Internship 2019

Food, Health, Wealth: Profiling and Causal Modelling

SUPERVISORS: Philippe Caillou, Michèle Sebag, Olivier Allais
LABS: TAO − CNRS − INRIA − Paris-Sud; ALISS− INRA; U. Paris-Saclay
MAIL: sebag@lri.fr, caillou@lri.fr, olivier.allais@inra.fr
URL: `http://www.lri.fr/~sebag/Stages/Stage_Nutri_19.pdf`

Abstract. Le contexte général est celui du projet Nutriperso (CNRS - INRIA - INRA - INSERM - CEA), cherchant à identifier les relations entre alimentation, santé et indicateurs socio-démographiques. Les données Kantar décrivent 20,000 familles: leur données socio-démographiques (statut familial, diplomes, niveau de vie) et l'ensemble des produits alimentaires achetés sur 20 ans (fréquence journalière); la précision est celle du code barre (180,000 produits). Seul l'indicateur *Indice de masse corporelle* est disponible pour les données Kantar ; les experts considèrent qu'il s'agit d'un indicateur fiable pour une partie des problèmes tels que diabète de type 2.

Objectifs du stage. L'une des difficultés est de prendre en compte les produits alimentaires au niveau du code barre au lieu de se ramener à quelques dizaines de catégories de produits ; la raison en est que la composition détaillée des aliments ("pizza ultra-processée Margherita ZX318" et non "pizza") a un impact sur la santé.

Les résultats précédents, abordant les données de nutrition selon une approche NLP (chaque ménage est vu comme un "document" composés de "mots"/produits alimentaires) permettent de caractériser certains topics (e.g. régime breton-normand; régime alsacien; régime bio). Ces régimes sont corrélés avec certains indicateurs de santé : l'objectif de ce stage est de déterminer s'il y a **causalité**. Ce stage peut déboucher sur une thèse.

Approche proposée. Idéalement, les relations causales sont établies en se fondant sur des essais randomisés contrôlés (ERC: on choisit deux populations identiques à tous égards, dont l'une reçoit un médicament et l'autre non; et on compare les indicateurs de survie des deux populations pour établir le bénéfice du médicament). Cependant, les ERC sont souvent impossibles à mettre en oeuvre pour des raisons de coût, de faisabilité ou d'éthique. En matière d'alimentation, les expériences de régime dit occidental (gras et sucre) sont faites essentiellement sur des souris.

La recherche proposée s'intéressera à la mise au point de procédure de type ERC à partir de données observationnelles. Par exemple, étant donné un échantillon de personnes suivant le régime R en année N, on cherchera (par exemple dans la base de données de l'année $N+1$) un échantillon de mêmes caractéristiques socio-démographiques, pour évaluer comparativement l'impact du régime R. L'une des difficultés vient de la présence des confondeurs ; certains sont connus et présents dans les données (niveau d'études) ; d'autres ne sont pas documentés (aucune information sur le sport ou le fait de fumer). On pourra ainsi chercher à reconstruire des confondeurs, en analysant la distribution des données [1]. Une autre option consiste à suivre les cohortes (les données de consommation de plusieurs années seront disponibles), et à voir l'évolution de leur état de santé.

Ce stage demande de très bonnes capacités statistiques et informatiques (programmation C++ ou Python, en partant de librairies existantes), ainsi que du goût pour la pluridisciplinarité (maths-informatique, biologie, économie, sociologie).

Références [1] The blessing of multiple causes. Yixin Wang, David M. Blei, Arxiv 2018.
[2] Jonas Peters, Dominik Janzing, Bernhard Schölkopf: Elements of Causal Inference: Foundations and Learning Algorithms, MIT Press, 2018.