7èmes Rencontres de Statistique

Science des Données

Environnement et Climat

1-2 décembre 2022

Université Bretagne Sud

Amphi B001 - Bâtiment SSI-ENSIBS

Campus de Tohannic - F-56000 Vannes

Comités d'organisation et scientifique

Présidents : Gilles Durrieu et François Septier

Arlette Antoni, Anne Cuzol, Thierry Dhorne, Evans Gouno, Ion Grama, Salim Lardjane, Audrey Poterie

Conférenciers et intervenants invités

Essoham Ali, UBS, LMBA, Vannes

Philippe Bastien, L'Oréal Paris

Liliane Bel, AgroParisTech, MIA-Paris-Saclay, Palaiseau

Anne Cuzol, UBS, LMBA, Vannes

Gilles Durrieu, UBS, LMBA, Vannes

Arthur Fétiveau, UBS, LMBA, Vannes

Philippe Naveau, Laboratoire des Sciences du Climat et de l'Environnement, Gif-Sur-Yvette

Evans Gouno, UBS, LMBA, Vannes

Naomi Ouachene, INRAE Centre Bretagne-Normandie, Rennes

Charlotte Pelletier, UBS, IRISA, Vannes

Tristan Senga-Kiesse, INRAE Centre Bretagne-Normandie, Rennes

Roméo Tayewo, UBS, LMBA, Vannes

Jeudi 1 décembre 2022

— **10h15 : Accueil des participants et café** dans l'amphi B001, Bâtiment SSI-ENSIBS, Université Bretagne Sud, Rue André Lwoff, 56000 Vannes.

— **10h45-11h : Introduction**

— **11h-12h : Liliane Bel, AgroParisTech**

Trend analysis of extremes for spatio-temporal fields. Application to climate variables.

In the framework of climate change a big concern is the evolution of extreme events, such as floods, droughts or heat waves. We consider spatio-temporal fields and thanks to ell-Pareto models that describe threshold exceedances for some kind of events determined by the ell function, we investigate the presence of a trend in the extremes. To evaluate the strength of the trend we produce return level maps in a non-stationary framework. Applications for several climate variables in different parts of the world are provided.

— **12h-14h : Déjeuner au Tableau**(https://www.au-tableau.com/)

— **14h-15h : Tristan Senga Kiesse et Naomi Ouachene, INRAE**

Modélisation statistique de systèmes environnementaux : cas d'étude de l'analyse des variations extrêmes des services écosystémiques des sols et de la modélisation des dépendances au sein des systèmes d'élevage.

De nombreuses études portent sur l'évaluation des performances environnementales des systèmes d'élevage car ils sont une source non négligeable d'émission de gaz à effet de serre dues à l'activité humaine. L'amélioration des impacts environnementaux de ces systèmes inclut, notamment, de mieux étudier leur environnement extérieur (par exemple, les propriétés

des sols, les conditions météorologiques) et aussi de mieux représenter les relations entre les variables descriptives de ces systèmes. Dans ce contexte, nous présentons d'abord une étude des indicateurs des services écosystémiques des sols (des services que les sols peuvent rendre à l'humanité) par la théorie des valeurs extrêmes. Cette étude vise à identifier les variations atypiques des services écosystémiques des sols en fonction de propriétés des sols et des conditions météorologiques. Nous présentons ensuite un travail de modélisation des relations de dépendance entre les variables descriptives des systèmes d'élevage bovins laitiers par l'approche des copules. Ce travail vise à étudier les synergies et les antagonismes qui peuvent améliorer ou détériorer l'efficacité des actions mises en place pour réduire les impacts environnementaux de ces systèmes.

— **15h-16h : Philippe Naveau, Laboratoire des Sciences du Climat et de l'Environnement**

Climate Extreme Event Attribution and Extreme Value Theory

Numerical climate models are complex and combine a large number of physical processes. They are key tools in quantifying the relative contribution of potential anthropogenic causes (e.g., the current increase in greenhouse gases) on high-impact atmospheric variables like heavy rainfall or temperatures. These so- called climate extreme event attribution problems are particularly challenging in a multivariate context, that is, when the atmospheric variables are measured on a possibly high-dimensional grid. In addition, global climate models like any in sillico numerical experiments are affected by different types of bias. In this talk, I will discuss about how to combine to two statistical theories to assess causality in the context

of extreme event attribution. In addition, the question of uncertainties quantification that remains a challenge in any climate attribution analysis will be explored from various directions. In particular, a simple model bias correction step for records will described in details. To illustrate our approach, we infer emergence times in precipitation from the CMIP5 and CMIP6 archives.

Joint work with Anna Kiriliouk, Paula Gonzalez, Soulivanh Thao and Julien Worms

— **16h-16h15 : Pause café**

— **16h15-17h : Témoignages d'anciens étudiants**

  — Baptiste Bouillon, Preligens Paris (diplômé en 2020)

  — Valentin Langrognat-Mulier, STEF Theix (diplômé en 2020)

  — Naomi Ouachene, INRAE Rennes (diplômée en 2021)

— **17h-18h : Débat sur le thème des Data Science en environnement et climat**

— **19h30 : Diner au Piano barge** http://www.pianobarge.com/Page/Accueil

Vendredi 2 décembre 2022

— **8h20 : Accueil des participants et café** dans l'amphi B001, Bâtiment SSI-ENSIBS, Université Bretagne Sud, Rue André Lwoff, 56000 Vannes.

— **8h30 : Roméo Tayewo, UBS LMBA**

Prédiction des émissions de CO2 par un modèle de régression sur graphe.

Le changement climatique est l'un des principaux défis actuels auxquels doit faire face l'humanité. Les gaz à effet de serre sont à la base du réchauffement climatique et toute politique de lutte doit s'évertuer à les réduire. Afin de planifier sereinement ces politiques, il faudrait pouvoir prédire correctement les émissions futures. Dans ces travaux, nous nous sommes attelés à prédire l'un d'entre eux, à savoir, le CO2. En considérant plusieurs comtés des États-Unis comme les nœuds d'un même graphe, nous avons appliqué une régression de type pénalisée basée sur des caractéristiques de ce graphe et combinée à la pénalisation ridge classique afin de prédire les émissions futures de CO2 de ces comtés.

Dans cette présentation, j'expliquerai comment l'estimation des paramètres est effectuée et je présenterai les résultats obtenus avec le modèle proposé en le comparant à d'autres modèles de régression pénalisée.

— **9h00 : Arthur Fétiveau, UBS LMBA**

Modèles à base de distance sur les permutations.

Lorsque l'on fait une même requête, dans une même base de données de n éléments, avec différents moteurs de recherche, on souhaiterait obtenir le classement parfait de la pertinence des réponses à notre requête. Cependant, tous les résultats ne sont pas identiques, certains algorithmes sont plus précis que d'autres, mais tous visent le résultat parfait. Ici, nous pouvons représenter un classement avec une permutation. L'ensemble des classements possibles

n'est autre que l'ensemble des permutations de n éléments.

Dans cette présentation, je présenterai différentes distances permettant de comparer des permutations. Je parlerai ensuite des modèles de Mallows, des lois de probabilités basées sur des distances entre permutations. J'expliquerai comment estimer les paramètres du modèle. Pour finir, j'aborderai les cas où l'on ne dispose que d'une partie du classement.

— **9h30 : Evans Gouno, UBS LMBA**

Inférence pour les processus auto-excités : une application à l'occurrence des impacts de foudre.

Les processus auto-excités (PAE) introduits dans les année 70 par Hawkes, sont caractérisés par une intensité qui dépend de toute ou d'une partie, de l'histoire du processus lui-même. Ils trouvent des applications dans de nombreux domaines : sismologie, neurophysiologie, génétique, épidémiologie, finance, fiabilité.

Nous présentons un travail sur l'estimation des paramètres de l'intensité d'un PAE, motivé par une étude industrielle concernant la fiabilité de matériels électriques. Il s'agit d'évaluer l'influence de l'activité orageuse sur la propension à la panne du matériel. Le travail s'appuie sur des données d'occurrences d'impacts de foudre collectées sur une période de plusieurs années. Nous développons des méthodes d'inférence par maximum de vraisemblance et envisageons une approche bayésienne. Nous explorons également la relation entre des variables exogènes et l'occurrence des impacts.

— **10h00 : Gilles Durrieu, UBS LMBA**

Modélisation des valeurs extrêmes en environnement.

Nous présentons une méthode statistique pour estimer les quantiles extrêmes. L'idée de

notre approche est d'ajuster la queue de la fonction de distribution de cette vitesse par une distribution de Pareto au delà d'un seuil que nous déterminerons. Le paramètre de la loi de Pareto est estimé en utilisant un estimateur à noyau non paramétrique de taille de fenêtre h à partir des observation plus grande que le seuil. Nous donnons sous des hypothèses de régularités les vitesses de convergence des estimateurs des quantiles extrêmes et du paramètre de la loi de Pareto. Une application de cette méthode dans le domaine de l'environnement fournit en temps réel une analyse du comportement d'un bioindicateur du milieu marin et apparaît comme un moyen efficace pour la surveillance de la qualité des eaux d'un système aquatique ainsi que pour la mesure des effets du réchauffement climatique.

— **10h30 : Essoham Ali, UBS LMBA**

Inférence statistique dans un modèle de régression Bell à inflation de zéros

Dans cet article, nous étudions les propriétés asymptotiques de l'estimateur du maximum de vraisemblance (EMV) pour un modèle de régression de Bell à inflation de zéros. Sous certaines conditions de régularité, nous établissons que l'estimateur est cohérent et asymptotiquement normal. Ceci apporte un soutien substantiel aux résultats empiriques qui ont déjà été obtenus par certains auteurs. Des simulations de Monte Carlo sont effectuées pour illustrer numériquement les principaux résultats. Le modèle est appliqué à un ensemble de données sur la demande de soins de santé aux états-Unis.

— **11h-12h : Philippe Bastien, L'Oréal R&D**

Un voyage à travers la notion de causalité

De façon surprenante la causalité est un domaine émergent en science de nos jours. Même si cette notion semble familière, poser les questions en terme de causalité jusqu'à récemment

pouvait être considéré comme non scientifique en dehors des essais randomisés. Cela est dû à plusieurs raisons : le manque d'un langage mathématique associé à la causalité, l'utilisation généralisée des équations algébriques en science, incapables d'exprimer une relation causale asymétrique, ou le rôle centrale conféré à la corrélation par Karl Pearson, disciple de Francis Galton qui la découvrit à la fin du 19ème siècle, niant par la même la nécessité d'un concept indépendant de causalité au-delà de la corrélation.

Malgré les critiques de leur paires, des approches causales ont été proposées au cours du 20ème siècle par des généticiens (Segal Wright, 1920), économistes (Haavalmo, Wold, 1960), ou statisticiens (Neyman 1920, Rubin 1970), mais il a fallu attendre les années 1980 pour voir émerger avec les travaux de Judea Pearl, prix Turing 1984, une formulation mathématique complète de la causalité avec le do-calculus associé à une approche graphique.

Il faut voir la causalité comme un enrichissement de la Statistique permettant d'accéder à une partie du monde que les méthodes statistiques traditionnelles ne peuvent approcher. Sans une vision causale des paradoxes apparaissent comme les paradoxes de Simpson ou de Berkson. Il peut sembler en particulier paradoxal pour un statisticien conventionnel que l'on puisse tirer des conclusions différentes, voire opposées, suivant que l'on utilise les mêmes données sous forme agrégée ou segmentée. En régression la notion de causalité transparaît à travers le choix des régresseurs qui peuvent biaiser ou non les résultats. Sans vision causale il n'existe pas de solution générale à ce problème, la solution étant dépendante d'hypothèses causales qui ne peuvent être exprimées uniquement à travers le langage des statistiques. On montrera à travers des exemples simples comment grâce à l'utilisation d'un modèle causal de représentation du domaine étudié, on peut à partir des données et de simples corrélations avoir une interprétation en terme de causalité sans recours à un essai randomisé.

Pearl a présenté une échelle de la causalité à 3 niveaux. Le premier niveau, dépourvu de

causalité, correspond aux données et à la corrélation, le second niveau correspond au premier niveau de causalité associé à une intervention qui doit permettre d'éliminer les chemins non causaux, enfin le troisième niveau correspond au contre factuel associé à l'imagination qui permet de répondre à des questions pour lesquelles on a pas d'observations. Cela revient à revenir dans le passé, le modifier, et en observer les conséquences. Le contrefactuel permet en particulier d'estimer des effets directs et indirects quand l'effet peut s'exprimer à travers un médiateur. Ce domaine en plein expansion trouve de nombreuses applications, en sociologie, justice, médecine, fusion de données, problème de transportabilité, et est présenté comme la prochaine révolution des réseaux profonds par la prise en compte de la causalité pour un modèle plus robuste, généralisable et plus facilement interprétable.

— **12h-14h : Déjeuner au Tableau** (https://www.au-tableau.com/)

— **14h-15h : Charlotte Pelletier, UBS IRISA**

Automatic mapping of the land surfaces from Earth observation time series

Nowadays, modern imaging sensors on board the satellites give (free) access to a large amount of Earth observation data. For example, the twin Sentinel-2 satellites from the European Space Agency (ESA) provide images of the entire Earth every two to five days, referred to as satellite image time series (SITS). A frequent acquisition of images is crucial for vegetation-related remote sensing applications such as crop type classification. The automatic transformation of spatio-temporal data cubes into meaningful information and products (e.g., land cover, crop-type, or deforestation maps) usually relies on supervised learning. Recent advances in this field have been marked by a shift towards deep learning methods due to

their state-of-the-art results in computer vision and natural language processing tasks. The ability of these techniques to deal with sequential data (e.g., text or audio) and to detect time-invariant characteristics results in various achievements for time series classification in several domains, including remote sensing. In this talk, I will present the potential and the limitation of deep learning techniques for the classification of SITS through several algorithms and datasets.

— **15h-16h : Anne Cuzol, UBS LMBA**

Markov Switching Multivariate Space Time model for weather variables

Hierarchical stochastic weather generators are statistical models aiming at quickly simulating realistic random sequences of atmospheric variables that include a latent categorical variable, referred to as the weather type. We propose a new Markov Switching model in a multivariate spatio-temporal setting that combines the ability of Markov Switching Auto Regressive models to account for multiple weather types with the flexibility of multivariate non-separable space-time covariance models, making it thus possible to predict and simulate weather variables at unmeasured location and time. The estimation of the parameters and the weather types is done using an EM type algorithm. The optimal number of weather types is selected using the BIC criterion. The model is then illustrated on a data set covering the Mediterranean region in France. The ability of this model to reproduce complex dependence structures is then discussed and compared to simpler models.

— **Clôture des 7èmes Rencontres.**