OFFRE DE STAGE EN *DIFFERENTIAL PRIVACY* ET APPRENTISSAGE SEQUENTIEL

Mission : Proposition d'algorithmes confidentiels pour la publication séquentielle de relevés de consommation électrique

Durée : 6 mois - Début du stage souhaité : avril 2026 - Lieu : EDF R&D Lab Saclay (91120)

Contexte

La R&D d'EDF (2000 chercheurs) a pour missions principales de contribuer à l'amélioration de la performance des unités opérationnelles du groupe EDF, d'identifier et de préparer les relais de croissance à moyen et long terme. Dans ce cadre, le département Services, Economie, Outils Innovants et IA (SEQUOIA) est un département pluridisciplinaire (sciences de l'ingénieur, sciences humaines et sociales) qui fournit un appui à l'élaboration et au portage des offres, des services et des outils de relation client aux directions opérationnelles du groupe EDF. Au sein de ce département, ce stage sera rattaché au groupe « Statistiques et Outils d'Aide à la Décision » (SOAD) : cette équipe compte une vingtaine d'ingénieur·es chercheur·es spécialisé·es en IA et *data science* avec des compétences fortes autour du *machine* et du *deep learning*, du *web* sémantique, de l'IA symbolique et de l'IA générative.

Dans ce contexte, l'équipe :

- Réalise la veille et le test des solutions émergentes dans le monde académique et industriel
- Oriente les entités du Groupe EDF vers les meilleurs choix technologiques pour les besoins opérationnels
- Élabore des méthodes et des outils permettant de gagner en performance sur l'analyse de données structurées (tabulaires, séries temporelles) et non structurées (texte, image, son, …)
- Réalise des études et des POC
- Valorise les résultats obtenus sous forme de démonstrateurs, articles scientifiques, brevets

Objectifs

Afin d'impulser la transition énergétique *via* la mise en œuvre de nouveaux services (optimisation des flexibilités électriques, *smart grids*, …), les gestionnaires des réseaux de distribution électrique sont encouragés à publier des données de consommation agrégée (que ce soit à l'échelle d'un quartier ou de quelques milliers de foyers) tout en garantissant la protection de leurs clients (en réduisant le risque de fuite d'information individuelle).

Pour répondre à ces deux exigences contradictoires (compromis utilité- confidentialité), une approche par *differential privacy*, adaptée aux spécificités des séries temporelles de relevés de consommation demi-horaire, a été proposée par Agoua *et al.* [1]. Dans un premier temps, la série temporelle est projetée sur une base d'ondelettes bien choisie. Le processus de *differential privacy* est ensuite appliquée sur les coefficients de la projection : du bruit est ajouté à ces derniers, de sorte à garantir la confidentialité de chaque client. La transformée en ondelettes inverse permet ensuite de reconstruire la série temporelle, désormais protégée.

L'objectif du stage est de poursuivre ces travaux dans un cadre de publication en ligne : nous imaginons une publication hebdomadaire ou mensuelle de la consommation d'un même agrégat de clients. Pour se faire, les coefficients de la projection dans la base d'ondelettes devront être mis à jour, et à chaque fois reprotégés. La publication en ligne à l'aide d'un mécanisme de *differential privacy* classique aura une utilité limitée puisque, pour garantir le même niveau de confidentialité qu'une publication *off-line*, l'écart-type du bruit à ajouter devra être multiplié par le nombre de publication.

Afin de maximiser l'utilité, l'algorithme FAST, proposé par Fan and Li [2], n'observe qu'à certains instants la série temporelle. Il estime un modèle espace-état (filtre de Kalman, par exemple) sur ces observations protégées (en ajoutant du bruit) et l'utilise pour prévoir les valeurs de la série aux instants non observés. Les instants d'observation sont choisis de manière adaptative en fonction de la dynamique de la série temporelle détectée. Adapter ce type d'approche sera la première piste à explorer.

Étapes du stage :

- Etude bibliographique et prise en main des codes déjà développés pour la publication unique d'une courbe de chargé agrégée
- Modélisation de l'aspect séquentiel de la publication et proposition d'algorithmes
- Test sur un jeu de données *open source*, optimisation des hyperparamètres de l'algorithme en fonction du budget de confidentialité et interprétation des résultats
- Rédaction du rapport de stage

Profil recherché :

- Etudiant·e en Master 2 ou équivalent école d'ingénieur
- Compétences en statistiques, apprentissage automatique et séquentiel et programmation (Python)
- Bon niveau de rédaction en français et en anglais
- Curiosité scientifique, intérêt pour la recherche

Références :

[1] Agoua, Ghislain, et al. "DIFFERENTIAL PRIVACY FOR ENERGY DATA PUBLICATION." *IET Conference Proceedings CP785*. Vol. 2021. No. 6. Stevenage, UK: The Institution of Engineering and Technology, 2021.

[2] Fan, Liyue, and Li Xiong. "Real-time aggregate monitoring with differential privacy." *Proceedings of the 21st ACM international conference on Information and knowledge management*. 2012.

Informations pratiques

Unité d'accueil : Groupe SOAD (Statistique et Outils d'Aide à la Décision), département SEQUOIA d'EDF Lab Paris-Saclay, 7 boulevard Gaspard Monge, 91120 Palaiseau.
Le stage sera encadré par des ingénieur·es chercheur·es du département SEQUOIA.

Transmettre par mail un CV et une lettre de motivation à : margaux.bregere@edf.fr (Département SEQUOIA).