

PROPOSITION DE SUJET DE THESE TH SAN26-13

RÉFÉRENCE DE PRÉ-ENREGISTREMENT à compléter :

Th SAN26-13**SUJET DE THESE**

Intégration de données multimodales avec des blocs de données manquantes dans l'étude des effets des rayonnements ionisants à faible dose

Intitulé en anglais (*optionnel mais recommandé*) :

Multi-modal data integration with block-wise missing data in the context of low-dose ionizing radiation studies

Lien avec les axes programmes

Ce projet s'inscrit dans l'Axe Programme Santé Environnement.

Contexte de la thèse et objectif en matière d'avancée(s) des connaissances Etat de l'art (références bibliographiques) et objectif de la thèse en matière d'avancée des connaissances

Malgré l'abondance croissante de données disponibles et l'émergence du Big Data au sein des laboratoires de biologie expérimentale (RNA-seq, métabolomique, lipidomique...), la gestion des données manquantes demeure un défi majeur pour exploiter pleinement l'information disponible. Pour relever ce défi, des approches méthodologiques spécifiques sont nécessaires pour intégrer différentes modalités de données et identifier les mécanismes d'action impliqués dans la réponse d'un organisme à un stress.

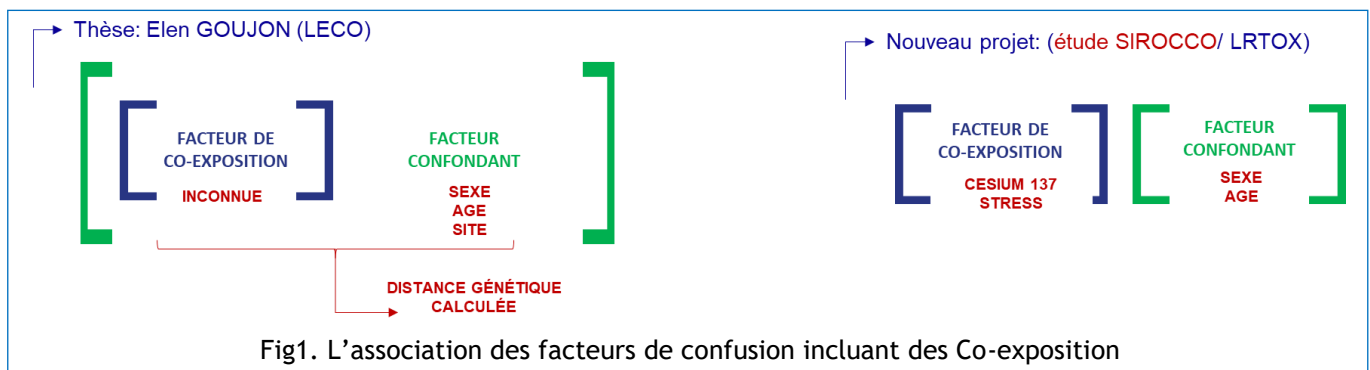
Eliminer les données manquantes, peut non seulement réduire la précision et la puissance statistique de l'étude, mais aussi introduire des biais significatifs dans les modèles d'analyse intégrative, conduisant finalement à des résultats erronés [1]. Ces problématiques sont étroitement liées aux caractéristiques intrinsèques des données générées et aux méthodes d'intégration utilisées pour traiter ces ensembles de données volumineux, souvent impactés par un nombre important de données manquantes. Par ailleurs, étant observées à différentes échelles biologiques, ces mêmes données présentent des niveaux de fluctuation pouvant être induits par des facteurs confondants et/ou des co-expositions, qui nécessitent l'application de facteurs correctifs dans l'analyse des données brutes.

Ce projet de thèse repose sur les données omiques d'une étude examinant les effets d'une exposition à de faibles doses de rayonnements ionisants couplée au facteur de risque du stress psychique sur la pathologie athéromateuse et ses conséquences vasculaires.

Les données d'analyse ayant déjà été produites pour différentes modalités, les deux premières années de la thèse seront dédiées au développement d'une méthodologie d'analyse intégrative (multimodale) prenant en compte les facteurs de co-exposition dans l'imputation des données. La dernière année sera consacrée à l'analyse des données et à l'interprétation des résultats à l'aide de méthodes bio-informatiques d'enrichissement des voies biologiques.

Ce projet permet de poursuivre l'axe méthodologique de la thèse d'Elen GOUJON sur les facteurs confondants [2] (projet MORDOR) en abordant la problématique de co-exposition dans un contexte différent comme illustré dans la figure (Fig 1). En effet, le projet MORDOR portait sur un cas d'étude accidentel Tchernobyl où les scénarios et la nature de la co-exposition n'étaient pas ou très partiellement connus (une matrice de distance génétique a été proposée afin de pallier ces limitations). Les données expérimentales qui seront analysées dans le présent projet ont en revanche l'avantage d'être issues d'une étude expérimentale menée au laboratoire LRTOX où les facteurs confondants et le type de co-exposition sont connus et maîtrisés. A terme, les résultats de ces deux thèses permettront la prise en compte de facteurs confondants dans l'exploitation de l'information biologique des données dans différents contextes d'acquisition des données à faibles doses d'exposition.

PROPOSITION DE SUJET DE THESE TH SAN26-13



Contenu de la thèse

Selon l'Organisation Mondiale de la Santé, les maladies cardiovasculaires (MCV) sont responsables de la mort de près de 17,9 millions de personnes dans le monde chaque année [3]. Les résultats des études épidémiologiques suggèrent une association entre les expositions aux rayons ionisants (RI) et les MCV (tels que l'athérosclérose et les accidents vasculaires cérébraux) dans des gammes de doses fortes à modérées (au-dessus de 500 mGy) [4]. Cependant, les connaissances des MCV dans les gammes de doses plus faibles restent largement débattues par manque de compréhension des effets des RI à faibles doses ainsi que la puissance statistique limitée pour obtenir des résultats robustes.

Le stress est un des premiers facteurs de risque cardiovasculaire. En effet, un stress psychique est capable d'activer d'une part, le système nerveux sympathique qui libère de l'adrénaline et de la noradrénaline, de l'autre, l'axe hypothalamo-hypophysaire - surrénale qui aboutit à la sécrétion du cortisol. L'activation de ces deux voies va générer des espèces réactives de l'oxygène et de l'inflammation au niveau des artères en activant les cellules immunitaires (globules blancs) qui vont générer des radicaux libres oxygénés (molécules pro-oxydantes) et des cytokines pro-inflammatoires (molécules aux propriétés inflammatoires). L'activation des voies inflammatoires et de stress oxydants va entraîner à terme des lésions vasculaires qui en cas de complications à plus ou moins long terme va aboutir aux symptômes ischémiques ou hémorragique (AVC, infarctus du myocarde ...) [5,6].

Aucune étude n'a cependant mis en évidence la réponse d'une co-exposition associant les rayonnements ionisants et un cofacteur de risque tel que le stress sur la pathologie athéromateuse et les conséquences sur la vascularisation. Pour cela une étude expérimentale a été effectuée au sein du LRTOX dans le cadre du projet post-Doc « SIROCCO ». Pour faire progresser la compréhension des conséquences de cette exposition faible, ce projet de thèse visait à développer des méthodes d'intégration de données multimodales afin de mieux décrypter les mécanismes sous-jacents impliqués dans la réponse multifactorielle d'un organisme vivant exposé à un ou plusieurs expositions (ici, radiations et stress odorant).

Dans ce contexte, l'emploi d'analyses biologiques omiques à larges spectres et à haut débit (telles que transcriptomique, le métabolomique ou le lipidomique) et leur intégration avec d'autre type de données (imagerie, physiologie...etc) nécessite des développements méthodologiques en biostatistique et bio-informatique pour caractériser l'impact potentiel des faibles doses de RI sur MCV et comprendre l'ensemble des mécanismes impliqués dans la réponse biologique d'un organisme vivant.

L'analyse intégrative de tels jeux de données, aussi appelés données multiblocs [7], tire parti de la complexité et de la richesse de l'information contenue par chaque bloc(modalité), ainsi que de la structure inhérente des blocs.

Plus précisément, ces approches reposent sur des méthodes de réduction conjointe de la dimensionnalité qui permettent d'apprendre des représentations de faible rang, appelées composantes latentes ou facteurs, qui résument l'information redondante à travers des motifs plus simple à interpréter [8]. Le cadre statistique proposé par le RGCCA [9,10], développée par notre partenaire de Centrale Supélec, permet de mener de telles analyses multiblocs. Cette méthode flexible et généraliste cherche des composantes latentes capturant les directions de variation communes entre les blocs.

Cependant, les données multiblocs ont souvent une structure manquante, c'est-à-dire que les données d'un ou plusieurs blocs peuvent être complètement inobservées pour un échantillon. Afin d'aborder correctement l'imputation des données manquantes il faut en distinguer les causes, surtout si elles ne sont pas le simple fruit du hasard. Les différents types de données manquantes peuvent être classés en trois catégories [11] : MCAR (missing completely at random), MAR (Missing at random) et MNAR (Missing not at random). De nombreuses méthodes ont été proposées dans la littérature pour prendre en compte ces données manquantes comme les

PROPOSITION DE SUJET DE THESE TH SAN26-13

approches basées sur les forêts aléatoires (CART et XGBoost) qui considèrent des surrogate splits ou divisions de substitution [12] ou L'algorithme NIPALS [13] de qui permet de gérer les données manquantes sans les imputer dans un formalisme de la régression PLS. Mentionnons également les méthodes d'imputation reposants sur des décompositions en valeurs singulières (SVD), inférence bayésienne ou imputation multiple (tel que missMDA [14,15]).

D'une façon générale, la performance de ces méthodes dépend du pourcentage des données manquantes se situant entre 20 % et 30 % pour des résultats fiables. Or, ces pourcentages de données manquantes sont généralement beaucoup plus importants dans le cadre des données multimodales où des blocs entiers peuvent être manquants pour la plupart des individus ou échantillons. Ainsi, peu de méthodes permettent à ce jour de gérer les données manquantes de manière intégrée dans une analyse multiblocs.

L'objectif de ce travail de thèse est de proposer des développements méthodologiques pour traiter correctement les structures de données manquantes par bloc en proposant une nouvelle version (dite globale) du modèle RGCCA [16,17].

Plus précisément, cette nouvelle approche permettra d'extraire « simultanément », en présence de données manquantes en blocs, les composantes latentes de chaque bloc contrairement à la version « séquentielle » de l'actuel RGCCA où les biais liés aux données manquantes se propagent d'une composante à une autre [18,19,20].

Ainsi, les résultats de cette thèse contribueront à une meilleure identification des facteurs de risque des MCV en explorant ses mécanismes induit par l'effet des doses faibles d'irradiation combiné à d'autres stressseurs.

Déroulement de la thèse et résultats attendus

- Année 1 : Développement méthodologique pour la gestion des blocs des données manquante

Objectif : Etablir le modèle statistique et la validation du modèle avec les données simulées.

- Etude bibliographique
- Formalisation mathématique du modèle et son implémentation
- Vérification et validation des résultats obtenus par simulation
- Comparaison des performances avec les approches méthodologiques existantes

- Année 2 : Développement de l'intégration multimodale ; adaptation du modèle aux facteurs confondants et co-expositions. Dans le domaine des données structurées nous pouvons analyser nos données sur plusieurs facettes méthodologiques tel que le multibloc (différents modalités), le multigroupe (la prise en compte des différents co-facteurs) et le multiway (images observées sous trois dimensions).

Objectif : Incorporer les nouvelles fonctionnalités méthodologiques dans l'environnement RGCCA.

- Vérification de l'intégration cohérente des deux modèles
- Rédaction d'un article méthodologique dédié à la gestion des données manquantes

- Année 3 : A la suite du développement méthodologique, le doctorant poursuivra par une phase d'analyse des données et interprétation biologique.

Objectif : Réaliser une étude sur les effets des rayonnements ionisant en présence de co-exposition (Exposition aux rayonnement ionisants et le stress) et facteurs confondants (sexe et âge) : Mise en œuvre sur les maladies cardiovasculaires (MCV) : analyse, intégration des résultats et valorisation scientifique

- Analyses complémentaires et écriture du deuxième papier de thèse
- Rédaction du manuscrit de thèse et soutenance.

Le doctorant travaillera au sein du laboratoire LRAcc. Le travail interdisciplinaire de la thèse sera articulé autour de réunions régulières avec Arthur TENENHAUSE (Centrale-Supelec) pour la partie méthodologique et avec EBRAHIMIAN CHIUSA Teni et GRISON Stéphane (PSE-Santé/LRTOX) pour l'interprétation des résultats biologiques.

PROPOSITION DE SUJET DE THESE TH SAN26-13

Jalons et livrables

1er année	2ème année	3ème année
- Etude bibliographique pour les données manquantes dans le cas des analyses multi-tableaux -Développement méthodologique pour l'imputation des données	-1er papier pour le développement méthodologique de l'imputation des données multi-tableaux -Participation à des conférences -Analyse et adaptation du modèle méthodologique pour la gestion des facteurs confondants et les facteurs de co-exposition	-2ème papier pour l'intégration multimodale pour étudier les effets des facteurs de co-exposition sur les maladies cardio-vasculaires - Participation à des conférences - Rédaction du rapport de thèse - Soutenance de la thèse

Partenariats envisagés

Collaboration avec TENENHAUS Arthur (L2S/CentraleSupélec), EBRAHIMIAN CHIUSA Teni (LRTOX/ ASNR) et GRISON Stéphane (LRTOX/ ASNR)

Apports respectifs de l'ASNR et des partenaires – Connaissances propres Savoir-faire et compétences mis en œuvre par chacune des parties

Le doctorant bénéficiera également de l'expertise du L2S/CentraleSupélec, qui possède une expertise considérable dans les méthodes statistiques/ML pour l'intégration des données.

Bibliographie

- [1] Alwateer, M., Atlam, E.-S., Abd El-Raouf, M.M., Ghoneim, O.A. and Gad, I. (2024) Missing Data Imputation: A Comprehensive Review. *Journal of Computer and Communications*, 12, 53-75.
<https://doi.org/10.4236/jcc.2024.1211004>
- [2] Elen Goujon, Olivier Armant, Clément Car, Jean-Marc Bonzom, Arthur Tenenhaus, and Imène Garali. Batch Effect Correction in a Confounded Scenario: a Case Study on Gene Expression of Chornobyl Tree Frogs. In Roberta Gori, Paolo Milazzo, and Mirco Tribastone, editors, *Computational Methods in Systems Biology*, pages 89-107, Cham, 2024. Springer Nature Switzerland.
- [3] Organisation mondiale de la santé. Maladies cardiovasculaires (MCV). Fiche d'information, 31 juillet 2025 (consultée le 5 novembre 2025). Disponible sur : [https://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
- [4] Little MP, Azizova TV, Bazyka D, Bouffler SD, Cardis E, Chekin S, Chumak VV, Cucinotta FA, de Vathaire F, Hall P, Harrison JD, Hildebrandt G, Ivanov V, Kashcheev VV, Klymenko SV, Kreuzer M, Laurent O, Ozasa K, Schneider T, Tapio S, Taylor AM, Tzoulaki I, Vandoolaeghe WL, Wakeford R, Zablotska LB, Zhang W, Lipshultz SE. Systematic review and meta-analysis of circulatory disease from exposure to low-level ionizing radiation and estimates of potential population mortality risks. *Environ Health Perspect.* 2012 Nov;120(11):1503-11. doi: 10.1289/ehp.1204982. Epub 2012 Jun 22. PMID: 22728254; PMCID: PMC3556625.
- [5] S Seddon, A Shantsila, I Lip, "Psychological stress and heart disease: fact or folklore?", *The American Journal of Medicine*, 135(2):138-45, 2022. *American Journal of Medicine*
- [6] A Xie, Y Zhang, "The role of oxidative stress in cardiovascular disease caused by psychosocial stress", *Journal of Cardiovascular Translational Research*, 13(2):113-126, 2020
- [7] Hotelling, Harold (1936). « Relations Between Two Sets Of Variates ».In: *Biometrika* 28.3-4, pp. 321-377. doi: 10.1093/biomet/28.3-4.321

PROPOSITION DE SUJET DE THESE TH SAN26-13

- [8] Jolliffe, Ian T.; Cadima, Jorge (2016-04-13). "Principal component analysis: a review and recent developments". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 374 (2016) 20150202.
- [9] Tenenhaus A., Tenenhaus M., (2014). Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *European Journal of Operational Research*, 238(2), 391-403.
- [10] Garali I., M. Adanyeguh I., Ichou F., Perlberg V., Seyer A., Colsch B., Moszer I., Guillemot V., Durr A., Mochel F., Tenenhaus A. (2018) A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia. *Briefings Bioinform.* 19(6): 1356-1369
- [11] Little R.J.A. et Rubin D.B., *Statistical Analysis with Missing Data*, Wiley series in probability and statistics, 1987.
- [12] A. J. Feelders, *Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation.*, PKDD, Lecture Notes in Computer Science, t. 1704, Springer, 1999, p. 329-334.
- [13] Preda C., Saporta G. et Hedi Ben Hadj Mbarek M., *The NIPALS algorithm for missing functional data*, *Romanian Journal of Pure and Applied Mathematics* 55 (2010), no 4, 315-326.
- [14] Josse J. & Husson F. (2016). *missMDA: a package for handling missing values in multivariate data analysis*. *Journal of Statistical Software*, 70(1), 1-31.
- [15] Rubin D.B., *Multiple Imputation for Nonresponse in Surveys*, Wiley, 1987.
- [16] Gloaguen, Arnaud (2020). « A Statistical and Computational Framework for Multiblock and Multiway Data Analysis ». PhD thesis. Université Paris-Saclay.
- [17] Girka, Fabien (2023). « Development of New Statistical/ML Methods for Identifying Multimodal Factors Related to the Evolution of Multiple Sclerosis ». PhD thesis. Université Paris-Saclay.
- [18] Peltier C, Le Brusquet L, Lejeune FX, Moszer I, Tenenhaus A (2022). "Missing Values in RGCCA: Algorithms and Comparisons." In 8th International Conference on Partial Least Squares Structural Equation Modeling (PLS'22).
- [19] Baena-Miret S, Reverter F, Vegas E. A framework for block-wise missing data in multi-omics. *PLoS One*. 2024 Jul 23;19(7):e0307482. doi: 10.1371/journal.pone.0307482. PMID: 39042603; PMCID: PMC11265675.
- [20] Tu T. Do, Mai Anh Vu, Tuan L. Vo, Hoang Thien Ly, Thu Nguyen, Steven A. Hicks, Michael A. Riegler, Pål Halvorsen, Binh T. Nguyen. "Blockwise Principal Component Analysis for monotone missing data imputation and dimensionality reduction" doi: <https://doi.org/10.48550/arXiv.2305.06042>