Proposition de stage de Master 2

Sujet

Classification de données fonctionnelles multivariées pour attester les mobilités écoresponsables

Contexte

Depuis une dizaine d'années, plusieurs cryptomonnaies ont été développées afin de promouvoir des échanges dématérialisés de pair à pair et en se passant d'un intermédiaire tels que le sont les banques ou autres organismes financiers. Pour la plupart de ces cryptomonnaies, la création de nouvelles unités repose généralement sur l'accomplissement de tâches virtuelles, comme la résolution d'un problème numérique. Ceci induit aujourd'hui une utilisation intensive de serveurs de calculs et dont le coût environnemental est problématique. Pour pallier ce problème, de récentes cryptomonnaies se développent sur la base de protocoles de création d'unités beaucoup moins énergivores, certains allant même jusqu'à valoriser des comportements écoresponsables. Par exemple Solarcoin récompense en unités monétaires les producteurs d'électricité par panneau photovoltaïques, ou encore Regen qui encourage les agriculteurs mettant en œuvre des pratiques régénératrices des terres. Toutefois, à notre connaissance, aucune tentative n'a encore été amorcée dans le domaine des transports, pourtant majoritairement responsables de l'émission de CO2 dans l'atmosphère, et donc du dérèglement climatique. Une piste innovante est la création d'une cryptomonnaie pour valoriser l'usage des mobilités moins énergivores et polluantes comme le vélo, le covoiturage ou l'utilisation de transports en commun. Pour l'établissement d'une telle cryptomonnaie, une question primordiale est de détecter à partir de données GPS comment se déplacent les individus, avec une précision suffisante pour exclure la possibilité de fraude (individus voulant faire croire à des déplacements écoresponsables). Bien que des solutions existent déjà, les contraintes suivantes du contexte de ce projet ne rendent pas utilisables ces solutions :

- Afin de respecter les recommandations du RGPD et la vie privée des citoyens, il ne doit pas être possible de déterminer la présence d'un citoyen à un endroit donné et à un instant donné. Cela implique que les données GPS des utilisateurs doivent être anonymisées.
- Le système établi ne doit pas faire appel à des informations fournies par des services tiers qui pourraient constituer autant de failles de sécurité. Cela implique de ne pas pouvoir recouper les données GPS avec des cartes.
- Si pour d'autres contextes les conséquences d'une fraude non-détectée (détecter à tort que la personne se déplace de manière écoresponsable) ne sont pas coûteuses, dans ce contexte la précision de détection est un élément important pour l'intérêt et la confiance du système proposé. Plus précisément, ce système économique ne serait pas viable si par des moyens détournés, un fraudeur arrive à générer une grande quantité d'unités monétaires sans avoir accompli de comportement écoresponsable.

Ces différentes contraintes impliquent la nécessité de développer une méthodologie spécifique pour ce projet de recherche.

Travail à effectuer

Ce stage est une des étapes de ce projet de recherche, et il a pour objectif de répondre à cette problématique avec une approche probabiliste des traces GPS. Plus précisément, la modélisation des traces GPS, comme des processus stochastiques ou des données fonctionnelles, permettrait de prendre en compte la nature fonctionnelle et temporelle des données. Une modélisation adaptée permettrait de considérer une structure temporelle prédéfinie aux données, de sorte à compenser la faible quantité de données, relativement à la complexité du problème, tout en offrant de bonnes performances de prédiction. Afin de décrire les pistes à envisager, il est nécessaire de donner plus de détails concernant les données. Pour chaque traces GPS d'un utilisateur effectuant un trajet, nous considérons qu'en prenant en compte les dérivées de la position (après de potentielles dilatations/translations du temps), il n'est pas probable de pouvoir retrouver des informations concernant la réelle localisation temporelle des utilisateurs, ce qui permet d'assurer la sécurité des individus concernant le données personnelles. Or, des travaux précédents [3] indiquent que les trois premières dérivées de la position (vitesse, accélération et secousse) font partie des grandeurs pertinentes pour différencier les modes de mobilité. Nous considérons donc disposer, pour chaque trajet d'un individu, d'une donnée fonctionnelle multivariée. L'objectif est alors de déterminer une règle d'affectation permettant d'attribuer un mode de mobilité à une donnée fonctionnelle multivariée. Une piste envisageable à ce stade consiste à adapter les travaux de classification de données fonctionnelles [1, 4] au cas

spécifique des données étudiées. Notammment, une spécificité à prendre en compte pour modéliser correctement le problème est la relation (de dérivation par rapport au temps) entre les différentes dimensions des données fonctionnelles. Une seconde piste à étudier consiste à modéliser ces données comme des processus stochastiques afin d'évaluer si la structure de ces données peut être plus efficacement modélisée ainsi, plutôt qu'avec des bases de fonctions. Des travaux récents indiquent que cette approche est adaptée pour modéliser des données GPS [2].

Parmi les tâches qui pourront être envisagées par le stagiaire, voici ci-dessous une liste des points (ordonnés chronologiquement) qui paraîssent les plus pertinents :

- prendre en main d'une base de données qui sera mise à disposition,
- récupérer et comprendre le code et les approches existantes au sein de l'équipe de recherche,
- se former à utiliser un git et un serveur de calcul (via `ssh`),
- s'approprier les éléments bibliographiques relatifs au sujet du stage,
- mettre au point une méthodologie et l'implémenter, et
- comparer la méthodologie proposée avec d'autres approches existantes.

Profil recherché

- Niveau équivalent Master 2 en mathématiques appliquées, orienté statistique ou probabilités.
- Connaissances en analyse de données fonctionnelles ou en processus stochastiques.
- Une bonne maîtrise de l'environnement Linux sera un plus non négligeable.
- Bonne maîtrise de *R* et/ou *python*.

Superviseurs

Paul-Marie Grollemund (paul_marie.grollemund@uca.fr), Pascal Lafourcade (pascal.lafourcade@uca.fr) et Kevin Atighehchi(kevin.atighehchi@uca.fr)

Organisme d'accueil

Ce stage se fera à l'Université Clermont Auvergne au Laboratoire de Mathématiques Blaise Pascal.

Indemnisation

Taux légal : une gratification de 536 euros net par mois.

Références

[1] Andrés M Alonso, David Casado, and Juan Romo. Supervised classification for functional data : A weighted distance approach. *Computational Statistics & Data Analysis*, 56(7) :2334–2346, 2012.

[2] R Barzaghi and Alessandra Borghi. Theory of second order stationary random processes applied to gps coordinate time-series. *GPS Solutions*, 22(3) :1–12, 2018.

[3] Sina Dabiri and Kevin Heaslip. Inferring transportation modes from gps trajectories using a convolutional neural network. *Transportation research part C : emerging technologies*, 86 :360–371, 2018.

[4] Julien Jacques and Cristian Preda. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71 :92–106, 2014.