Prédiction du besoin de réanimation à partir de l'enregistrement de constantes respiratoires

---

Mots-clé : séries temporelles, données manquantes, intégration de données multi-sources, hétérogènes, covid19

Contexte. L'un des défis majeurs de la crise sanitaire provoquée par l'épidémie de COVID-19 est la disponibilité de lits de réanimation et d'une ventilation mécanique invasive (VMI). Le nombre limité de lits et de ventilateurs exige une utilisation stricte de ces ressources rares. Étant donné qu'une proportion importante de patients initialement admis dans le service pour la COVID-19 nécessiteront une VMI, il est essentiel de prévoir ces besoins à l'échelle individuelle pour aider à gérer la pénurie de lits et de ventilateurs dans les unités de soins intensifs [1,2].

Notre objectif est d'établir et de valider un score pour chaque patient, reposant sur les données de surveillance continue de constantes comme la fréquence respiratoire, oxygène pulsé saturation, rythme cardiaque, tension artérielle, etc. mais aussi sur des données cliniques, qui permettra d'identifier 48 heures à l'avance les patients qui devront être transférés en soins intensifs avec ventilateur.

Pour relever ce défi, nous devons donc agréger des sources de données hétérogènes (séries temporelles, et données numériques ou catégorielles) pour établir des modèles prédictifs. L'agrégation de ces données multiples favorise en général les données manquantes : celles-ci peuvent même être très structurées et ne concerner qu'exclusivement les données tabulaires (quand dans des situations d'urgence, certaines mesures sur le patient ne peuvent être faites) ou les données temporelles (si par exemple le moniteur est retiré pour des raisons d'hygiène quotidienne, ou de situation d'urgence de l'état du patient). L'innovation est donc ici plurielle : elle réside dans le fait de développer une méthodologie dédiée aux données manquantes dans le cadre de séries temporelles, puis de gérer des données de type hétérogène (tabulaire et chronologique), et enfin de développer une méthodologie gérant les données manquantes structurées par type de données.

Plan de travail. Dans un premier temps, nous développerons une méthodologie statistique pour traiter les valeurs manquantes dans les données temporelles. Nous nous attacherons à établir les garanties théoriques nécessaires à valider la méthode. Peu de travaux théoriques ont été entrepris dans ce domaine, les développements méthodologiques sont très spécifiques aux domaines d'application, e.g. [3,4] et concernent souvent des séries temporelles univariées [5,6]. Pour le cas multivarié, des approches récentes basées sur des approximations en rang inférieur [7] ou des gan [8, et références à l'intérieur] ont été proposées. Après un état de l'art approfondi [9, 10] et en particulier [8] et la comparaison des approches disponibles, nous explorerons différentes approches. Une idée consiste à adapter les algorithmes funFEM [11] initialement

prévus pour le clustering de données fonctionnelles à la complétion de séries temporelles. Une autre est d'utiliser des méthodes à noyau (de type SVM ou Gaussian Processes), combinées à des signature kernels [12, 13]. Ces méthodes permettent de conserver la forme fonctionnelle des données.
Puis, dans un second temps, nous étendrons la méthode pour gérer les données hétérogènes, de données cliniques et temporelles, présentant potentiellement des données manquantes structurées. Le type de données manquantes généralement considéré dans la littérature peut se classer en 3 catégories (MCAR, MAR, MNAR) [14] ; les données manquantes structurées par bloc et type (clinique ou temporel) représentent un nouveau champ de l'analyse statistique des données manquantes, qui jusqu'ici n'a été ni étudié, ni traité. En pratique, les utilisateurs considèrent que ce cas extrême de données manquantes ne peut pas être traité et suppriment les observations ce qui conduit à une perte considérable d'information. Néanmoins, la prise en compte des relations entres variables de différents types et des relations entre observations peut permettre d'éviter cette suppression. Cette contribution fondamentale sera illustrée et appliquée sur les données de besoin de ventilation mécanique.

Résultats attendus. Les résultats prendront la forme de publications (en apprentissage statistique pour les contributions fondamentales, et dans le domaine médical pour les études réalisées) et de briques logicielles en licence libre. Il y a deux parties dans le sujet, un travail plus méthodologique, pouvant mener à des considérations théoriques, et un travail très appliqué sur les données. Les deux directions sont importantes mais selon l'étudiant.e, l'accent pourra être mis sur un des aspects.

Composition de l'équipe & collaborations.
Équipe: Julie Josse (Inria), Antoine Liutkus (Inria), Claire Boyer (LPSM, Sorbonne Université), Pierre Lafaye de Micheaux (Université de Montpellier).
Ce projet est issu d'une collaboration entre Inria et le CHRU de Nancy représenté par le PUPH réanimateur Antoine Kimmoun. Les données sont stockées à l'Inria et continuent d'arriver au cours de cette deuxième vague. Le travail sera donc interdisciplinaire et l'équipe statistique/ML/informatique travaillera en étroite collaboration avec Antoine Kimmoun.

Compétences requises.
Nous recherchons dans un premier temps un stagiaire niveau M2 en statistique/machine learning ayant un intérêt fort pour l'application médicale et l'interaction avec les médecins. Une thèse en machine learning pour la santé pourrait être envisagée par la suite. Le candidat doit déjà avoir traité des séries temporelles et des capacités de programmation importantes.
Les candidats doivent envoyer CV, relevés de notes des deux dernières années ainsi que le nom d'une personne référente à julie.josse@inria.fr, copie à antoine.liptkus@inria.fr et claire.boyer@sorbonne-universite.fr

[1] R.D. Truog, et al. The Toughest Triage — Allocating Ventilators in a Pandemic. *The New England Journal of Medicine 2020*
[2] E.L. Biddision et al. Too Many Patients…A Framework to Guide Statewide Allocation of Scarce Mechanical Ventilation During Disasters. *Chest 2019*
[3] W. Velicer, et al. A Comparison of Missing-Data Procedures for Arima Time-Series Analysis. *Educational and Psychological Measurement 2005*
[4] W. Junger, et al.. Imputation of missing data in time series for air pollutants. *Atmospheric Environment 2015*
[5] W. Dunsmuir et al. Estimation of Time Series Models in the Presence of Missing Data. *JASA 1981*
[6] https://steffenmoritz.github.io/imputeTS/
[7] P. Alquier, et al. Matrix factorization for multivariate time series analysis. Electronic journal of statistics, 2019
[8] Jarrett et al. Clairvoyance: A Pipeline Toolkit for Medical Time Series. ICLR2021. (https://www.youtube.com/watch?v=NJ700fg-0YM&feature=youtu.be)
[9] Stephanie Clark, Rob J Hyndman, Dan Pagendam, Louise M Ryan (2020) Modern strategies for time series regression. International Statistical Review, to appear.
[10] Shanika L Wickramasuriya, George Athanasopoulos, Rob J Hyndman (2019) Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. J. American Statistical Association, 114(526), 804-819.
[11] C. Bouveyron, et al. funFEM: an R package for functional data clustering. *CRAN*
[12] Király, F. J., & Oberhauser, H. Kernels for sequentially ordered data. Journal of Machine Learning Research, (20). 2019.
[13] Toth, C., & Oberhauser, H. Bayesian learning from sequential data using gaussian processes with signature covariances. *(ICML 2020).*
[14] R. Little, D. Rubin. Statistical analysis with missing data. *2019*