

Stage de 6 mois (possibilité d'enchaîner après sur une thèse CIFRE)*de préférence, à partir du 1^{er} octobre 2019***Intitulé** : « **Méthodologie de sélection et de hiérarchisation des déterminants d'absence et de stress au travail** »**Lieu** : Laboratoire MESuRS « Modélisation, Epidémiologie et Surveillance des Risques Sanitaires », Conservatoire national des arts et métiers, 292, rue Saint-Martin, 75003 Paris.**Encadrement** :

- Mounia N. Hocine, MCF en biostatistique, laboratoire MESuRS, Cnam ;
- Tom Duchemin, Ingénieur ENSAE-doctorant, Malakoff-Médéric-Humanis (Paris 09).

La question des arrêts maladie occupe une place de plus en plus grandissante dans les débats de santé publique avec la place dominante des risques psychosociaux dans les dépenses de l'état et des entreprises. D'une part, l'Etat doit faire face à une augmentation des coûts liés aux arrêts maladie à cause, notamment, du vieillissement de la population ; et d'autre part, les employeurs se préoccupent de plus en plus de la santé de leurs salariés et de nouveaux besoins surgissent pour comprendre les arrêts maladie. Des outils statistiques performants sont ainsi requis pour répondre à ces demandes et accompagner les chercheurs et décideurs intéressés par ces problématiques.

Les données d'arrêt maladie sont cependant des données complexes à étudier et le choix d'outils statistiques adapté s'avère compliqué. Les données d'arrêt maladie peuvent en effet être censurées, *zero-inflated* ou bien auto-corrélées mais elles sont principalement déterminées par de nombreuses variables, parfois très corrélées les uns aux autres. Afin d'identifier les déterminants de ces arrêts maladie, l'utilisation de régressions linéaires et de méthodes de sélection de variables classiques comme les régressions LASSO ou Elastic Net s'est avérée peu efficace. De premiers travaux publiés proposent l'utilisation de *forêts aléatoires conditionnelles* et de mesures d'importance de variable pour la hiérarchisation des déterminants des arrêts maladie [1]. Une nouvelle méthode de sélection de variable par construction d'indicateurs synthétiques, à nouveau fondée sur des méthodes de forêts aléatoires, a depuis été publiée [2] et semble encore plus adapté à l'étude de nos données d'absence ou de stress au travail qui offrent une grande quantité de variables explicatives très corrélées les unes aux autres.

Ce travail doit permettre de tester la validité et l'efficacité de nouvelles approches (dont la méthode citée) pour hiérarchiser l'impact de diverses variables sur l'occurrence d'arrêt maladie et sur des variables de santé à partir des données d'une enquête représentative, le Baromètre Santé et bien-être au Travail (BST). Cette enquête annuelle permet de documenter, par auto-questionnaire, une soixantaine de variables explicatives dont la moitié décrivant le contexte et l'environnement du travail. Au total, 40 000 salariés du secteur privé ont complété le questionnaire de l'étude sur les dix vagues de 2009 à 2018.

Références

1. Duchemin T., Bar-Hen A., Lounissi R., Dab W., Hocine M.N. Hierarchizing determinants of sick leave: insights from a survey on health and wellbeing at the workplace. *Journal of Occupational and Environmental Medicine*. In Press.
2. M. Chavent, R. Genuer, J. Saracco (2019). Combining clustering of variables and feature selection using random forests. *Communications in Statistics – Simulation and Computation*. DOI : 10.1080/03610918.2018.1563145.

Contact : Merci d'envoyer votre cv et votre lettre de motivation à Mounia N. Hocine mounia.hocine@cnam.fr, et Tom Duchemin t.duchemin@malakoffmederic-humanis.com.