Analyse de données **2023-2024**
Eléments de modélisation statistique
4modIA

Projet d'étude

Intervenante : Cathy Maugis-Rabusseau

Organisation du projet et documents à rendre

- Le projet sera réalisé par groupe de 4 étudiant-e-s. La constitution des groupes sera faite lors de la première séance.
- 6 séances de 2h45 sont dédiées dans votre emploi du temps au travail du projet. Je serai présente lors de ces séances pour répondre à vos questions.
- Livrables : vous devrez déposer sur la page Moodle de l'UF EMS au plus tard le **vendredi 26 janvier 2024 minuit** les 2 documents suivants :
  1. un fichier Rmarkdown (*nom1-nom2-nom3-Rapport.Rmd*) contenant les codes R et générant le rapport au format pdf.
  2. un rapport au format .pdf (*nom1-nom2-nom3-Rapport.pdf*) généré par la compilation du fichier .Rmd précédent.
     Attention : le rapport est limité à 25 pages, figures incluses.
- Un dossier "ModeleRapport", disponible sur Moodle, vous donne un exemple avec des consignes pour la rédaction de votre rapport. Il est important d'en prendre connaissance dès la première séance !

Evaluation du projet

Pour chaque UF, la note de projet compte pour un tiers de la note finale de l'UF. Elle sera issue de l'évaluation des critères suivants :

Critère	**UF EMS**	**UF AD**
Maitrise de la rédaction avec Rmarkdown	X	
Rédaction générale du rapport (légende des figures, références croisées, mise en page, orthographe, ...)	X	X
Choix et rendu des graphiques illustratifs	X	X
Analyse ($\neq$ lecture !) des résultats obtenus	X	X
Programmation en R	X	X
Formalisation mathématique des questions, modèles considérés, ...	X	X
Utilisation pertinente des méthodes d'exploration de données		X
Utilisation pertinente de l'analyse discriminante linéaire		X
Utilisation pertinente de méthodes de clustering		X
Utilisation pertinente de modélisations ML et MLG	X	
Utilisation de méthodes de sélection de variables	X	
Bonus pour des choix originaux adaptés	X	X

Jeu de données étudié

Les données sont issues du site web Atmo-Occitanie [1]. On dispose de la mesure des émissions de polluants atmosphériques tous secteurs d'activités confondues des EPCI (Etablissements Publics de Coopération Intercommunale) de la région Occitanie de 2014 à 2019. On s'intéresse ici aux polluants suivants :

- nox_kg : oxyde d'azote en kg
- so2_kg : oxyde de soufre en kg
- pm10_kg : particules en suspension dans l'air de diamètre inférieur à 10 μm
- pm25_kg : particules en suspension dans l'air de diamètre inférieur à 2.5 μm
- co_kg : monoxyde de carbone
- c6h6_kg : benzène
- nh3_kg : Ammoniac
- ges_teqco2 : gaz à effet de serre
- ch4_t : méthane
- co2_t : dioxyde de carbone
- n2o_t : protoxyde d'azote

On a aussi à disposition l'année de mesure (2014 à 2019).
Pour chaque EPCI, on dispose de

- son nom (lib_epci)
- son code d'identification (code_epci)
- son (ses) département(s) d'appartenance
- sa latitude
- sa longitude
- son Type (TypeEPCI) : CC (communauté de commune), CA (communauté d'agglomération), Métrople et CU (communauté urbaine)

Questions à aborder

Dans votre rapport final, vous devez avoir abordé par une/des méthodes adaptées les questions suivantes :

- Faites une analyse descriptive des données et préparez le jeu de données pour la suite de l'étude.
  En particulier, vous justifierez vos choix de transformation potentielle des données.
- Proposez une visualisation des individus dans un espace de faible dimension à partir des émissions des polluants. Vous pouvez interpréter les résultats vis-à-vis de l'année et du type EPCI.
- Proposez une réduction de dimension à partir des émissions des polluants et du type EPCI.
- Classifiez les EPCI en fonction des émissions de polluants.
  Vous mettrez en place différentes méthodes que vous comparerez.
  Vous pouvez utiliser la librairie `ggmap` pour visualiser vos résultats sur une carte en utilisant les latitudes et longitudes.
  Vous pouvez croiser vos résultats avec les variables qualitatives décrivant les EPCI.
- Par une analyse discriminante linéaire,
  — explorer / prédire le dépassement d'émission de méthane de 1000 t par an
  — explorer / predire le type d'EPCI
- Expliquez par un modèle linéaire ou modèle linéaire généralisé :
  — le gaz à effet de serre en fonction des variables Type et années
  — le gaz à effet de serre en fonction de tous les autres polluants
  — l'émission de méthane en fonction de l'ammoniac, le protoxyde d'azote, le type d'EPCI et l'année. Vous pouvez considérer des interaction entre variables.
  — le dépassement d'émission de méthane de 1000 t par an en fonction de l'ammoniac, le protoxyde d'azote, le type d'EPCI et l'année.

  Au travers de ces questions vous devez essayer de simplifier les modèles au mieux, de mettre en place au moins une régression régularisée.

1. https ://data-atmo-occitanie.opendata.arcgis.com/search ?collection=Dataset