



financé par
IDEX Université Grenoble Alpes



LABORATOIRE
JEAN KUNTZMANN
MATHÉMATIQUES APPLIQUÉES - INFORMATIQUE

Stage de M2 en statistique - sciences des données

Classification croisée pour tests de positionnement en langues

L'Université Grenoble Alpes a développé un test de positionnement appelé SELF (Cervini et al., 2015), pour évaluer les compétences en 6 langues (anglais, italien, japonais, mandarin, espagnol et français) des étudiants entrant à l'université à partir du Cadre Européen Commun de Référence pour les Langues (CECRL), de A1 (niveau élémentaire) à C2 (niveau expert). Ce dispositif est maintenant également utilisé par d'autres universités, ainsi que par les Centres Universitaires d'Études Françaises, regroupés au sein du réseau ADCUEFE, pour le niveau en Français Langue Étrangère (FLE). Tous ces étudiants de niveaux hétérogènes ont besoin d'être dirigés vers des groupes différents afin d'être placés dans un cours qui corresponde à leur niveau et leur permette de progresser au mieux.

Pour analyser les matrices-résultats issus de ces tests, nous avons appliqué le modèle des blocs latents ou LBM (Govaert et Nadif, 2003) qui permet de construire une partition des lignes et des colonnes dans le but de faire ressortir des blocs de cases homogènes à l'intérieur d'un bloc et hétérogènes entre deux blocs.

A la suite de ce premier travail, la première partie du stage consistera à étudier le modèle des blocs latents avec classe de bruit ou NFLBM (Laclau et Brault, 2018) qui est un dérivé du LBM incluant une classe particulière pour les items qui n'auraient pas d'intérêt dans la classification des lignes. Une comparaison avec les résultats avec ceux du LBM devra être menée.

La deuxième partie du stage sera consacrée à l'analyse des résultats en fonction des profils des étudiants. Il s'agira d'identifier les différences significatives entre les publics testés (spécialistes/non spécialistes, public en formation initiale/formation continue, etc.) à l'aide de tests statistiques d'indépendance entre les groupes de niveaux (créés par SELF, LBM ou NFLBM) et les groupes de langues maternelles ou les nationalités.

Références :

- Cervini C., Jouannaud M.P. (2015). *Ouvertures et tensions liées à la conception d'un système d'évaluation numérique multilingue en ligne dans une perspective communicative et actionnelle* ». ALSIC – Apprentissage des langues et systèmes d'information et de communication. Numéro spécial 'Des machines et des langues'.
- Govaert, G., & Nadif, M. (2003). *Clustering with block mixture models*. Pattern Recognition, 36(2), 463-473.
- Laclau, C. & Brault, V. (2018) *Noise-free latent block model for high dimensional data*. Data Mining and Knowledge Discovery. <https://doi.org/10.1007/s10618-018-0597-3>

Profil recherché :

- Étudiant en M2 de statistique ou sciences des données

Compétences recherchées :

- Modèles de mélange, classification non supervisée
- Algorithme EM
- Programmation R

Date et lieu du stage :

- Durée : 6 mois
- A partir de février 2020
- Laboratoire LJK de l'Université Grenoble Alpes, dans le cadre du projet IRS COPOLangues.

Contacts :

- Vincent Brault : vincent.brault@univ-grenoble-alpes.fr
- Frédérique Letué : frederique.letue@univ-grenoble-alpes.fr
- Marie-José Martinez : marie-jose.martinez@univ-grenoble-alpes.fr