## SUJET

Titre : Prédiction de l'environnement d'un quartier

Superviseur : Franck Favetta, Fabien Duchateau (enseignants-chercheurs UCBL et LIRIS)

Mode de financement : gratifications de stage, Labex IMU

Partenaires : sociologues du Centre Max Weber (Lyon)

## DETAIL

L'INSEE découpe la France en 50000 IRIS (zones géographiques d'environ 5000 personnes), que nous considérons comme des « quartiers ». Chaque IRIS possède plusieurs centaines d'indicateurs (e.g., revenu moyen, nombre de boulangeries, nombre de logements construits avant 1950 ou encore nombre de résidents par catégorie socio-professionnelle). Dans le cadre du projet IMU-HiL qui porte sur la recommandation immobilière, des sociologues ont défini cinq variables environnementales, avec pour chacune une liste de valeurs possibles (e.g., une variable situation morphologique avec des valeurs *centre, urbain, péri-urbain, rural* ou une variable classement social allant de *populaire* à *très aisé*). Ces variables permettent de décrire simplement un quartier sous une forme pertinente pour un utilisateur en recherche d'un bien immobilier, et facilitent la comparaison de plusieurs quartiers lors d'études sociales par exemple. Les sociologues ont ensuite annoté 300 IRIS au moyen de ces variables.

L'objectif principal du stage consiste à prédire les valeurs de ces variables d'environnement pour les IRIS restants :

- La première tâche permet de sélectionner et développer un ou plusieurs classifieurs pertinents et fiables. Des résultats préliminaires ont montré que des classifieurs basiques (non paramétrés et génériques) obtenaient des résultats moyens (30 % à 55 % de prédictions correctes selon la variable considérée) ;
- De plus, il sera nécessaire d'étudier la normalisation des indicateurs (surface, quartiles, quantités, etc.). En effet, certains classifieurs nécessitent d'utiliser des informations « comparables » entre elles. La densité de population d'un IRIS est une piste pour la normalisation de ces indicateurs ;
- Si le choix de produire un classifieur par variable est privilégié, il faudra s'intéresser à la possible corrélation entre les variables. Intuitivement, il apparaît que certaines variables ne sont pas complètement indépendantes (e.g., la situation morphologique avec l'environnement paysager). Si des relations entre variables sont confirmées, la prédiction d'une variable pourrait donc faciliter la détection d'une autre ;
- Enfin, si les données d'apprentissage ne sont pas suffisantes, une piste consiste à générer un jeu de données synthétique de plus grande taille (mais partageant les mêmes caractéristiques) afin d'améliorer la qualité des prédictions.

Un prototype sera développé sur la base de l'outil interne *mongiris*, qui permet d'afficher et de rechercher des IRIS et leurs informations sur une carte. L'outil permettra de prédire et d'afficher les variables d'environnement d'un IRIS pour différents classifiers (ainsi que l'expertise humaine disponible pour les 300 IRIS). Une page spécifique permettra en plus de choisir, paramétrer, tester un classifieur et d'afficher la qualité de ses prédictions. Le code sera générique afin de faciliter l'intégration de nouveaux classifieurs et le prototype sera soumis au journal JOSS.
Les sociologues du projet IMU-HiL ont déjà commencé la rédaction d'un article pluridisciplinaire décrivant la définition des variables, et le stage devrait permettre de finaliser cet article en présentant la méthode et les résultats de prédiction pour les autres IRIS.