Sujet de la thèse	Intelligence artificielle explicable dans des séries temporelles hétérogènes : étude de l'impact des événements environnementaux et sociaux sur la consommation énergétique	
Spécialité de la thèse	Intelligence Artificielle/Machine Learning, Statistique, Technologies de l'information de la communication	
Mots-clefs	Machine Learning, eXplianable Artificial Intelligence (XAI), statistique, auto-régression, causalité	
Type de thèse	cotutelle	
Pays	France : Brest-Rennes	Tunisie
Laboratoire d'accueil	L@bISEN Yncréa Ouest, Equipe Vision - LSL	Université de Carthage, SUP'COM, Laboratoire COSIM, Equipe Traitement et analyse de l'information pour l'environnement
Directeurs de thèse	Noms des directeurs de thèse	Maher JRIDI, Amel BENAZZA
Encadrants	Encadrant1, Encadrant2	Lina FAHED, Matthieu SAUMARD

Contexte

Depuis de nombreuses années, comprendre l'impact des événements (phénomènes) environnementaux et sociaux sur la consommation énergétique devient un enjeu majeur et attire beaucoup l'attention dans les champs sociétaux, industriels et scientifiques. En effet, comprendre un événement et surtout connaître les relations de cause à effet (la causalité) permet d'assister les experts du domaine et les conforter dans leurs décisions. Prenons l'exemple de la prédiction d'un pic de consommation d'électricité : il est important de prédire au plus tôt ce phénomène dans une ville/région, de détecter ses causes directes environnementales et sociétales (éléments déclencheurs) et indirectes (signaux faibles). Ainsi, pour ce type de problème, les moyens technologiques mises en place ne cessent de se développer. A titre d'exemple, des études récentes [Palme et al, 2017] ont monté le lien entre les événements environnementaux, sociaux et la consommation énergétique. L'analyse de l'évolution des données environnementales (températures, humidité, pollution, vent, etc.) et sociaux (fêtes, rassemblements, vacances, événements imprévus/inédits) couplée avec les données de consommation énergétique, permet de détecter les corrélations cachées et de prédire les pics de cette dernière. Cette prédiction peut être utilisée par les experts afin d'anticiper, au moindre coût, la production énergétique en termes de quantité et de type (renouvelable, …). Ce qui rentre dans le cadre de ville intelligente.

Projet

De ce fait, et dès qu'il s'agit de modèles d'intelligence artificielle, la problématique de la collecte de données numériques fait surface. Les séries temporelles font partie des types de données largement collectées et étudiées. Leur succès revient principalement au développement du marché des capteurs. C'est ainsi que sont mis à disposition de la communauté scientifique des séries temporelles de relevé de température, d'humidité, de pression, de consommation énergétique, d'événements sociaux [data-ref][Owayedh et al, 2000,] [Grolinger et al, 2016], etc.

Cependant, depuis plusieurs années, un nouveau phénomène lié aux données numériques émerge : des données de plus en plus volumineuses et hétérogènes, apparaissent. La diversité des points/méthodes de collecte (capteurs) fait émerger un nouveau défi : la fusion de ces sources de données hétérogènes. Par conséquent, la modélisation de ce type de données s'impose.

D'autre part, de nouvelles exigences sociétales apparaissent, il s'agit des demandes pressantes pour rendre le processus de modélisation transparent afin de fournir des explications claires aux experts du domaine d'application [Goodman et Flaxman, 2017].

Problématiques

Dans ce projet, nous nous focalisons sur un type particulier de données complexes : les séries temporelles et posons la question suivante : **Comment faire des séries temporelles un outil pour l'aide à la prédiction et à la détection de l'émergence des phénomènes comme la consommation énergétique ? Quels modèles explicables aux experts faut-il mettre en place ?**

Pour répondre à cette question, nous proposons un système intelligent qui sera validé sur des données actuellement disponibles [data-ref] et par des experts du domaine.

Dans ce projet, nous nous focalisons sur deux défis majeurs :
(I) **La prédominance de nouveau type de données représenté par des séries temporelles complexes** : En effet, la fusion ou l'agrégation des informations apportées par chaque série (environnement, social, énergie) devient une tâche complexe car avec le grand volume de données disponibles, la portée temporelle et l'impact d'une variable deviennent plus importants. Mais cet impact devient moins visible et moins facilement détectable surtout à cause de chevauchement de plusieurs phénomènes [Fahed et al, 2018]. Par conséquent, deux difficultés se présentent : (i) La détection de causalité : Souvent les méthodes proposées sont purement statistiques (comme la causalité au sens de Granger) [Saumard, 2017] et sont peu performantes face à la complexité des données et quand il s'agit de détection des liens sur le long terme entre plusieurs séries [Mavrotas et Kelly, 2001]. (ii) La détection de l'émergence au plus tôt et sur le long terme.
(II) **Les exigences actuelles de la société pour rendre le processus de modélisation explicable et transparente** : Parmi les méthodes existantes de prédiction dans des séries temporelles complexes, nous pouvons citer les méthodes de fouille de données telles les réseaux de neurones, méthodes ensemblistes, etc. Ces méthodes sont très performantes, mais elles restent des boîtes noires, à savoir le processus de modélisation est opaque et plusieurs questions se posent sur son explicabilité et la compréhension du résultat. En revanche, des méthodes de prédiction basées sur des modèles auto-régressifs, sont relativement moins performantes mais peuvent être considérées comme étant transparentes et explicables. Récemment, différentes méthodes d'explication [Došilović et al, 2018], indépendantes des modèles de prédiction, ont été proposées. Cependant, très rares sont les méthodes d'explicabilité qui s'intéressent à la causalité dans les séries.

Approches méthodologique et technique envisagées

Nous proposons un système qui représente un cycle complet : les données, qui sont à notre disposition, sont pré-traitées et analysées, puis une "modélisation prédictive explicable" à base de fouille de séries temporelles sera proposée. Par conséquent, nous pouvons définir deux tâches principales et inséparables :
(I) La modélisation prédictive : Nous trouvons qu'il est prometteur d'exploiter les méthodes statistiques existantes de détection de ruptures (abrupts) ou de pics dans des séries temporelles (comme les méthodes paramétriques, non-paramétriques ou semi-paramétrique. Notre objectif sera ensuite de proposer un nouveau modèle de prédiction qui aura pour originalité de tirer profit de la transparence des méthodes statistiques et de la performance des méthodes d'intelligence artificielle opaques. Afin d'étudier la causalité dans notre modèle, nous proposons d'exploiter et d'adapter les méthodes statistiques existantes de causalité au sein des méthodes de fouille de données. Notre objectif est de détecter une causalité temporelle distante en effectuant une modélisation temporelle pour détecter l'impact sur le long terme.
(II) La modélisation explicable : Parmi les questions qui se pose : à quel moment de la modélisation doit-on intégrer l'explicabilité ? Contrairement aux méthodes existantes qui fournissent des explications partielles, nous proposons un modèle complet dans lequel des explications sont fournies tout au long du processus de la fouille.

Références succinctes

- [data-ref] Sites des données environnementales (consultés le 01/02/2020) :
  - Observation météorologique historiques France (SYNOP) https://opendata.paris-saclay.com/explore/dataset/observation-meteorologique-historiques-france-synop-orly/table/?sort=date
  - Météo France, données publiques https://donneespubliques.meteofrance.fr/
  - Data Gouv. Energie : https://www.data.gouv.fr/fr/datasets/r/e9cc5f55-b74f-418d-9361-a15629648472
- [Došilović et al, 2018] Došilović, Filip Karlo, Mario Brčić, and Nikica Hlupić. "Explainable artificial intelligence: A survey." 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE, 2018.
- [Fahed et al, 2018] Fahed, Lina, Armelle Brun, and Anne Boyer. "DEER: Distant and essential episode rules for early prediction." *Expert Systems with Applications* 93 (2018): 283-298.
- [Goodman et al, 2017] Goodman, Bryce, and Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"." *AI magazine* 38.3, 2017.

- [Palme et al, 2017] Palme, M., et al. "From urban climate to energy consumption. Enhancing building performance simulation by including the urban heat island effect." Energy and buildings 145, 2017.
- [Mavrotas et Kelly, 2001] Mavrotas, G. et Kelly, R. Old Wine in New Bottle : Testing Causality between Savings and Growth. Manchester School, 69, 2001.
- [Saumard, 2017] Saumard, Matthieu. "Linear causality in the sense of Granger with stationary functional time series." *Functional Statistics and Related Fields*. Springer, Cham, 2017.
- [Grolinger et al, 2016] Grolinger, Katarina, Miriam AM Capretz, and Luke Seewald. "Energy consumption prediction with big data: Balancing prediction accuracy and computational resources." 2016 IEEE International Congress on Big Data (BigData Congress). IEEE, 2016.
- [Owayedh et al, 2000] Owayedh, M. S., A. A. Al-Bassam, and Z. R. Khan. "Identification of temperature and social events effects on weekly demand behavior." Power Engineering Society Summer Meeting. IEEE, 2000.

Profil recherché

- La ou le candidat(e) doit avoir un diplôme de Master et/ou Ingénieur dans des domaines liés à l'informatique, mathématiques appliquées, statistique, science des données ou traitement de signal
- Avoir une aptitude au développement de méthodes d'intelligence artificielle, machine learning, statistique, analyse des données.
- Avoir un vif intérêt pour la recherche scientifique et être familier au moins avec les outils informatiques/langages suivants : python (scikit-learn), R, …
- Avoir un bon niveau d'anglais écrit et oral

Financement de la thèse en cotutelle

Le sujet de thèse est financé par les fonds propres du laboratoire L@bISEN de l'ISEN Yncréa Ouest. Le doctorant, au cours de sa période de présence en France (période de 4 à 6 mois par année), est financé avec une bourse de thèse à hauteur de 1000€/mois.

Une bourse d'alternance pour un séjour prolongé en France peut être également demandée au niveau de l'Université de Carthage.

Au cours de son séjour en Tunisie, le(a) candidat(e) bénéficiera de la bourse du Ministère de l'Enseignement Supérieur.

Modalités de candidature

Le dossier de candidature doit comprendre, en un seul PDF, votre CV, lettre de motivation, Relevés de notes de L3, M1, M2 (ou années équivalentes) et les noms de 2-3 référents à contacter ou éventuellement des lettres de recommandation.

L'ensemble du dossier (un seul PDF) doit être adressé le plus tôt possible à :

- Amel BENAZZA : benazza.amel@supcom.tn
- Maher JRIDI : maher.jridi@isen-ouest.yncrea.fr
- Lina FAHED : lina.fahed@isen-ouest.yncrea.fr
- Matthieu SAUMARD : matthieu.saumard@isen-ouest.yncrea.fr