Master Internship 2021-2022

Spatio-temporal causal learning: Health outcomes of residual pesticides

Supervisors: Michèle Sebag, Olivier Allais, Julia Mink
sebag@lri.fr, olivier.allais@inrae.fr, julia.mink@sciencespo.fr
Lab: TAU − CNRS − INRIA − LISN, U. Paris-Saclay & ALISS − INRAE

Abstract:
Pesticide impacts due to their professional use (e.g., on farmers) are increasingly well identified. Their impacts on other people have been much less studied. The Horapest study aims to investigate the causal relationships [1,2] between the diffusion of pesticides (depending on the crops, the period of the year, the meteo) and the health-related events in the population of newborns and their mothers in France, during the 2018-2019 year. While causal relationships are harder to discover and assess than correlations, they are required to provide a sound basis for interventions [3].

The evidence will include: i) the medical events, as available from the Health data hub, reporting all medical visits, drug prescriptions and hospitalizations for circa 98% of the French population; ii) the concentration in the air of 75 substances involved in crop protection products, in the vicinity of 50 stations, reported on a weekly basis (CNEP study); iii) the amount of products sold, reported every year by distributors at the zip code level.

The challenge is manifold. Firstly, the identification of causal relationships is adversely affected by the presence of hidden confounders [3,4], which are pervasive in real-world studies. For instance in the Horapest context, the educational background of the household members might have an impact on their home location and its distance to fields. The socio-economic category of the head of household has an impact on the health of the household members. On the other hand, the impact of the crop protection products might be hard to assess (e.g. masked by alimentary products; blurred by the fact that the same molecules are present in different products).

The internship :
The intern will participate in the Horapest pilot study, focusing on a few candidate zip codes selected for presenting diverse types of risks (proximity to specific crops; other sources of pollution; proximity to CNEP stations).

The methodology will firstly leverage the main approaches in causal discovery, based on linear modeling and differences-in-differences (among households with high/low risk; among newborns exposed *in utero*/not exposed to given molecules). Significant extensions will be needed to account for hidden confounders, taking inspiration from the Deconfounder approach [4]. Generative models based on independent sources might be used [5] to consider a low-dimensional search space and guide the search.

An option is to consider an augmented representation of the households (e.g. enriched through societal average indicators available from INSEE) to reflect the societal confounders. Independently, latent compressed representations will be considered to reflect and compress the medical evidences and handle their diversity.

The internship requires excellent creativity, as well as theoretical and algorithmic skills (programming environment in Python).

Références

[1] J. Pearl. Causality: models, reasoning, and inference. Econometric Theory, 19(675-685):46, 2003.
[2] J. Pearl, D. Mackenzie. The book of why: the new science of cause and effect, 2018

[3] J. Peters, D. Janzing, B. Schölkopf. Elements of Causal Inference, Foundations and Learning Algorithms. MIT Press, 2017.*Recommender Systems*, Aggarwal, Springer 2016.
[4] Y. Wang, D.M. Blei. The blessings of multiple causes. Journal of the American Statistical Association, 114(528), p. 1574-1596, 2019 [5] J. Chen et al., Supercharging Imbalanced Data Learning With Energy-based Contrastive Representation Transfer, NeurIPS 2021.