

Données à disposition



- précipitations journalières
- durée de plusieurs années
- mesurées sur plusieurs stations météo

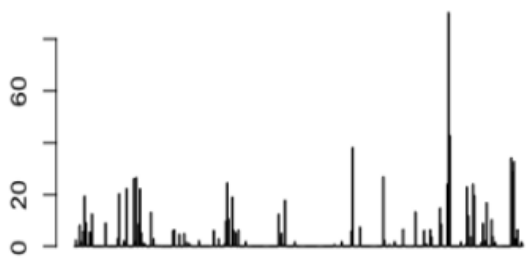


Figure: Données pluviomètres météo France

Objectif : Avoir un modèle génératif de données de pluie

Etape 1 : Proposer un modèle paramétrique qui génère des données de pluie

1. Modèles Markoviens censurés
2. Modèles semi-markoviens

$$(x_1....x_n) \sim G_\phi$$

Modifier les paramètres changera les propriétés de la série générée.

Etape 2 :Inférer les paramètres correspondant aux mieux aux données observées

On note L_ϕ la likelihood de la série générée par le générateur, et $(x_1^*...x_n^*)$ les données de pluie observées, on cherche donc

$$\phi^* = argmax_\phi L_\phi((x_1^*...x_n^*))$$

Etape 3 : On peut alors générer des données de pluie similaires aux données observées

1. Etudier les régimes de pluie
2. Etudier la dépendance de ces régimes avec des covariables climatiques (température moyenne terrestre, température à la surface des océans...)

Approche entamée : Modèles markoviens censurés

On a commencé par utiliser des modèles de Markov censurés

Exemple 1

Exemple de modèle : Censored Stochastic Reccurent Equation

$$X_t = [A^{\xi}_{1,\sigma_1,t}\mathbb{1}(X_{t-1} > 0) + A^{\xi}_{2,\sigma_2,t}\mathbb{1}(*X_{t-1} > 0)] \times X_{t-1} + B_t$$

$$\tilde{X}_t = \begin{cases} 0 \text{ if } X_t \leq 0 \\ X_t \text{ if } X_t > 0 \end{cases} .$$

- $A_{1,\sigma_1,t}$ follows a log-normal distribution
- $B_t = +/-1$ avec probabilité $\frac{1}{2}$

Difficulté : estimation de la likelihood en cadre censuré

$$X_t = [A^{\xi}_{1,\sigma,t}\mathbb{1}(X_{t-1} > 0) + A^{\xi}_{2,\sigma',t}\mathbb{1}(*X_{t-1} > 0)] \times X_{t-1} + B_t$$

$$\tilde{X}_t = \begin{cases} 0 \text{ if } X_t \leq 0 \\ X_t \text{ if } X_t > 0 \end{cases} .$$

On note $f(x_1 \dots x_n|\phi)$ avec $\phi = (\xi, \sigma, \sigma')$ la vraissemblance de la chaîne non observée $(X_1 \dots X_n)$.Son estimation est rendue difficile par la censure.

Cadre non-censuré

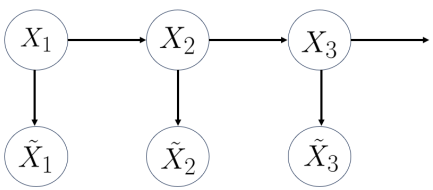


Figure: HMM Non censuré

$f(x_1, x_2, x_3|\phi)$

Cadre censuré simple

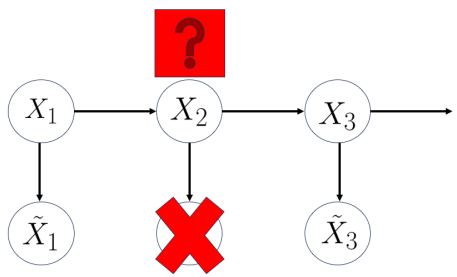


Figure: HMM censuré

$\int_{-\infty}^{0} f(x_1, x_2, x_3|\phi)dx_2$

Cas général En notant $c_t = (x_1 \dots x_t)$ la totalité de la chaîne et $c_t^{\leq 0}$ le vecteur de variables censurées

$$\int_{-\infty}^{0} f\left(\mathbf{c}_t \mid \boldsymbol{\phi}\right) \mathrm{d}\mathbf{c}_t^{\leq 0}$$

Cette intégrale intractable quand il y a beaucoup de censure, ce qui est notre cas.

Pistes de solution : estimation de la likelihood en cadre censuré

1. **Expectancy-Maximisation algorithm.** Essayer successivement d'estimer et de maximiser de façon approximative la likelihood. Méthode rapidement limitée.
2. **Pairwise Composite likelihood methods.** Utiliser une approximation de la likelihood. Celle-ci ne travaille pas sur la totalité de la chaîne mais sur des "paires" successives d'observations : $\mathcal{L}_{C,pair} = \prod_{t=1}^{n} f_{\alpha,\alpha',\sigma,\sigma'}(x_t, x_{t+1})$
3. **Méthodes likelihood-free.** Utiliser des méthodes "black-box" qui font une estimation des paramètres en fonction de la chaîne. (Deep Learning : Multilayer Percepron, Convolutionnal Neural Networks, Long-Short-Term-Memory nets, CNN-LSTM architectures ...)

Conclusion

Questions ?