Predicting Future Contaminant Levels with Machine Learning

Scott L., Jina W., Candan M., Aiden G.
The University of Texas at San Antonio, San Antonio TX, 78249

Abstract

The EPA's UCMR datasets [1] span three time periods: UCMR3 (2013–2015), UCMR4 (2018–2020), and UCMR5 (2023–2025). These datasets include contaminants such as per- and polyfluoroalkyl substances (PFAS), lithium, strontium, and chromium. Each entry provides facility location, measurement method, contaminant, and metrics like the Minimum Reporting Level (MRL) and Analytical Result Value (ARV). ARVs, reported only when exceeding the MRL, leave blank cells for undetected levels, influencing predictions. Our work trained models using ARVs as the target variable and explored regression algorithms, including linear regression for its simplicity and ability to generate outputs such as heatmaps, charts, graphs, and prediction result files. To address missing ARVs, we tested two dataset versions: one filling missing ARVs with zero and another excluding them. While UCMR1 and UCMR2 lacked zip code data, UCMR3–5 provided location-based predictions and visualizations, supporting the development of actionable insights.

Data Preprocessing and Model Approaches

To prepare the UCMR datasets for modeling, we focused on relevant columns: Contaminant, ARV, Collection Date, and Zip Code. Latitude and longitude data were added from a US Zip Codes Database [2] to enable geographical analysis and heatmap generation. Zip codes with missing latitude and longitude information were excluded. The Collection Date values were converted to Unix timestamps to facilitate the analysis of changes over time and optimize data storage and retrieval.

To address missing ARVs, two dataset versions were created:

- ➢ Version 1: Blank ARVs were assigned a value of zero, and zip codes missing latitude or longitude data were excluded. This version retained 85 contaminants across 18,601 zip codes, encompassing 87.74% of the original data.
- ➢ Version 2: Blank ARVs were excluded entirely, along with zip codes missing latitude or longitude data. This version retained 72 contaminants across 16,405 zip codes, representing 13.89% of the original data.

Both dataset versions were used for model training and predictions. The training data were organized into nested dictionaries, with zip codes as primary keys. Each zip code contained inner dictionaries detailing contaminants, collection dates, and ARVs. These dictionaries were saved in JSON format, making them portable and compatible with various platforms for further analysis [3].

Smaller datasets were used for testing to manage the large size of the UCMR datasets and ensure efficient program performance. Larger datasets caused memory issues and crashes. These smaller datasets were created by filtering zip codes from the UCMR3-5 data, which included latitude and longitude information, into CSV files. Each file contained six to ten zip codes. The datasets include:

- ➢ **Dataset 1:** Six random zip codes (6339, 35005, 48858, 82520, 98221, 56633) with missing ARVs set to zero.
- ➢ **Dataset 2:** The same six random zip codes, but with missing ARVs excluded.
- ➢ **Dataset 3:** Ten Texas zip codes (77008, 78201, 78249, 79901, 76705, 75205, 77703, 78409, 79705, 78840) with missing ARVs set to zero.
- ➢ **Dataset 4:** The same ten Texas zip codes, but with missing ARVs excluded.

Model Training

Python was used for data preprocessing and implementing machine learning models targeting ARVs. Models applied included:

- Linear Regression: The primary model, chosen for its simplicity, interpretability, and consistent performance in environmental predictions [4].
- Random Forest Regression: Used for complex data systems [5].
- Gradient Boosting: Applied to enhance accuracy by correcting prior errors [6].
- Gaussian Processes: Explored but faced memory limitations.

Linear regression emerged as the preferred model for its consistent performance and straightforward interpretability.

Preprocessing Outputs

- ➢ Folium Heatmaps: Generated for specific zip codes, these visualizations provide a geographical overview of contaminant concentrations (Figs. 1-5, 9, 13, 17). They also serve as a baseline for comparison with heatmaps of predicted data after model training.
- ➢ JSON Files: Training data, organized by zip code, were saved as nested dictionaries in JSON format. Each zip code entry included contaminants, collection dates, and ARVs. These files are shareable for further analysis and can serve as a reference for comparing with prediction outputs.

Exploring the Impact of Missing ARVs

As described in the Data Preprocessing and Model Approaches section, two dataset versions were created to address missing ARVs:

- ➢ Version 1: 85 contaminants, with missing ARVs set to zero.
- ➢ Version 2: 72 contaminants, excluding entries with missing ARVs.

Figures 1–4 illustrate the impact of handling missing ARVs differently, affecting both the visual representation of contaminant distribution and the accuracy of predictions.

- ➢ Figures 1 and 3: Heatmaps based on Version 1, where missing ARVs were set to zero.
- ➢ Figures 2 and 4: Heatmaps from Version 2, excluding missing ARVs.

Figures 1 and 2 compare lithium data with missing ARVs either set to 0 or excluded.

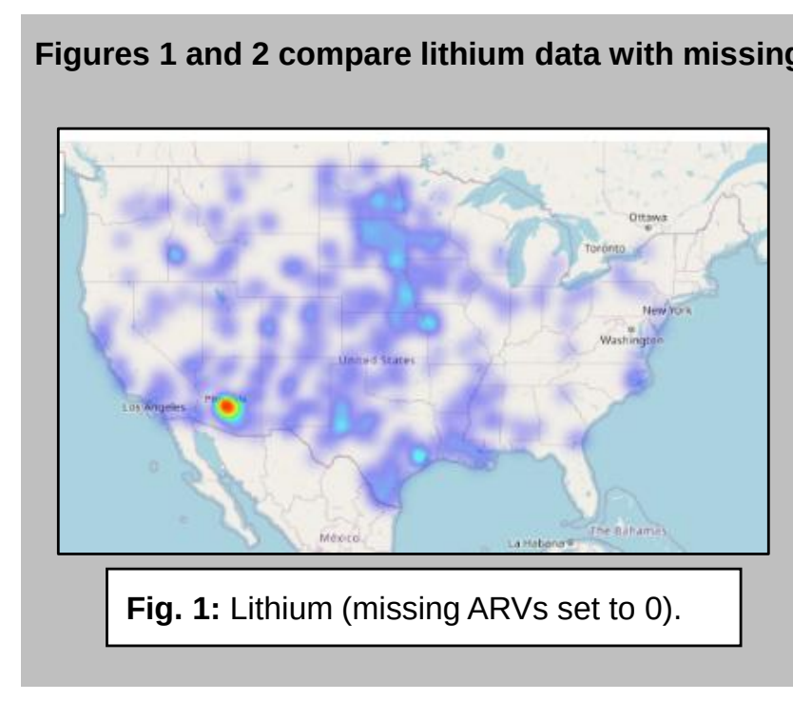


Fig. 1: Lithium (missing ARVs set to 0).

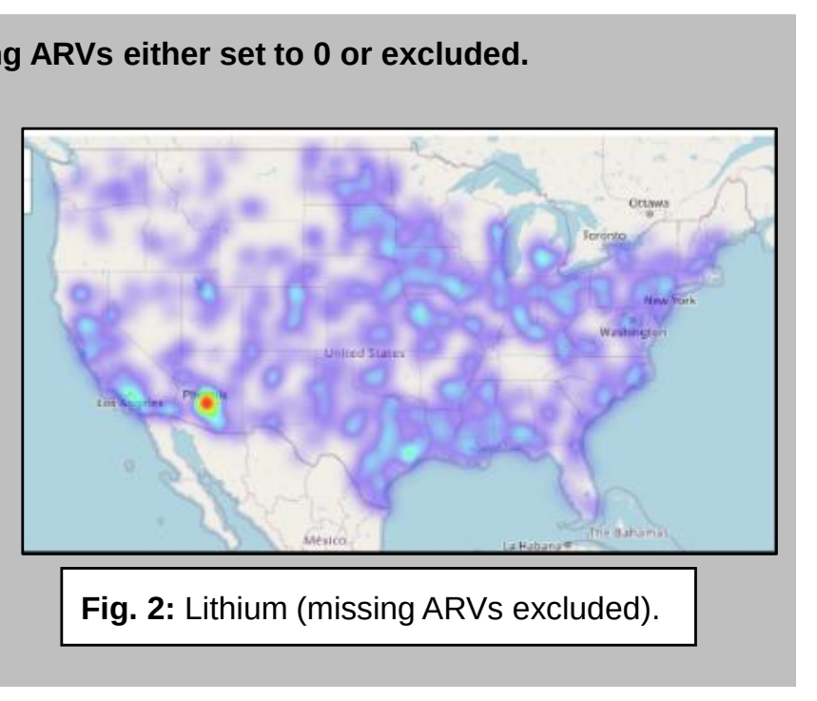


Fig. 2: Lithium (missing ARVs excluded).

Figures 3 and 4 compare PFBA data with missing ARVs either set to 0 or excluded.

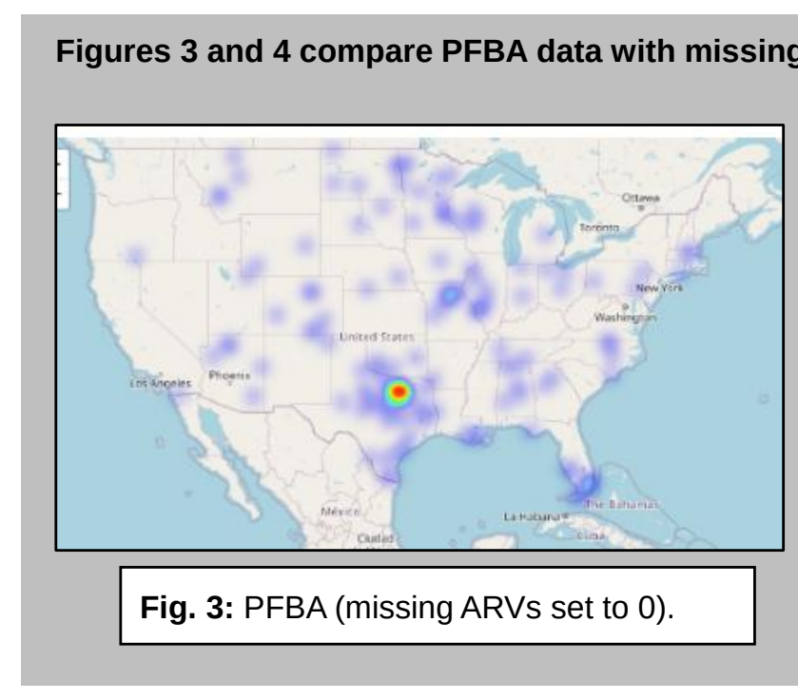


Fig. 3: PFBA (missing ARVs set to 0).

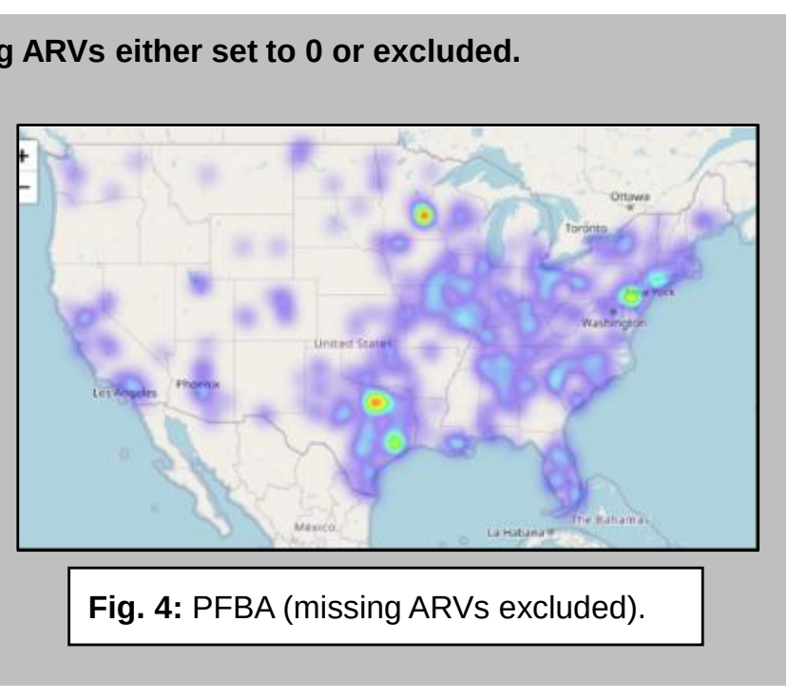


Fig. 4: PFBA (missing ARVs excluded).

Results: Predicting Contaminants by Zip Code

Figures 5-26 display model training and prediction results for the four datasets, which include a combination of heatmaps, scatter plots, line graphs, and bar charts of the top ten predicted ARVs for various contaminants.

The figures compare outputs from linear regression and random forest regression, displaying Mean Squared Error (MSE) and R² scores for select models. For example, Figures 6, 10, 14, and 18 highlight predictions for zip code 6339 with HAA9 and 75205 with PFHxA respectively. The scatter and line plots (Figs. 7, 11, 15, 19) compare recorded ARVs with predicted values for each zip code and contaminant, while the bar charts (Figs. 8, 12, 16) focus on the top ten predicted ARVs for each contaminant.

Figures 5-8: Dataset (1) - Six Random Zip Codes with Missing ARVs Set to 0 Using Linear Regression

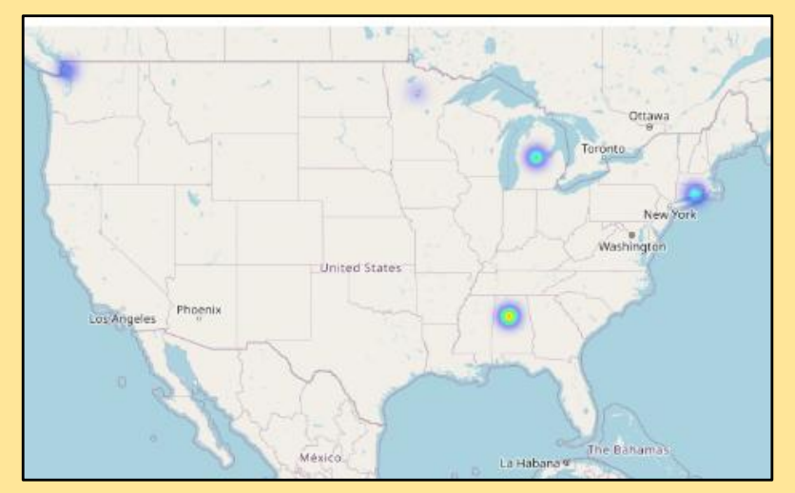


Fig. 5: Heatmap of HAA9 based on reported data for six zip codes. Heatmaps for all 85 contaminants were produced, but only HAA9 is shown here.

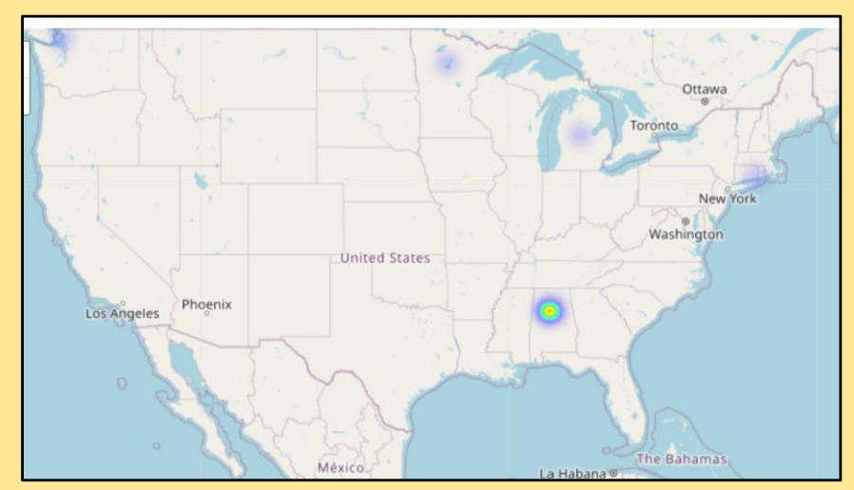


Fig. 6: Heatmap for six zip codes of HAA9 predictions.

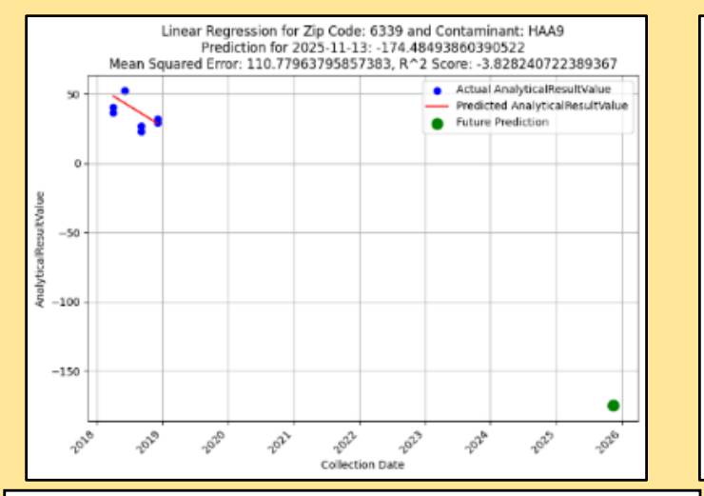


Fig. 7: Linear regression model results for zip code 6339, showing actual ARV, training prediction, and future predictions for HAA9. Graphs were created for all contaminants, but only HAA9 is shown here.

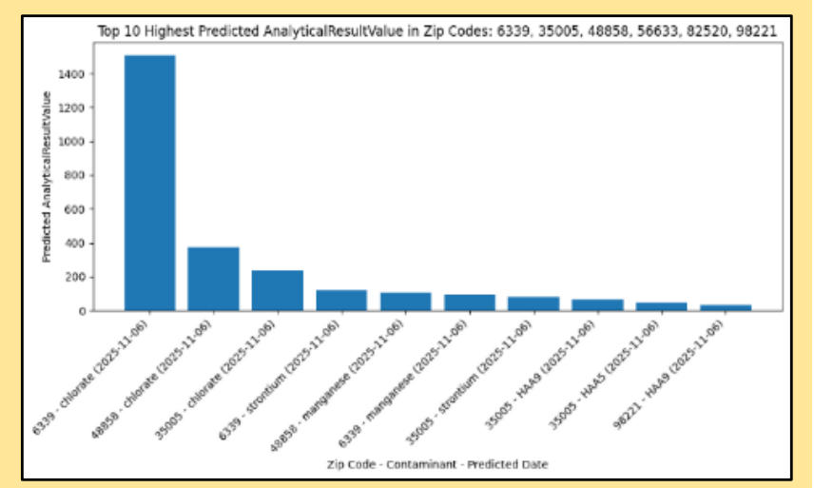


Fig. 8: Top ten contaminant ARV predictions from linear regression, using 85 contaminants.

Figures 9-12: Dataset (2) - Six Random Zip Codes with Missing ARVs Excluded Using Random Forest Regression

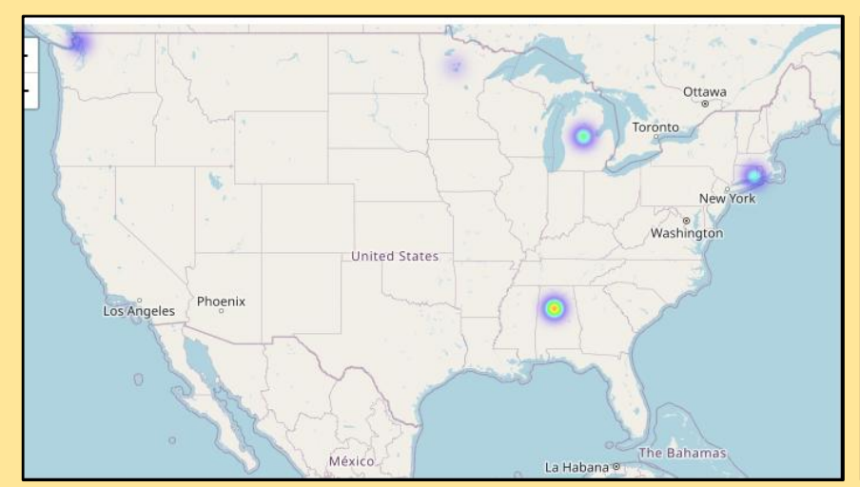


Fig. 9: Heatmap of HAA9 based on reported data for six zip codes. Heatmaps for all 72 contaminants were produced, but only HAA9 is shown here.

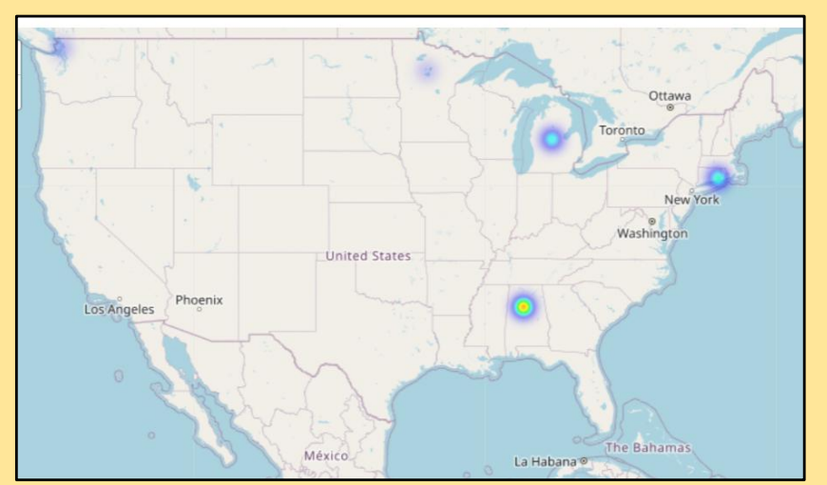


Fig. 10: Heatmap for six zip codes of HAA9 predictions.

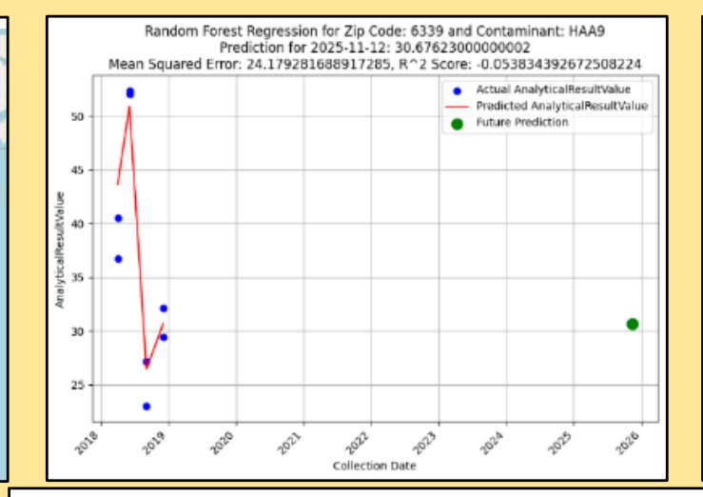


Fig. 11: Random forest regression model results for zip code 6339, showing actual ARV, training prediction, and future predictions for HAA9. Graphs were created for all contaminants, but only HAA9 is shown here.

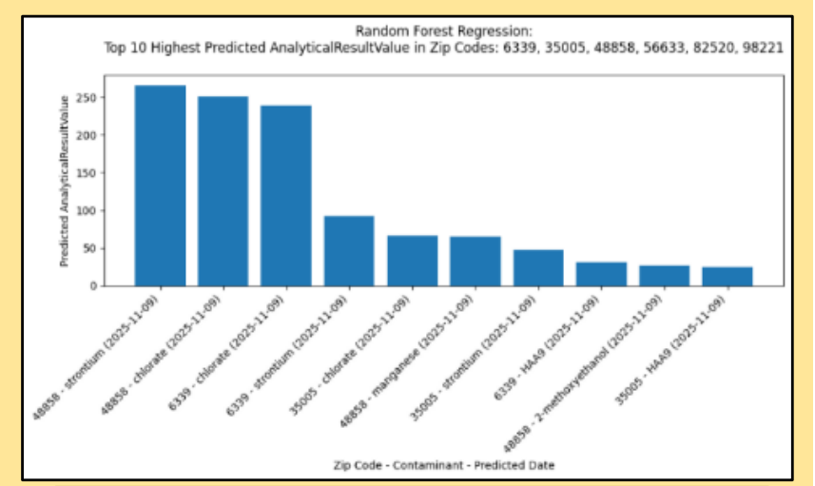


Fig. 12: Dataset (2) top ten Contaminant ARV predictions from random forest regression. Top ten predictions using 72 contaminants.

Figures 13-16: Dataset (3) - Ten Texas Zip Codes with Missing ARVs Set to 0 Using Linear Regression

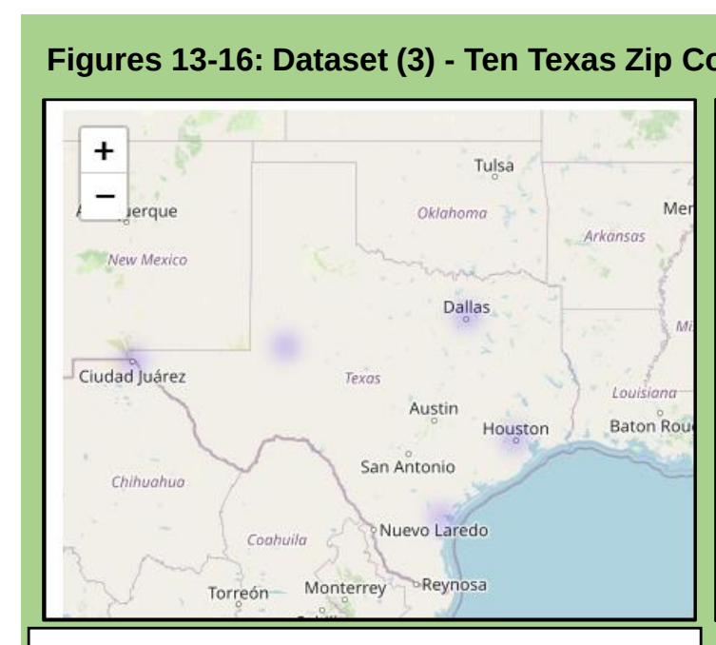


Fig. 13: Heatmap of PFHxA based on reported data for ten TX zip codes. Heatmaps for all 85 contaminants were produced, but only PFHxA is shown here.

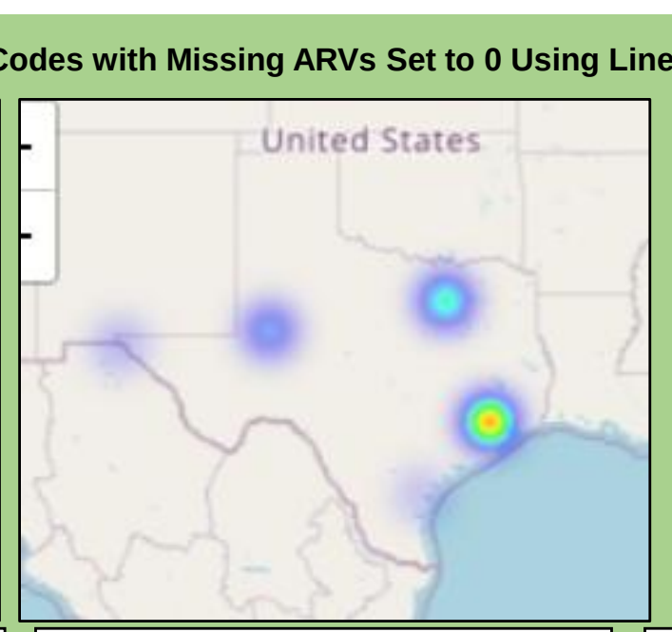


Fig. 14: Heatmap for ten TX zip codes of PFHxA predictions.

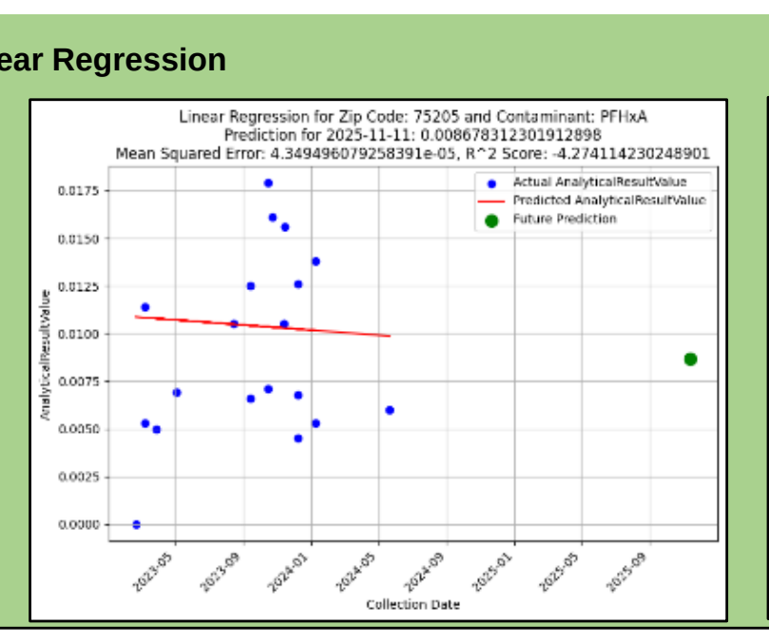


Fig. 15: Linear regression model results for zip code 75205, showing actual ARV, training prediction, and future predictions for PFHxA. Graphs were created for all contaminants, but only PFHxA is shown here.

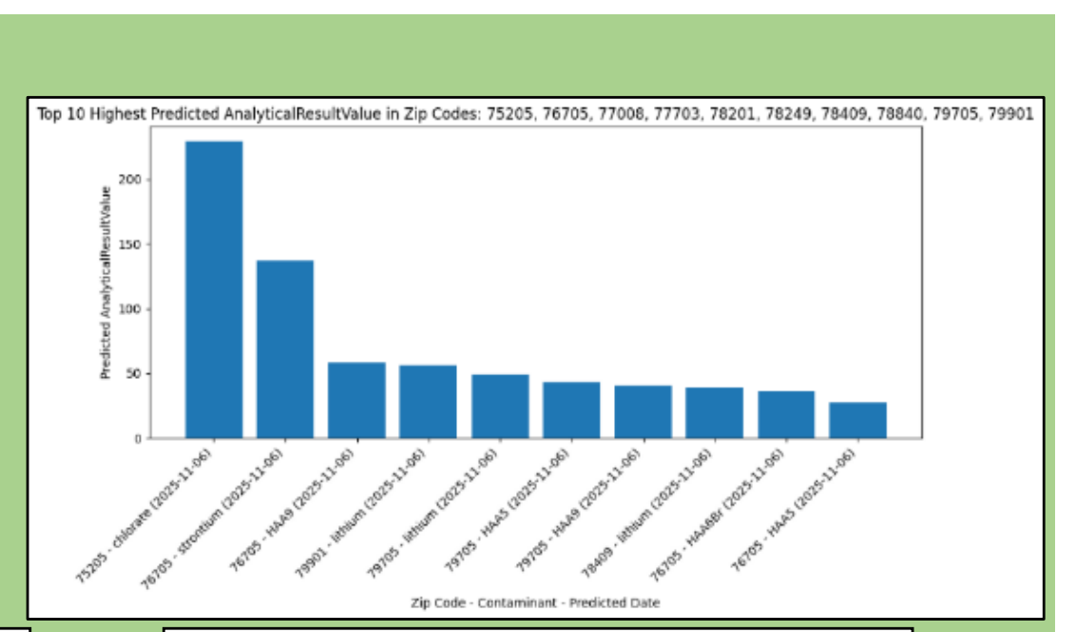


Fig. 16: Top ten contaminant ARV predictions from linear regression, using 85 contaminants.

Figures 17-20: Dataset (4) - Ten Texas Zip Codes with Missing ARVs Excluded Using Linear Regression

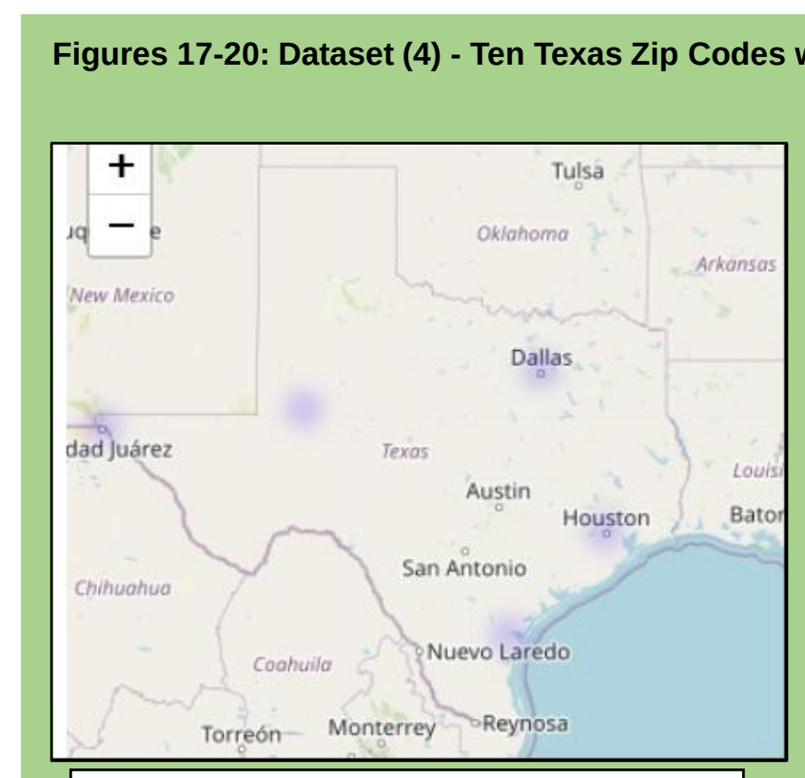


Fig. 17: Heatmap of PFHxA based on reported data for ten TX zip codes. Heatmaps for all 72 contaminants were produced, but only PFHxA is shown here.

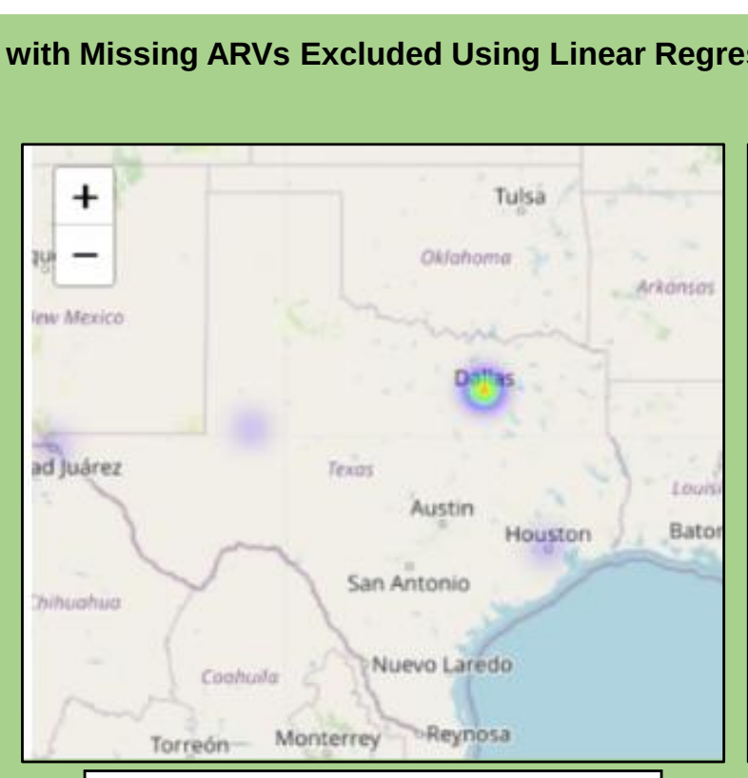


Fig. 18: Heatmap for ten TX zip codes of PFHxA predictions.

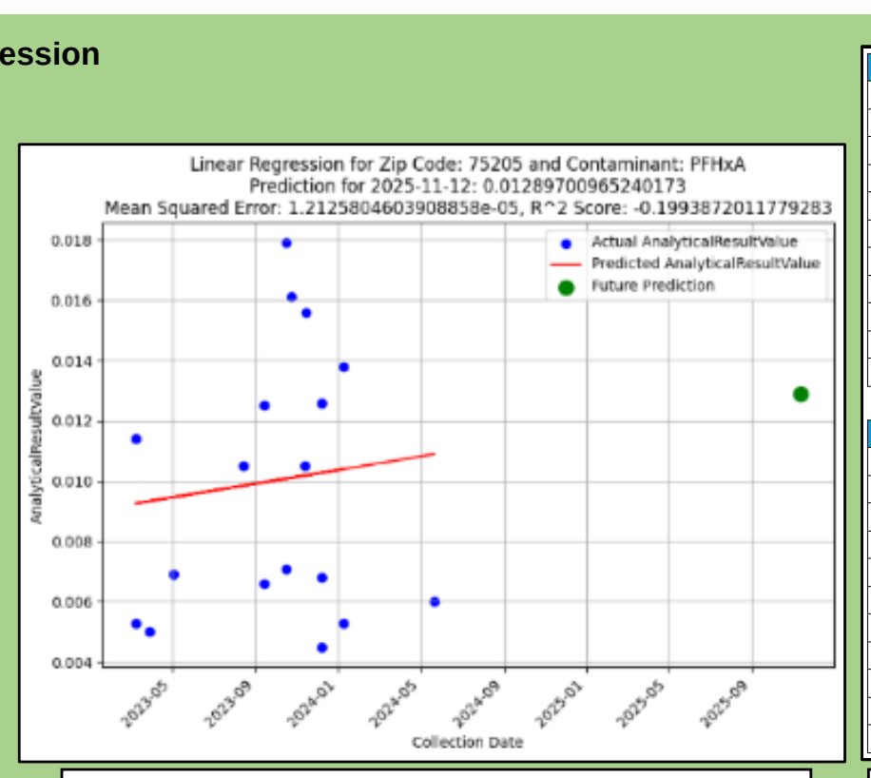
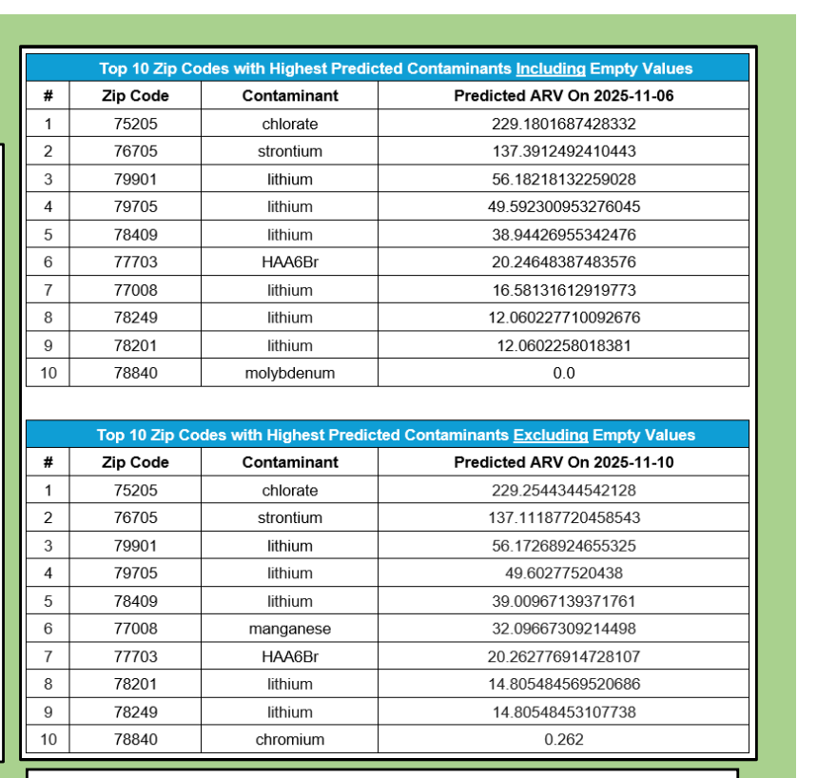


Fig. 19: Linear regression model results for zip code 75205 with missing ARVs excluded, showing actual ARV, training prediction, and future predictions for PFHxA. Graphs were created for all contaminants, but only PFHxA is shown here.



Top 10 Zip Codes with Highest Predicted Contaminants Including Empty Values

#	Zip Code	Contaminant	Predicted ARV On 2025-11-06
1	75205	chlorate	229.1801687428332
2	76705	strontium	137.3912492410443
3	79901	lithium	56.18218132259028
4	79705	lithium	49.592300953276045
5	78409	lithium	38.94426955342476
6	77703	HAA6Br	20.24648387483576
7	77008	lithium	16.58131612919773
8	78249	lithium	12.060227710092676
9	78201	lithium	12.0602258018381
10	78840	molybdenum	0.0

Top 10 Zip Codes with Highest Predicted Contaminants Excluding Empty Values

#	Zip Code	Contaminant	Predicted ARV On 2025-11-10
1	75205	chlorate	229.2544344542128
2	76705	strontium	137.11187720458543
3	79901	lithium	56.17268924655325
4	79705	lithium	49.60277520438
5	78409	lithium	39.00967139371761
6	77008	manganese	32.09667309214498
7	77703	HAA6Br	20.262776914728107
8	78201	lithium	14.805484569520686
9	78249	lithium	14.80548453107738
10	78840	chromium	0.262

Fig. 20: Top ten contaminant ARV predictions from linear regression, comparing missing ARVs set to 0 versus excluded. Subtle differences in predictions were noted, though not visibly apparent in bar graph format.

Figures 21–24 present heatmaps showcasing ARV predictions one year ahead across Datasets 1–4, combining all contaminants. These heatmaps were generated for smaller datasets due to program limitations on larger data.

- Figures 21 and 22 compare prediction heatmaps for six random zip codes, illustrating the impact of different regression methods
- Figures 23 and 24 compare prediction heatmaps for ten Texas zip codes using linear regression

Figs 21 and 22 compare six zip codes with either blank ARVs set to 0 or excluded.

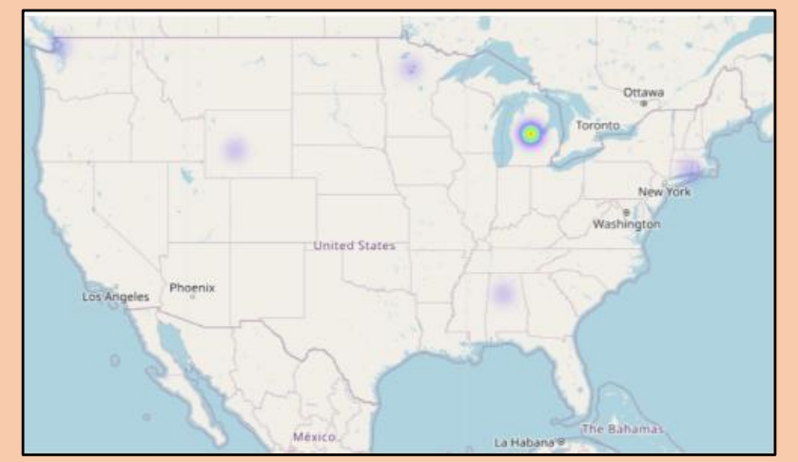


Fig. 21: Dataset 1 heatmap for six random zip codes using linear regression, with missing ARVs set to 0.

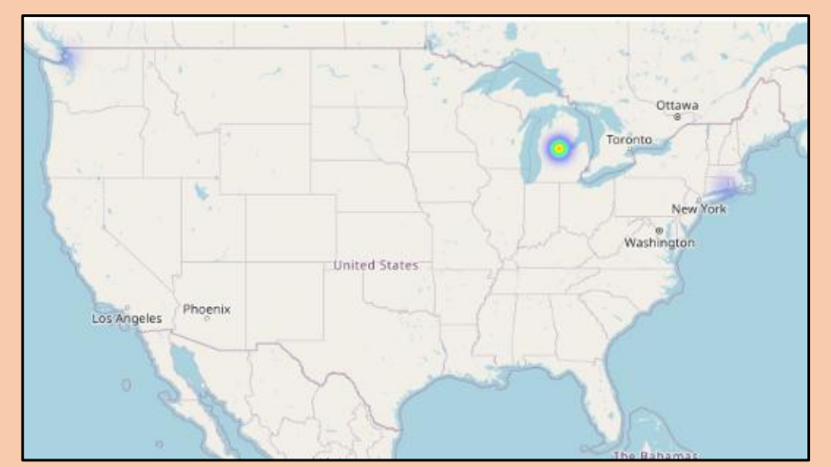


Fig. 22: Dataset 2 heatmap for six random zip codes using random forest regression, with missing ARVs excluded.

Figs. 23 and 24 compare ten TX zip codes with either blank ARVs set to 0 or excluded.

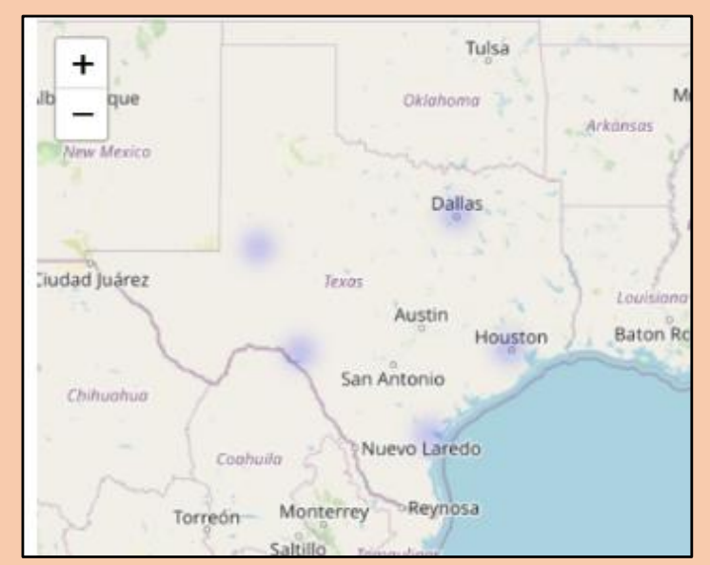


Fig. 23: Dataset 3 heatmap for ten Texas zip codes using linear regression, with missing ARVs set to 0.

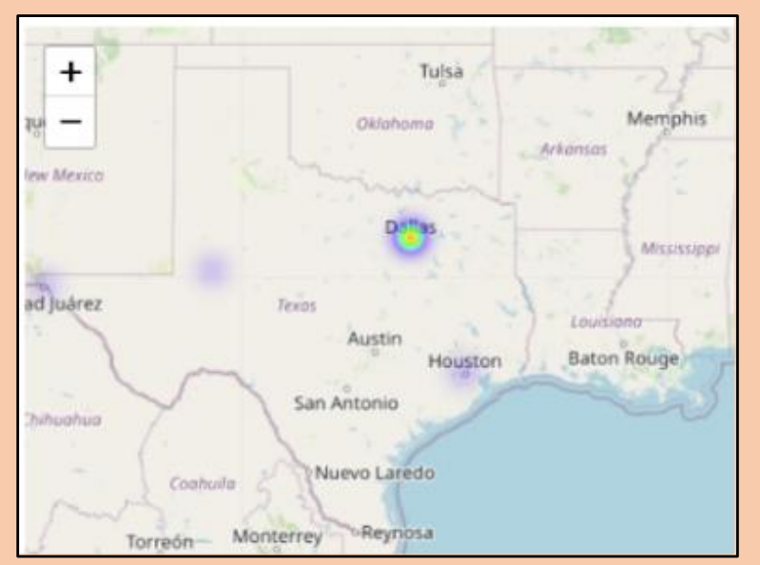
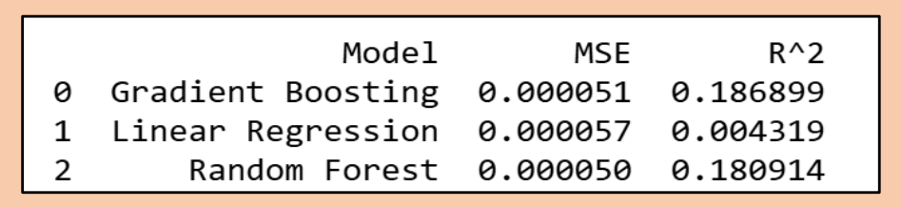


Fig. 24: Dataset 4 heatmap for ten Texas zip codes using linear regression, with missing ARVs excluded.

Results (cont'd)

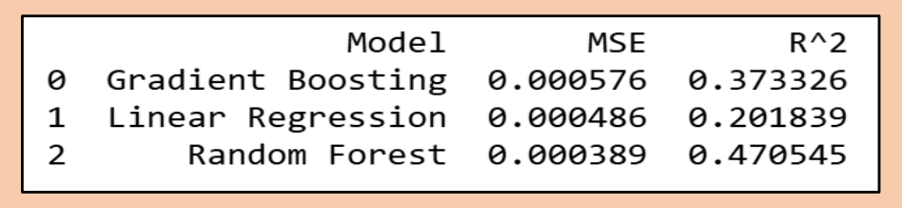
- Figures 25 and 26 further explore the impact of handling missing ARVs on model performance, specifically focusing on PFAS data.
- The most effective solution was processing the dataset by excluding ARVs missing due to being below the MRL, which achieved a significantly higher R² score in the predictions; however, this approach poses a risk of overestimation if the model is used to predict contaminant levels in areas with insufficient measurements.
- The best predictive accuracy was achieved with random forest regression with missing ARVs excluded, as shown in Figure 26.

Figs. 25 and 26: Averaged performance of predictions for the three selected models across exclusively PFAS readings



```
               Model       MSE       R^2
0  Gradient Boosting  0.000051  0.186899
1  Linear Regression  0.000057  0.004319
2      Random Forest  0.000050  0.180914
```

Fig. 25: Average effectiveness metrics of the three explored models when missing values are filled with 0. Note the low R² value and the relative superiority of gradient boosting.



```
               Model       MSE       R^2
0  Gradient Boosting  0.000576  0.373326
1  Linear Regression  0.000486  0.201839
2      Random Forest  0.000389  0.470545
```

Fig. 26: Average effectiveness metrics of the three explored models when missing values are excluded. Note the more appreciable R² values, and the far better performance of random forest regression.

Deliverables

This project produced outputs designed to facilitate data sharing and analysis:

- ➢ **Heatmaps**: Visualizations of recorded and predicted ARVs for various contaminants across zip codes (see Figs. 1-6, 9-10, 13-14, 17-18, and 21-24).
- ➢ **Scatter Plots, Line Graphs, and Bar Charts:** Detailed graphs of predicted ARVs for contaminants (see Figs. 7-8, 11-12, 15-16, and 19-20).
- ➢ **JSON Files:**
  - Reported Data: Organized by zip code, contaminant, ARV, and collection date.
  - Predictions: Contaminant forecasts for 2025, with predicted ARV, MSE, and R² scores.

The JSON files are easily converted to CSV (Comma Separated Values) format but are also optimized for web applications and other data systems. All files and visualizations are portable, enabling seamless sharing and collaboration across platforms.

Conclusions and Recommendations

The zip code-based approach, utilizing nested dictionaries to organize data by zip codes, contaminants, locations, and ARVs, shows promise for continued work. This method, applied to specific zip code datasets, produced varying results depending on whether missing ARVs were set to zero or excluded entirely, emphasizing the need to carefully address blank values to avoid affecting prediction accuracy. The inclusion or exclusion of blank ARVs and the choice of regression algorithm led to noticeable variations in forecasted predictions across all datasets. While MSE and R² scores were calculated, they were not ideal, suggesting room for model improvement.

- ❖ Recommendations:
  - ➢ Address the handling of blank data cells by deciding whether to exclude them or assign values to prevent potential data bias.
  - ➢ Ensure all ARVs are recorded during data collection, even those below the Minimum Reporting Level (MRL), to improve the accuracy of machine learning predictions.
- ❖ Future Work :
  - ➢ For enhanced predictive accuracy, future models should explore XGBoost and LightGBM algorithms. LightGBM, for example, can treat missing values as a unique category, which could enhance predictions of future contamination levels [7].
  - ➢ Using different models could show more promise. A trail test done with Bayesian showed potential, but more testing is needed.

This project delivers comprehensive tools for forecasting contaminant levels, including JSON files, heatmaps, and predictive models, that can be valuable for future analysis and policy development.

References

[1] U.S. Environmental Protection Agency, "Occurrence Data from the Unregulated Contaminant Monitoring Rule [Online]. Available: https://www.epa.gov/dwucmr/occurrence-data-unregulated-contaminant-monitoring-rule ," Accessed: Nov. 10, 2024.
[2] SimpleMaps.com, "US Zip Codes Database," [Online]. Available: https://simplemaps.com/data/us-zips Accessed: Nov. 10, 2024.
[3] "Introducing json," JSON. [Online]. Available: https://www.json.org/json-en.html Accessed: 14-Nov-2024.
[4] Linear Regression: H. Amini, S. M. Taghavi-Shahri, S. B. Henderson, and I. Burstyn, "Using Linear Regression to Estimate Ambient Air Pollution Levels at Unmeasured Locations: Application to Ozone," Environ. Sci. Technol., vol. 53, no. 9, pp. 4785–4794, 2019.
[5] Random Forest: Y. Zhan, Y. Luo, X. Deng, M. L. Grieneisen, and M. Zhang, "Prediction of Dissolved Oxygen in River Systems Using Gradient Boosting Decision Trees," J. Hydrol., vol. 559, pp. 301–311, 2018.
[6] "Gradient boosting," Wikipedia, 02-Oct-2024. [Online]. Available: https://en.wikipedia.org/wiki/Gradient_boosting Accessed: 14-Nov-2024.
[7] J. M. Ahn, J. Kim, and K. Kim, "Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting," *Toxins*, vol. 15, no. 10, Art. no. 608, Oct. 2023, doi: 10.3390/toxins15100608. [Online]. Available: https://www.mdpi.com/2072-6651/15/10/608

Acknowledgments

- UTSA CS 6463, EE 5993 with Dr. Miltiadis Alamaniotis
- FEDITC