LatinX in AI at ICML 2025

Vancouver, Canada

Evaluation of Machine Learning Regression Techniques for Analyzing Contaminated Soils

Rosa Virginia Encinas Quille, Felipe Valencia de Almeida

University of Sao Paulo, São Paulo, Brazil

July 13-19, 2025

Abstract

Environmental monitoring of contaminated urban soils is essential for risk management and decision-making. This study applies supervised Machine Learning regression models to predict CH4 concentrations using co-measured gases (CO2, O2, H2S, CO) from 128 gas monitoring wells across 14 buildings at EACH-USP (São Paulo, Brazil), collected from 2014 to 2022.

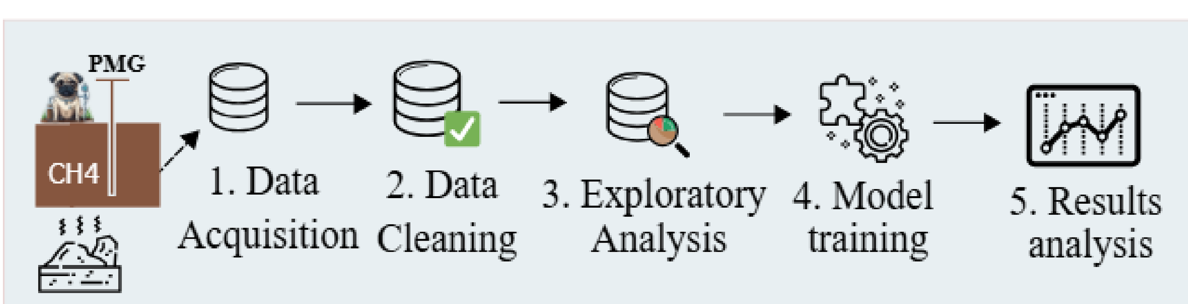
Five models were tested: Linear Regression, k-NN, Decision Tree, Random Forest, and XGBoost. Model performance was evaluated using R^2, MAE, and RMSE.

Random Forest showed superior results in the per-well analysis (Experiment 1), while model performance in the combined-well setting (Experiment 2) varied depending on feature composition, with DT, RF, and XGBoost performing best in different scenarios.

Introduction

- Soil contamination poses major risks to public health and urban management, particularly in densely populated areas. Polluted sites—often known as brownfields—can devalue land, pollute water resources, and threaten ecosystems.
- Environmental investigations generate large volumes of data during the diagnosis, remediation, and monitoring phases. However, conventional analysis methods are often costly, time-consuming, and dependent on laboratory procedures.
- Machine Learning (ML) regression techniques offer a promising alternative for analyzing contaminated soils. Prior studies have shown the potential of models like Random Forest, Decision Trees, and Gradient Boosting for predicting pollutant behavior.
- This study explores the use of ML regression to predict methane (CH4) concentrations in contaminated soils, using other co-measured gases (CO2, O2, H2S, CO) as input features. The goal is to support more accurate and scalable environmental diagnostics.

Methodology



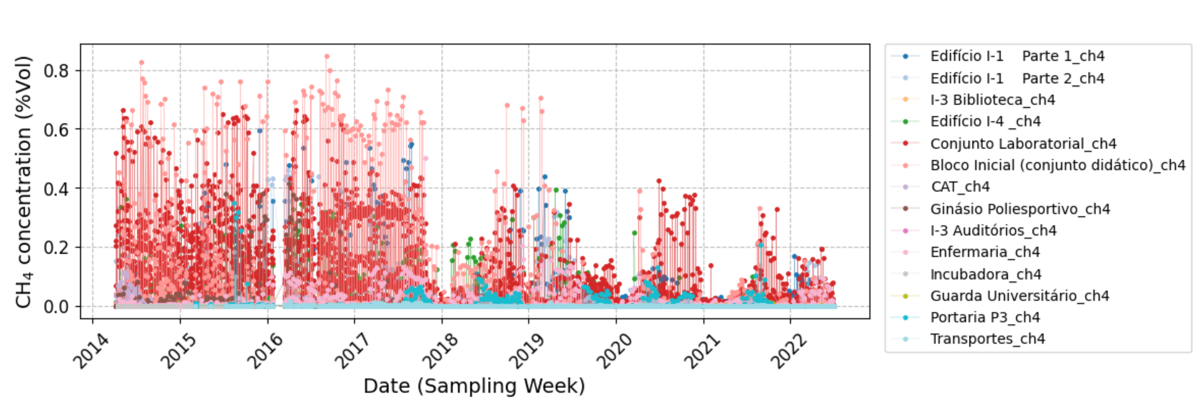
1. Data Collection

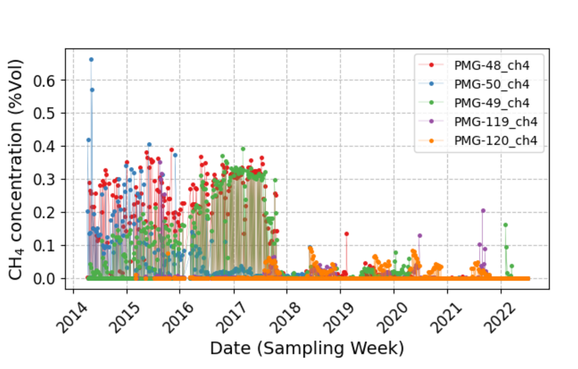
We used in situ gas concentration data collected from 128 monitoring wells across 14 buildings at EACH-USP (São Paulo) between 2014 and 2022 (totaling 100,570 records). Gases monitored include:
CH4, CO2, O2, H2S, and CO.

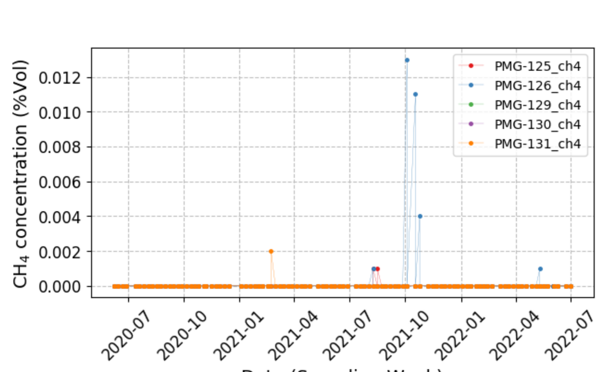
2. Data Cleaning

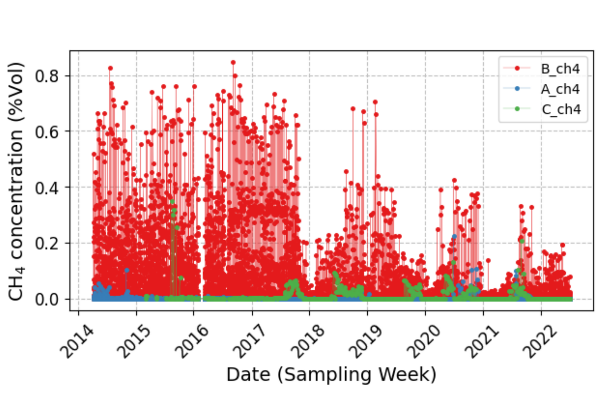
- Merged fragmented files into a unified dataset
- Standardized missing values (NA, #N/A, empty strings)
- Normalized formats (percent to decimals)
- Removed irrelevant columns
- Treated outliers and unified categorical labels
- Converted timestamps to standard format

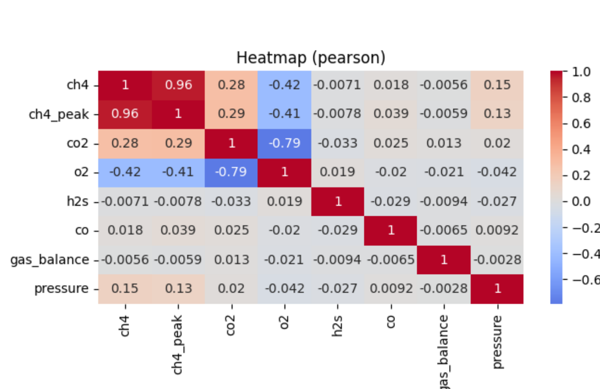
3. Exploratory Analysis











A (shallow, 0.30 m, B (intermediate, 1.00 m), and C (deep, 1.30 m).

Results

Experiment 1:

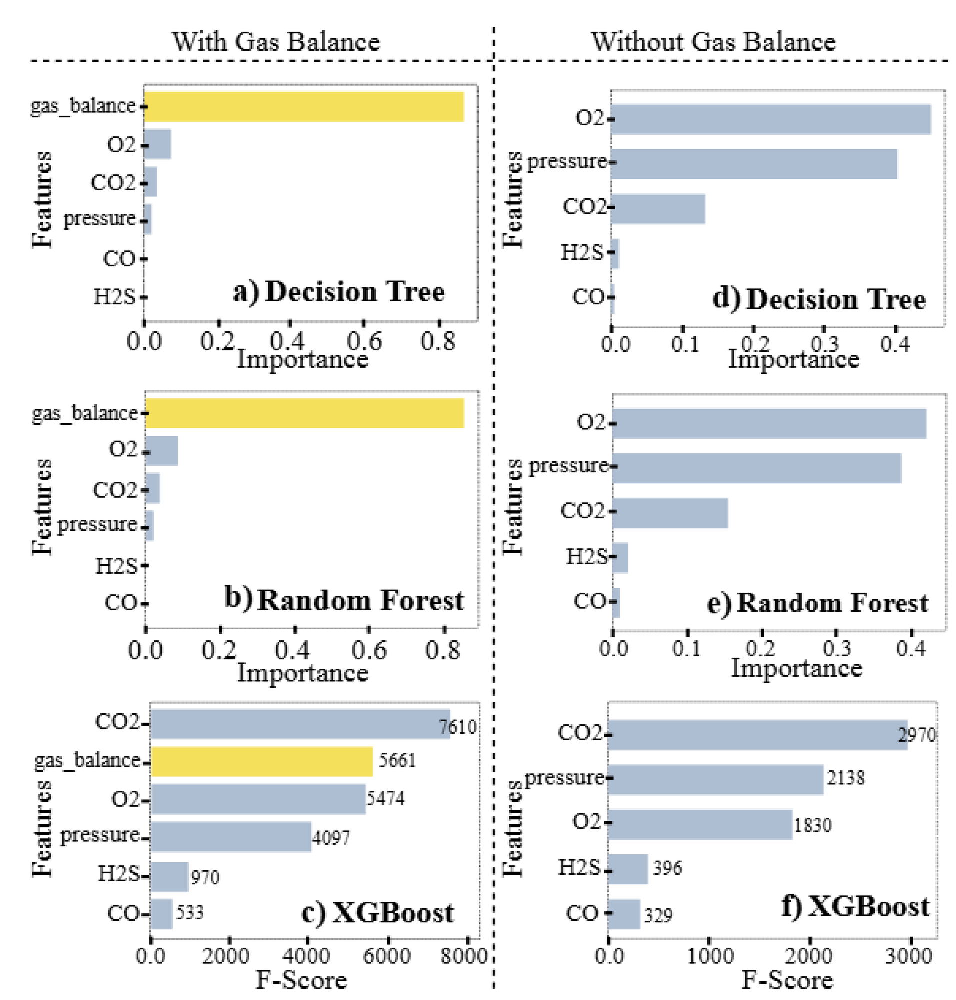
- Models were trained per well to evaluate their ability to capture local gas dynamics.
- Tree-based models (Random Forest, Decision Tree, XGBoost) consistently outperformed others across most wells.
- XGBoost achieved the highest accuracy when using the gas balance feature (R^2 > 0.92 in high-density wells).
- Some low-data wells showed perfect or negative R^2, indicating potential overfitting or noise.
- Excluding gas balance led to reduced accuracy but improved interpretability.

With Gas Balance

	Top PMGs						Bottom PMGs					
	PMG-119			PMG-120			PMG-126			PMG-130		
Model	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
LR	0.0057	0.0199	0.2880	0.0030	0.0090	0.5350	0.0000	0.0000	0.9594	0.0000	0.0000	1.0000
kNN	0.0048	0.0229	0.0616	0.0029	0.0096	0.4641	0.0003	0.0008	-25.8308	0.0000	0.0000	1.0000
DT	0.0027	0.0092	0.8466	0.0016	0.0065	0.7514	0.0001	0.0002	-1.3077	0.0000	0.0000	1.0000
RF	0.0020	0.0066	0.9217	0.0015	0.0065	0.7549	0.0001	0.0002	-0.4892	0.0000	0.0000	1.0000
XGB	**0.0019**	**0.0064**	**0.9268**	**0.0015**	**0.0061**	**0.7819**	**0.0001**	**0.0002**	**-0.0229**	**0.0000**	**0.0000**	**1.0000**
	PMG-48			PMG-49			PMG-129			PMG-131		
Model	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
LR	0.0348	0.0522	0.7632	0.0302	0.0502	0.6592	0.0000	0.0000	1.0000	0.0001	0.0003	-0.0256
kNN	0.0173	0.0450	0.8243	0.0211	0.0494	0.6692	0.0000	0.0000	1.0000	0.0001	0.0003	-0.0256
DT	0.0061	0.0165	0.9763	0.0061	0.0147	0.9706	0.0000	0.0000	1.0000	0.0001	0.0003	-0.0256
RF	0.0061	0.0137	0.9838	0.0063	0.0159	0.9658	0.0000	0.0000	1.0000	0.0001	0.0003	-0.0256
XGB	**0.0053**	**0.0137**	**0.9838**	**0.0059**	**0.0145**	**0.9715**	**0.0000**	**0.0000**	**1.0000**	**0.0001**	**0.0003**	**-0.0256**
	PMG-50						PMG-125					
Model	MAE	RMSE	R^2				MAE	RMSE	R^2			
LR	0.0258	0.0692	0.2178				0.0000	0.0001	0.4936			
kNN	0.0094	0.0450	0.6697				0.0000	0.0002	-0.1077			
DT	0.0039	0.0145	0.9656				0.0001	0.0003	-2.0769			
RF	0.0039	0.0139	0.9684				0.0001	0.0002	-0.5694			
XGB	**0.0037**	**0.0145**	**0.9657**				**0.0000**	**0.0002**	**-0.0142**			

Without Gas Balance

	Top PMGs						Bottom PMGs					
	PMG-119			PMG-120			PMG-126			PMG-130		
Model	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
LR	0.0057	0.0199	0.2880	0.0030	0.0090	0.5350	0.0003	0.0005	-8.6053	**0.0000**	**0.0000**	**1.0000**
kNN	0.0052	0.0216	0.1652	0.0030	0.0099	0.4329	0.0003	0.0006	-15.7452	**0.0000**	**0.0000**	**1.0000**
DT	**0.0037**	**0.0173**	**0.4625**	**0.0025**	**0.0091**	0.5215	**0.0001**	**0.0002**	-1.3077	**0.0000**	**0.0000**	**1.0000**
RF	0.0048	0.0191	0.3445	0.0022	0.0080	**0.6310**	**0.0001**	**0.0002**	-0.2628	**0.0000**	**0.0000**	**1.0000**
XGB	0.0053	0.0230	0.0519	0.0024	0.0086	0.5737	**0.0001**	**0.0002**	**-0.0229**	**0.0000**	**0.0000**	**1.0000**
	PMG-48			PMG-49			PMG-129			PMG-131		
Model	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
LR	0.0348	0.0523	0.7630	0.0302	0.0502	0.6589	**0.0000**	**0.0000**	**1.0000**	0.0001	0.0003	-0.0256
kNN	0.0231	0.0535	0.7513	0.0307	0.0680	0.3739	**0.0000**	**0.0000**	**1.0000**	0.0001	0.0003	-0.0256
DT	0.0250	0.0563	0.7253	0.0244	0.0611	0.4942	**0.0000**	**0.0000**	**1.0000**	0.0001	0.0003	-0.0256
RF	**0.0157**	**0.0385**	**0.8713**	0.0176	0.0399	0.7850	**0.0000**	**0.0000**	**1.0000**	0.0001	0.0003	-0.0256
XGB	0.0166	0.0418	0.8488	**0.0153**	**0.0342**	**0.8420**	**0.0000**	**0.0000**	**1.0000**	0.0001	0.0003	-0.0256
	PMG-50						PMG-125					
Model	MAE	RMSE	R^2				MAE	RMSE	R^2			
LR	0.0258	0.0693	**0.2160**				**0.0000**	**0.0002**	-0.0020			
kNN	**0.0197**	0.0633	**0.3461**				**0.0000**	**0.0002**	-0.2536			
DT	0.0182	0.0653	0.3041				**0.0000**	**0.0002**	-0.1111			
RF	0.0183	**0.0609**	**0.3948**				0.0001	0.0002	-0.8212			
XGB	0.0187	0.0660	0.2889				**0.0000**	**0.0002**	**-0.0142**			



Experiment 2:

- Random Forest delivered the best performance (R^2 = 0.95) when gas balance was included.
- Removing gas balance caused a drop of over 50% in R^2 across all models.
- Linear Regression performance remained unchanged, reflecting its limitations in capturing nonlinear relationships.

Model	With Gas Balance			Without Gas Balance		
	MAE	RMSE	R^2	MAE	RMSE	R^2
LR	0.0133	0.0393	0.1993	0.0133	0.0394	0.1990
k-NN	0.0026	0.0186	0.8217	0.0086	0.0385	0.2331
DT	0.0014	0.0127	0.9170	0.0081	0.0361	0.3249
RF	**0.0011**	**0.0098**	**0.9501**	**0.0078**	**0.0333**	**0.4276**
XGB	0.0015	0.0111	0.9366	**0.0078**	0.0336	0.4178

4. Modeling Approach

- We trained five regression models to predict CH4 using other gases as features:
- Linear Regression, k-Nearest Neighbors (k-NN), Decision Tree, Random Forest, and XGBoost.

Model	Selected Hyperparameters
LR	No hyperparameters (standard least squares)
kNN	n_neighbors = 3
DT	max_depth = 10, min_samples_split = 2, min_samples_leaf = 1, max_features = 'sqrt'
RF	n_estimators = 300, max_depth = 20, min_samples_split = 2, min_samples_leaf = 1
XGB	n_estimators = 100, max_depth = 3, learning_rate = 0.1

5. Experimental Setup

- Experiment 1: Per-well models (individual time series)
- Experiment 2: All-well combined dataset
- Evaluation metrics: R2, MAE, RMSE

Conclusion and recommendation

We evaluated machine learning regressors to predict CH_4 concentrations in contaminated soils using co-measured gases. Random Forest outperformed others in well-specific models, while Decision Tree, Random Forest, and XGBoost varied in performance in aggregated data. Although the gas_balance feature improved predictions, it may cause information leakage.
Future work should explore deep learning and temporal modeling to enhance environmental gas monitoring.

Acknowledgements

R.Q. is supported by the São Paulo Research Foundation (FAPESP) under project grant 2019/21693-0 and by the Institute for Technological Research (IPT) of the State of São Paulo and F.A. is supported by National Council for Scientific and Technological Development (CNPq) under project grant 140253/2021-1.