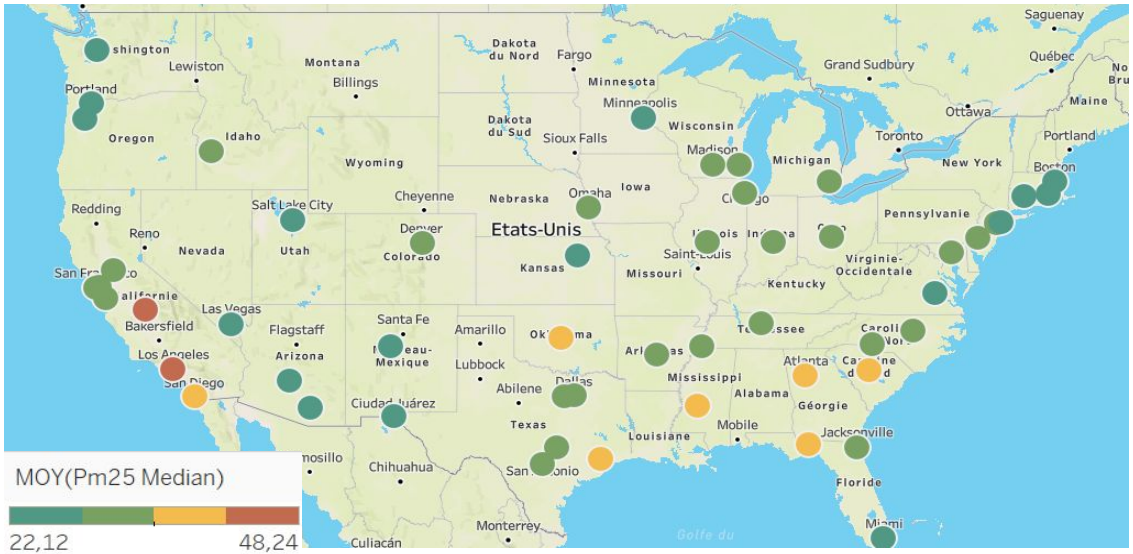

Prédire la pollution de l’air

Une forte pollution de l'air avec des conséquences clés

Concentration de PM2.5 aux USA *

Moyenne des mesures journalières période 2019-2020



< 5 MG/M3 / AN

Air bien ventilé

Concentration maximale dans l'air en moyenne annuelle (source : OMS)

> 15 MG/M3 / JOUR

Air mal ventilé

Concentration dans l'air maximale en moyenne journalière (source : OMS)

> 25 MG/M3 / JOURS

Air Vicié

Limite à ne pas dépasser plus de 3 jours par an (OMS)

> 50 MG/M3 / JOUR

Air dangereux

Limite de danger immédiat, actions correctives obligatoires (HCSP)

8.1 millions de morts par an dans le monde liés à la pollution de l'air*

*Sources: *IJCAI 2022 ** Health Effects Institute (HEI), 2021*

Expliquer et prédire: un double objectif de santé publique

Population:

Prévoir la pollution de l’air dans les prochains jours afin de leur permettre d’adapter leurs comportements

Décideurs politiques:

Déterminer les facteurs de pollution pour mettre en place des mesures de réduction

Notre démarche: 4 étapes pour prédire la pollution

Notre hypothèse:

La pollution de l'air peut être prédite par la densité du trafic urbain, les émissions domestiques, l'influence des centrales électriques et les conditions météorologiques

Notre méthodologie:

Appropriation du Dataset → Tests de models → Adaptation du scope de l'étude → Imaginer d'autres variables

Les données: riches, mais non uniformes

Variables disponibles:

- 2 ans de mesures journalières (+35000 lignes)
- 54 villes dans 34 états américains
- Jusqu'à 6 poiluants mesurés
- Jusqu'à 10 variables mesurées

Défis:

- Mesures météorologiques manquantes
- Mesures de polluants manquantes
- Pas de mesure globale de pollution

Travail préliminaire: cleaning & création d'indices

Pre-cleaning:

- Suppression des colonnes météo inutiles;
- Conservation des lignes où le **pm25** est présent
- Nettoyage et conversion des colonnes démographiques en numériques.
- Calcul et ajout d'un indicateur **Share_At_Home** représentant la proportion de personnes restant à domicile

Création d'indices pour mesurer la pollution globale:

- **Méthodologie** :
  - Sélection des polluants principaux : $PM_{2.5}$, PM_{10}, NO_2, CO, O_3.
  - SO_2 exclu en raison de données insuffisantes ou peu fiables.
  - Moyenne des concentrations normalisées
- **Formules utilisées** :

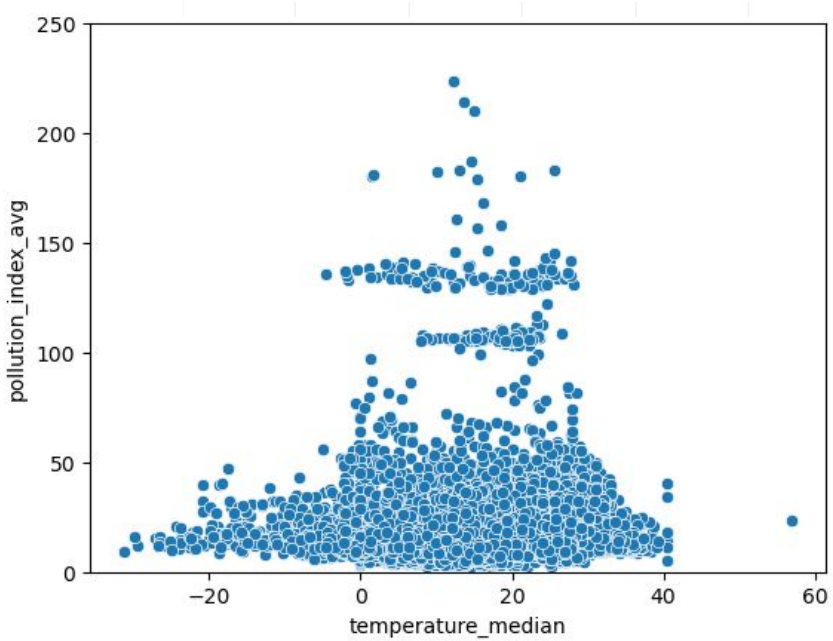
$$\text{Indice Composite}_{\text{mean}} = \frac{PM2.5 + PM10 + NO2 + O3 + CO}{n}$$

$$\text{Indice Composite}_{\text{max}} = \max(PM2.5, PM10, NO2, O3, CO)$$

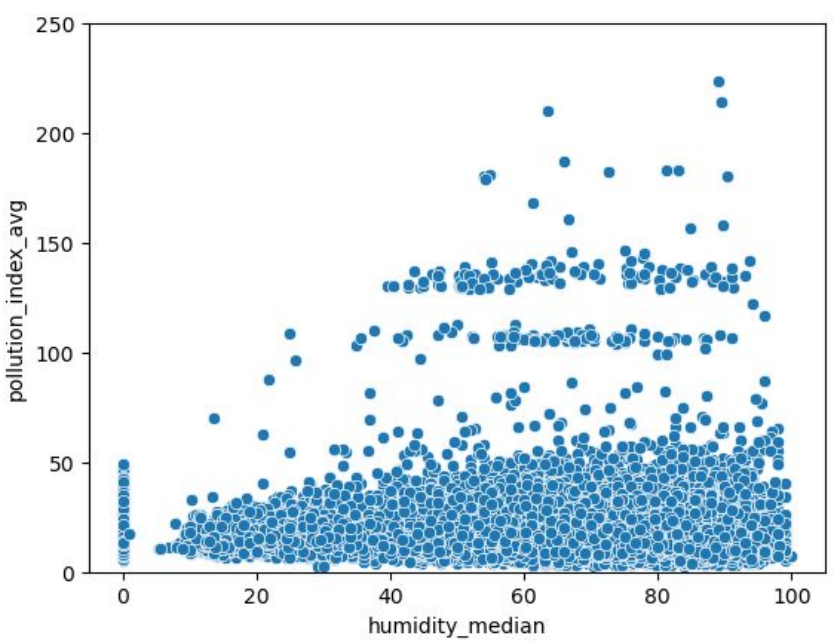
Appropriation du Dataset | Tests de models | Adaptation du scope de l'étude | Imaginer d'autres variables

Correlations visuelles: peu convaincant

Evolution de l'indice de pollution en fonction de la température



Evolution de l'indice de pollution en fonction de l'humidité



Appropriation du Dataset

Tests de models

Adaptation du scope de l'étude

Imaginer d'autres variables

2 modèles testés sur 2 variables: mauvais résultats

R2 - Test	Pollution Moyenne	Pollution Max
Linear Regression	0,03	0,02
Random Forest Regressor	0,2	0,26

Mauvais résultats

Appropriation du Dataset | **Tests de models** | Adaptation du scope de l'étude | Imaginer d'autres variables

Adaptation du périmètre pour augmenter la fiabilité


Quoi	Retrait des outliers	Modèle ville par ville	Modèle année par année
Pourquoi	*Erreurs manifestes*	*Grande variabilité*	*2020 Covid*
Comment	*3 Std*	*Los Angeles*	*2019 seulement*
Impact	*Aucun*	***Positif***	*Aucun*

Appropriation du Dataset → Tests de models → **Adaptation du scope de l'étude** → Imaginer d'autres variables

Ajout de variables temporelles avec un certain succès

Hypothèse	Tests	Conclusions
La pollution suit une logique de stocks: certains peuvent mettre plus d'une journée à se dissiper	- Ajout pollution D-1 comme prédicteur	- **Impact R-score, pas MEPA**
	- Remplacement par Résultats jour + 1	- **Pas d'impact**

Appropriation du Dataset → Tests de models → Adaptation du scope de l'étude → **Imaginer d'autres variables**

Résultats: amélioration mais pouvoir prédictif faible

Meilleur modèle : RandomForestRegressor qui nous permet de **prédire à 16% près**, malgré un fort overfit

Verdict : le modèle permet seulement de prédire la pollution dans une ville donnée

Certaines variables permettent quand même de mieux prédire que d'autres:

- Bourrasques de vent
- Humidité
- Temperature
- % de la population restant à la maison

Améliorations pour augmenter l’utilité du modèle

Solutions

- Réduire l’échelle d’étude
- Sélectionner d’autres paramètres significatifs structurels des villes
- Une meilleure régularité dans la collecte de données
- D’autres modèles prédictifs

Jedha

Des questions ?


Annexes

RandomForestRegressor - Los Angeles

R2
- Train = 0.68
- Test = 0.38

MEPA
- Train = 15%
- Test = 17%

	feature_names	coefficients
0	numerical_pipeline__wind-gust_median	0.264635
3	numerical_pipeline__share_at_home	0.204644
7	numerical_pipeline__temperature_median	0.134230
6	numerical_pipeline__humidity_median	0.121478
5	numerical_pipeline__mil_miles	0.114509
2	numerical_pipeline__dew_median	0.054780
1	numerical_pipeline__pp_feat	0.046844
4	numerical_pipeline__pressure_median	0.040556
13	categorical_pipeline__day_of_week_Wednesday	0.007126
12	categorical_pipeline__day_of_week_Tuesday	0.003021
10	categorical_pipeline__day_of_week_Sunday	0.002774
9	categorical_pipeline__day_of_week_Saturday	0.002740
8	categorical_pipeline__day_of_week_Monday	0.001401
11	categorical_pipeline__day_of_week_Thursday	0.001264

RandomForestRegressor - Los Angeles - augmenté de la pollution de la veille

R2
- Train = 0.69
- Test = 0.44

MEPA
- Train = 14%
- Test = 17%

	feature_names	coefficients
0	numerical_pipeline__wind-gust_median	0.264635
3	numerical_pipeline__share_at_home	0.204644
7	numerical_pipeline__temperature_median	0.134230
6	numerical_pipeline__humidity_median	0.121478
5	numerical_pipeline__mil_miles	0.114509
2	numerical_pipeline__dew_median	0.054780
1	numerical_pipeline__pp_feat	0.046844
4	numerical_pipeline__pressure_median	0.040556
13	categorical_pipeline__day_of_week_Wednesday	0.007126
12	categorical_pipeline__day_of_week_Tuesday	0.003021
10	categorical_pipeline__day_of_week_Sunday	0.002774
9	categorical_pipeline__day_of_week_Saturday	0.002740
8	categorical_pipeline__day_of_week_Monday	0.001401
11	categorical_pipeline__day_of_week_Thursday	0.001264