Développement d'une stratégie d'analyse statistique multibloc pour l'intégration de données exposomiques et métabolomiques au service de l'étude du lien entre exposition chimique et santé humaine.

Doctorant : Étienne BABIN

Oniris, INRAE, LABERCA, 44300 Nantes, France.

Contexte

Tout commence avec l'Exposome. Un concept assez récent proposé par Christopher Paul Wild en 2005 (Wild 2005). Que nous le voulions ou non nous sommes exposé.e.s à toutes sortes de choses. Des polluants et autre produits chimiques, dans l'air, l'eau, les aliments… Mais pas seulement. Vous vivez peut-être dans une belle maison avec un grand jardin ou dans une petite passoire thermique toute humide. Vous êtes peut-être du genre sport, eau, salade verte ou plutôt canapé, sodas, fast-food. Travail chille ou burn-out en approche. Tout cela a un impact sur vous, sur votre risque de développer des pathologies. C'est cela l'Exposome, l'ensemble des expositions que nous subissons au cours de notre vie.

Il est donc intéressant de se pencher, sans tomber dans la paranoïa, sur les effets que ces expositions ont sur la santé. Mais étudier expositions et santé ne suffit pas, il y a un intermédiaire. Oui, tout le monde a son petit intérieur propre, personne n'est fait pareil, et cela influe sur nos réactions aux expositions. Ce bloc de l'organisme, dit « omique » (car il rassemble les molécules de groupes en « omique », épigénomique, transcriptomique, protéomique, métabolomique…), peut donc nous donner des informations sur la façon dont les expositions agissent sur la santé, les voies d'action biologique.

Mais comment prendre en compte ces 3 blocs, exposition, organisme et santé ? C'est ici que j'interviens, moi un petit statisticien, venant des mathématiques fondamentales, je tente, aidé de mes encadrant.e.s, de trouver les meilleurs outils pour comprendre les voies d'action de l'exposition sur la santé humaine par l'intermédiaire de notre organisme. Pour cela, je m'intéresse en particulier à ce que l'on appelle des modèles multibloc. Comme leur nom l'indique plutôt bien, ils servent à modéliser les données regroupées en plusieurs blocs, en plusieurs groupes de variables.

Étude des méthodologies classiquement employées

L'idée de la première partie de la thèse était de faire l'état de l'art des méthodes statistiques utilisées dans le cadre de nos 3 blocs susnommés. Nous avons procédé à une bibliographie méticuleuse, écumants Pubmed, armés de critères précis dont je vous épargne le détail, pour rassembler les articles traitant de notre sujet d'intérêt et observer quelles analyses statistiques été effectuées.

Nous avons ressorti de cette revue (je ne résiste pas à une petite auto citation (Babin et al. 2023)) que la plupart des articles utilisaient des méthodes ne prenant pas en compte la structure en bloc et traitant les variables individuellement, voir en les considérants comme indépendantes.

Évaluation de modèles

Pour considérer cette structure en bloc nous avons étudié des modèles multibloc basés sur la Partial Least Squares (PLS). Son principe est de résumer un grand nombre de variables en quelques variables synthétiques permettant à la fois de capter le maximum d'information dans les blocs dit explicatifs, dans notre cas ceux d'exposition et omique, mais aussi donnant le maximum d'informations sur la ou les variables à expliquer, dans notre cas caractérisant la santé. Nous avons comparé la PLS à sa version multibloc la plus classique, la Multiblock Partial Least Squares (MB-PLS) et à un modèle fonctionnant davantage par étape sur les différents blocs explicatifs afin de récupérer des informations complémentaires, la Sequential and Orthogonalized Partial Least Squares (SO-PLS). Nous avons également étudié des versions dîtes sparse de ces modèles, permettant d'obtenir des coefficients d'importance des variables un petit peu différent apportant entre autre une sélection de variables plus stricte.

Nous avons utilisé comme données pour nos évaluations, à la fois des données réelles provenant d'un article publié (Aung et al. 2020), et des données simulées par nos soins afin de maîtriser la structure des données et ainsi voir si nos modèles la retrouve.

Conclusion

Après 3 années de thèse, nous avons pu mettre en évidence l'intérêt de l'utilisation de modèles basés sur la PLS pour l'étude de données d'épidémiologie environnementale. Bien que nos résultats doivent être confirmés et affinés davantage, et que chaque méthode ai des avantages et inconvénients selon la situation, la plus grande utilisation de ce type de modèles dans ce domaine est une recommandation que nous pouvons faire.

Aung MT, Song Y, Ferguson KK, Cantonwine DE, Zeng L, McElrath TF, et al. 2020. Application of an analytical framework for multivariate mediation analysis of environmental data. Nat Commun 11:5624.
Babin É, Cano-Sancho G, Vigneau E, Antignac JP. 2023. A review of statistical strategies to integrate biomarkers of chemical exposure with biomarkers of effect applied in omic-scale environmental epidemiology. Environmental pollution (Barking, Essex : 1987) 330:121741.
Wild CP. 2005. Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 14:1847-1850.