

Master 2 internship proposal

February to July 2024 (4 to 6 months)

Internship topic: Development of a machine learning model for annual predictive monitoring of airborne micro-organisms in French Guiana.

Summary :

The Airfungui project aims to establish epidemiological surveillance of risk factors for respiratory diseases and pollution episodes, using innovative technologies such as environmental DNA metabarcoding and artificial intelligence. The result is the determination of the potential influence of physico-chemical and meteorological factors on aerial microbial biodiversity in French Guiana. UMR Espace-Dev and UMR-CIIL-TBIP will be supervising a master's student to develop a machine learning model for predicting the taxonomy of micro-organisms present in the air, based on physico-chemical, meteorological and sequencing data obtained at two sampling sites in French Guiana. The sequencing data is used to train and validate the model.

Tasks :

- Data pre-processing: this involves cleaning up abiotic parameters, generating missing values and extracting significant characteristics from the data.
- Selecting a model: using the bibliography as a starting point, select an initial machine learning model.
- Model training: this involves training the selected model using one part of the pre-processed data. The abiotic data are the characteristics and the taxonomy of the micro-organisms is the target.
- Model validation: once the model has been trained, it needs to be validated on the other part of the data. It takes the abiotic data as input and predicts the taxonomy of the micro-organisms. The aim is to assess the model's performance on these data by comparing the predicted taxonomy of micro-organisms with the actual taxonomy of micro-organisms.
- Selection and implementation of the best model: this involves applying the 3 previous tasks to other possibly relevant models and then selecting the best performing model for the purposes of the study.
- Deploying the model: once the model has been validated, it can be deployed in a real environment.

Bibliography :

- Cordero, J. M., Núñez, A., García, A. M., & Borge, R. (2021). Assessment and statistical modelling of airborne microorganisms in Madrid. *Environmental Pollution*, 269, 116124.
- Liu, Z., Li, H., & Cao, G. (2017). Quick estimation model for the concentration of indoor airborne culturable bacteria: an application of machine learning. *International journal of environmental research and public health*, 14(8), 857.
- Liu, Z., Cheng, K., Li, H. et al. Exploring the potential relationship between indoor air quality and the concentration of airborne culturable fungi: a combined experimental and neural network modeling study. *Environ Sci Pollut Res* 25, 3510–3517 (2018). <https://doi.org/10.1007/s11356-017-0708-5>
- Meno, L.; Escuredo, O.; Abuley, I.K.; Seijo, M.C. Importance of Meteorological Parameters and Airborne Conidia to Predict Risk of Alternaria on a Potato Crop Ambient Using Machine Learning Algorithms. *Sensors* 2022, 22, 7063. <https://doi.org/10.3390/s22187063>
- Morera, A., Martínez de Aragón, J., Bonet, J.A. et al. Performance of statistical and machine learning-based methods for predicting biogeographical patterns of fungal productivity in forest ecosystems. *For. Ecosyst.* 8, 21 (2021). <https://doi.org/10.1186/s40663-021-00297-w>

Proposition de stage de Master 2

De février à juillet 2024 (4 à 6 mois)

Sujet du stage : Développement d'un modèle d'apprentissage automatique pour une surveillance prédictive annuelle des micro-organismes aéroportés en Guyane française.

Résumé :

Le projet Airfungui vise à établir une surveillance épidémiologique des facteurs de risque des maladies respiratoires et les épisodes de pollution, à l'aide de technologies innovantes telles que le métabarcoding d'ADN environnemental et l'intelligence artificielle. Cela se traduit par la détermination de l'influence potentielle des facteurs physico-chimiques et météorologiques sur la biodiversité microbienne aérienne en Guyane Française. L'UMR Espace-Dev et l'UMR-CIIL-TBIP encadreront un étudiant en master pour développer un modèle d'apprentissage automatique permettant de prédire la taxonomie des micro-organismes présents dans l'air, à partir des données physico-chimiques, météorologiques et des données de séquençage obtenues sur les deux sites d'échantillonnage en Guyane Française. Les données de séquençage sont utilisées pour entraîner et valider le modèle.

Tâches :

- Prétraitement des données : il s'agit de nettoyer les paramètres abiotiques, de générer les valeurs manquantes et d'extraire les caractéristiques significatives des données.
- Sélection d'un modèle : en s'inspirant de la bibliographie, il convient de sélectionner un premier modèle d'apprentissage automatique.
- Entraînement du modèle : il s'agit d'entraîner le modèle sélectionné à l'aide d'une partie des données prétraitées. Les données abiotiques sont les caractéristiques et la taxonomie des micro-organismes est la cible.
- Validation du modèle : une fois le modèle entraîné, il doit être validé sur l'autre partie des données. Il prend en entrée les données abiotiques et prédit la taxonomie des micro-organismes. Il s'agit d'évaluer les performances du modèle sur ces données en comparant la taxonomie des micro-organismes prédite avec la taxonomie réelle des micro-organismes.
- Sélection et implémentation du meilleur modèle : il s'agit d'appliquer les 3 tâches précédentes à d'autres modèles possiblement pertinents et ensuite sélectionner le plus performant dans le cadre de l'étude.
- Déploiement du modèle : une fois le modèle validé, il peut être déployé dans un environnement réel.

Bibliographie :

- Cordero, J. M., Núñez, A., García, A. M., & Borge, R. (2021). Assessment and statistical modelling of airborne microorganisms in Madrid. *Environmental Pollution*, 269, 116124.
- Liu, Z., Li, H., & Cao, G. (2017). Quick estimation model for the concentration of indoor airborne culturable bacteria: an application of machine learning. *International journal of environmental research and public health*, 14(8), 857.
- Liu, Z., Cheng, K., Li, H. et al. Exploring the potential relationship between indoor air quality and the concentration of airborne culturable fungi: a combined experimental and neural network modeling study. *Environ Sci Pollut Res* 25, 3510–3517 (2018). <https://doi.org/10.1007/s11356-017-0708-5>
- Meno, L.; Escuredo, O.; Abuley, I.K.; Seijo, M.C. Importance of Meteorological Parameters and Airborne Conidia to Predict Risk of Alternaria on a Potato Crop Ambient Using Machine Learning Algorithms. *Sensors* 2022, 22, 7063. <https://doi.org/10.3390/s22187063>
- Morera, A., Martínez de Aragón, J., Bonet, J.A. et al. Performance of statistical and machine learning-based methods for predicting biogeographical patterns of fungal productivity in forest ecosystems. *For. Ecosyst.* 8, 21 (2021). <https://doi.org/10.1186/s40663-021-00297-w>