Introduction

Le *Data Challenge Bertin Technologies 2025* porte sur la détection et la quantification de gaz toxiques à partir de mesures issues du dispositif **ChemProX**, un détecteur portable développé par **Bertin Technologies**. Ce capteur repose sur le principe de l'**Ion Mobility Spectrometry (IMS)**, une technologie de détection rapide permettant d'identifier la nature chimique d'un échantillon gazeux en analysant la mobilité d'ions dans un champ électrique.

Chaque observation du jeu de données associe un vecteur de **13 capteurs** à un ensemble de **23 cibles continues**, correspondant à des niveaux d'alarme pour différentes familles d'agents chimiques. La tâche formulée est donc une **régression multi-sorties**, évaluée à l'aide d'une *Root Mean Squared Error* (RMSE) pondérée, attribuant une importance accrue aux erreurs sur les classes actives afin de favoriser la détection correcte des situations critiques.

L'objectif de cette étude est d'analyser en profondeur la structure des signaux IMS, de comprendre l'influence des facteurs environnementaux tels que l'humidité, et de concevoir un pipeline de modélisation robuste face aux dérives expérimentales.

Analyse exploratoire des données

Structure des données. Le jeu d'entraînement comprend 202,933 échantillons et le jeu de test 134,673. Les données d'entrée comportent **quatorze variables** : un identifiant unique (`ID`) reliant les fichiers d'entrée et de sortie ; huit capteurs IMS (`M4−M7` et `M12−M15`) correspondant à des intensités de dérive issues de deux chambres de mesure ; quatre capteurs auxiliaires (`S1−S3` et `R`) fournissant des mesures complémentaires ; et enfin la mesure d'humidité absolue (`Humidity`), exprimée en unités arbitraires, qui influence fortement la réponse IMS.

Un dispositif d'Ion Mobility Spectrometry (IMS) mesure le *temps de dérive* d'ions soumis à un champ électrique à travers un gaz neutre (Figure 1). Ce temps de dérive dépend directement de la *mobilité ionique*, elle-même influencée par les interactions entre ions et molécules neutres du milieu.

L'analyse de la matrice de corrélation entre les canaux IMS (`M4−M7`, `M12−M15`) confirme une forte interdépendance entre capteurs adjacents, cohérente avec leur disposition physique connue. Les capteurs voisins enregistrent des dérives ioniques partiellement communes, traduisant des réponses instrumentales corrélées. Cette structure interne justifie la construction de *descripteurs de forme et de contraste inter-blocs*, capables de capturer la géométrie du profil IMS (pentes, décalages, courbures) et de modéliser plus finement la *dynamique spectrale* tout en atténuant l'influence de l'humidité.

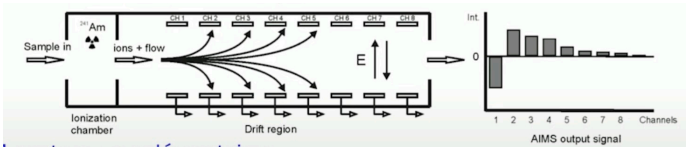

FIGURE 1 – Fonctionnement du dispositif IMS : les ions dérivent sous un champ électrique à travers le gaz porteur et sont détectés sur plusieurs canaux.

Problématique majeur. L'humidité joue ici un rôle critique : la présence de vapeur d'eau modifie la composition du *gaz porteur* et favorise la formation de *clusters hydratés*, dont la mobilité est plus faible. Cette modification se traduit par un *décalage temporel* et une *atténuation* de l'intensité des signaux IMS, même à concentration chimique constante. Autrement dit, une variation d'humidité peut produire une réponse instrumentale différente pour un même agent chimique, phénomène appelé *dérive hygrométrique*.

Dans le contexte du challenge, cette dérive se traduit par un *covariate shift* marqué entre les ensembles d'entraînement et de test (Figure 2). Le jeu d'entraînement se concentre principalement sur des régimes d'humidité *faibles* et, dans une moindre mesure, *élevés*, tandis que le jeu de test présente une distribution plus homogène, avec une nette sur-représentation des régimes *intermédiaires* et *élevés*. Un tel déséquilibre compromet la *généralisation des modèles* : les estimateurs appris sous des régimes secs tendent à se dégrader lorsque les conditions expérimentales changent, apprenant des corrélations parasites entre les canaux IMS et l'humidité plutôt que la signature chimique réelle des gaz.

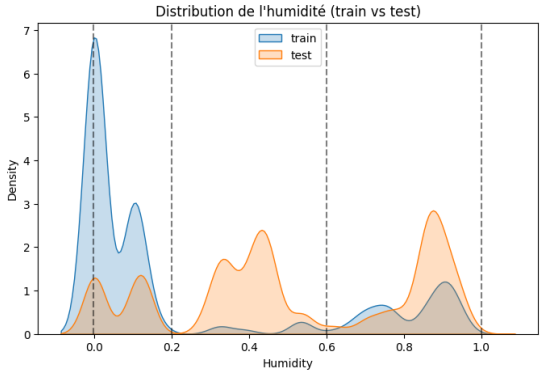

FIGURE 2 – Distribution d'humidité (train vs test). Le test surreprésente les régimes *mid/high*.

Mise en évidence du *covariate shift* et du *domain shift*. Pour quantifier le désalignement entre `train` et `test`, trois indicateurs ont été calculés sur la variable `Humidity` : le *Population Stability Index* (PSI, divergence de distribution), la distance de Kolmogorov–Smirnov (KS, écart maximal cumulé) et la distance de Wasserstein (W, déplacement moyen entre distributions). Les résultats révèlent une dérive marquée ($PSI = 0{,}86$, $KS = 0{,}59$, $W = 0{,}29$), avec environ $60\,\%$ d'observations du jeu `test` en régimes moyen ou élevé, contre seulement $26\,\%$ dans `train`. À l'échelle multivariée, un classifieur de domaine atteint une AUC moyenne de $0{,}78$, confirmant un *domain shift* global entre les deux ensembles.

Pour évaluer son impact sur la relation fonctionnelle $X \to Y$, nous avons comparé les matrices de corrélations $\mathrm{corr}(X_j, Y_k)$ dans trois zones d'humidité (*low*, *mid*, *high*), en considérant l'ensemble des variables d'entrée (`M4−M7`, `M12−M15`, `S1−S3`, `R`, `Humidity`). La figure 3 met en évidence des variations localisées, surtout en régime humide, montrant que l'humidité affecte à la fois les distributions marginales et la relation $X \to Y$.

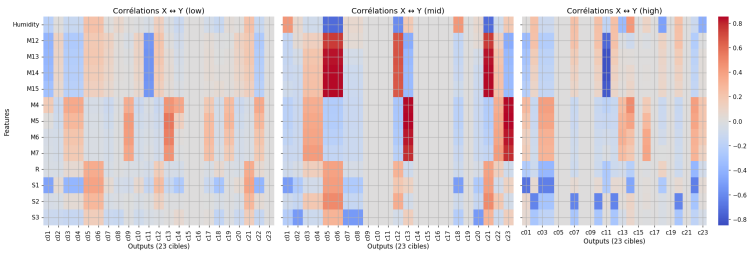

FIGURE 3 – Corrélations $X \leftrightarrow Y$ selon les zones d'humidité (*low*, *mid*, *high*). Les canaux IMS présentent une sensibilité accrue aux régimes hygrométriques élevés.

Prétraitement et construction des variables

Principe issu de l'EDA. L'analyse exploratoire a montré que (i) l'humidité constitue un facteur de dérive majeur perturbant plusieurs canaux IMS, et (ii) les huit canaux IMS se structurent en deux blocs corrélés présentant une organisation spatiale spécifique (*M4–M7* et *M12–M15*). Sur cette base, le pipeline de modélisation corrige d'abord l'effet de l'humidité avant de construire des descripteurs de forme et de contraste, exploitant la structure interne observée des canaux IMS.

Étape 1 - Dé-humidification capteur-par-capteur

Pour chaque capteur j, on modélise la tendance moyenne du signal selon l'humidité H et l'on travaille sur le résidu :

$$X_{ij} = f_j(H_i) + \varepsilon_{ij}, \qquad R_{ij} = X_{ij} - \widehat{f}_j(H_i).$$

La fonction f_j est estimée par une *régression polynomiale régularisée* (Ridge) sur H :

$$\widehat{f}_j(h) = \beta_{j,0} + \beta_{j,1}h + \beta_{j,2}h^2 + \cdots + \beta_{j,d}h^d,$$
$$\widehat{\beta}_j = \arg\min_{\beta\in\mathbb{R}^{d+1}} \sum_{i=1}^{n} \big(X_{ij} - \beta^\top \phi(H_i)\big)^2 + \lambda\|\beta\|_2^2,$$

avec $\phi(H) = (1, H, H^2)^\top$ et $\lambda = 30$. La correction est *externe* (fit sur la distribution d'humidité conjointe train+test sans labels), puis les colonnes résiduelles R_j sont ajoutées au jeu de données (suffixe `_resid`) *sans supprimer* les bruts X_j.

Important : pour prévenir tout raccourci statistique lié à l'humidité (le modèle apprenant à prédire les gaz à partir de H plutôt que des signaux), les variables H et H^2 sont supprimées avant l'entraînement du modèle final.

Choix de la méthode. J'ai comparé plusieurs stratégies de correction (splines régulières, kNN lissé, polynôme non régularisé), évaluées sur des indicateurs de dérive (PSI, KS) et un classifieur de domaine (AUC). La Ridge polynomiale d'ordre 2 s'est montrée la plus stable et la plus efficace pour réduire le *covariate shift* sans déformer la structure physique du signal ; c'est donc l'option retenue.

Étape 2 - Descripteurs de forme IMS (intra-échantillon)

On résume la structure spatiale des deux blocs IMS $\big(B_1,\ B_2\big)$ au sein de chaque échantillon afin de caractériser la forme du profil plutôt que sa valeur absolue :

— **Vues normalisées.** Normalisation L2 sur les 8 canaux pour analyser la forme relative plutôt que l'amplitude.

— **Dérivées locales.** Différences adjacentes $\Delta M_k = M_{k+1} - M_k$ dans chaque bloc.

— **Courbure centrale.** Indice Δ^2_{cent} par bloc pour capter la concavité locale du profil.

— **Ratios adjacents.** Log-ratios locaux au centre de chaque bloc (version minimale, robuste).

— **Contraste inter-blocs.** $\log\big(\sum B_1 / \sum B_2\big)$ pour mesurer le transfert d'énergie entre chambres.

— **Moments de forme.** Centroïde et largeur effective de chaque bloc, ainsi que des statistiques globales (moyenne, écart-type, énergie, IQR, amplitude de pic).

— **Cosine inter-blocs.** Similarité directionnelle entre chambres sur la vue normalisée L2.

Les capteurs auxiliaires $S1$–$S3$ et R ont été conservés sous leur forme brute. D'une part, leur signification physique n'étant pas documentée, aucune transformation interprétable ne pouvait être justifiée, contrairement aux signaux IMS dont la structure spatiale permet une interprétation physique directe. D'autre part, la forêt aléatoire utilisée en aval capture déjà des interactions et des effets de seuil non linéaires, rendant inutile l'ajout de transformations arbitraires sur ces variables.

Modélisation et sélection du modèle

Candidats évalués. Plusieurs variantes de *Random Forest* multi-sorties et de *Gradient Boosting* ont été testées. Les modèles de Boosting se sont révélés instables sous changement d'humidité, tandis que la Random Forest offrait la meilleure robustesse inter-domaines et le meilleur score public.

Configuration de features retenue. Les signaux IMS ont été corrigés par une *dé-humidification externe* Ridge(Poly(H)) capteur-par-capteur, appliquée aux seuls canaux IMS ($M4$–$M7$, $M12$–$M15$), puis enrichis de descripteurs géométriques (L2, Δ, Δ^2, contrastes inter-blocs, moments de forme). Les auxiliaires ($S1$–$S3$–R) ont été conservés tels quels, la forêt aléatoire étant insensible aux différences d'échelle. Un *clipping quantile* (0.5–99.5 %) a été appliqué pour limiter l'influence des valeurs extrêmes sur les signaux bruts et résiduels. Des variantes explorées (prototypes k-means pour la distribution spatiale des signaux, vue robuste standardisée par MAD, ratios étendus) n'ont pas amélioré les performances et ont été abandonnées, la configuration retenue privilégiant la simplicité et la stabilité.

Ajustement des hyperparamètres. Les hyperparamètres ont été fixés empiriquement afin de trouver un compromis entre capacité d'apprentissage et régularisation. Un nombre d'arbres élevé ($n_{\text{estimators}} = 2200$) assure une faible variance, tandis qu'une profondeur modérée (`max_depth` $= 16$) et un seuil minimal de feuilles (`min_samples_leaf` $= 65$) limitent le surapprentissage. Les valeurs `max_features` $= 0.36$ et `max_samples` $= 0.74$ favorisent la diversité des arbres et améliorent la stabilité sur différents régimes d'humidité. Cette configuration s'est montrée la plus cohérente et régulière sur les validations internes et le *leaderboard* public.

Pistes exploratoires

Afin d'améliorer la robustesse face aux régimes hygrométriques hétérogènes, plusieurs approches en plus ont été explorées.

Une première consistait à entraîner des *experts locaux* spécifiques à chaque zone d'humidité, puis à combiner leurs prédictions selon la valeur de H observée.

Une seconde tentative a porté sur une pondération par importance, visant à accorder davantage de poids aux régimes hygrométriques sous-représentés dans le jeu d'entraînement mais fréquents dans le test. En pratique, le ré-équilibrage du poids des observations a légèrement amélioré la stabilité sur validation interne, mais a dégradé la généralisation publique, suggérant une sensibilité accrue du modèle aux zones extrêmes d'humidité.

Enfin, comme mentionné dans l'étape 1, des essais de déhumidification par d'autres méthodes ont été testés, ainsi qu'une modélisation conjointe IMS + Humidity avec interactions polynomiales explicites. Ces variantes n'ont pas montré de gain notable : la correction interne tendait à sur-corriger certaines composantes IMS, tandis que les termes d'interaction réintroduisaient indirectement le biais hygrométrique.

Aucune de ces stratégies n'a finalement apporté de bénéfice durable, confirmant la pertinence de la dé-humidification externe capteur-par-capteur retenue dans le pipeline final.