Closing a Gap in Emissions Inventories: Using Machine Learning to Extrapolate Geographically-Limited Combination Truck Idling Data to a Wider Domain

Damian Vincent Dobrowolski

Department of Computer Science

Northwestern University

Advisor

Mohammed Alam, Ph.D.

An Undergraduate Senior Thesis (TR#: NU-CS-2026-23)

submitted to the Weinberg College of Arts and Sciences

in partial fulfillment of the Honors requirement for the degree of

Bachelor of Arts in Computer Science

April 15, 2026

Acknowledgements

I would like to express my sincere gratitude to my advisor, Prof. Mohammed Alam, Ph.D., for his continued support, encouragement, and smiles throughout this year-long project. Our weekly check-in meetings never failed to inspire me, lighten my mood, and provide that extra boost to keep me moving forward. I also thank both Prof. Daniel Horton, Ph.D., and Victoria Lang, Ph.D.-candidate, of the Climate Change Research Group for granting me access to this proprietary dataset and allowing me to continue the work that Lang started, shaping it into my own personal first research experience. I would also like to thank Prof. Connor Bain, Ph.D., for being a second reader on my approval committee.

The presentation of results from this research at the American Geophysical Union Annual Meeting 2025 (AGU25) was assisted by Conference Travel Grants from both the Office of Undergraduate Research, which is administered by Northwestern University's Office of the Provost, and from the Baker Program in Undergraduate Research, which is administered by Northwestern University's Weinberg College of Arts and Sciences.

The continuation of this research during summer quarter 2025 was enabled by the Summer Internship Grant Program (SIGP), administered by Northwestern Career Advancement (NCA). However, the conclusions, opinions, and other statements in this presentation are the author's and not necessarily those of the sponsoring institutions.

Abstract

Combustion engines are a major contributor of greenhouse gas (GHG) and air pollutant emissions, which each disproportionately impact marginalized communities. While just a small percentage of vehicles, combination trucks (CT) are one of the highest-polluting vehicle classes. Fueled by just-in-time delivery, the goods-transport sector has become the most rapidly-growing contributor to GHG emissions. To design effective remediation policies, a better understanding of CT emission processes is needed. One key CT uncertainty involves constraining the location and duration of CT idling emissions. At present, the U.S. EPA SMOKE-MOVES emission modeling platform spatially distributes CT idling emissions using a land cover development intensity spatial surrogate, which may miss idling hot spots. Here, to better constrain CT idling activity we leverage telemetry data from the seven-county Chicago metropolitan area. The telemetry data indicates that the current EPA methodology may distribute CT idling emissions too widely and may substantially underestimate CT idling emission magnitudes. While the telemetry data is revealing, it is also cost prohibitive to obtain for large geographic regions. Machine learning can help—by learning relationships between the telemetry-based idling data and other data sources (e.g., location of truck-stops, traffic, warehouses, intermodal terminals, etc.), we extrapolate our limited idling activity data over a larger geographic area. The telemetry-based idling data is zero-inflated and highly right-skewed. As such, we implement a two-part Gradient Boosting (XGBoost) model. This architecture is designed to handle both data characteristics: a classifier first predicts the probability of idling, and if positive, a second-stage regressor—trained only on non-zero data—predicts its duration and location. Our framework aims to better capture CT idling hotspots and magnitudes over wide geographic domains using only a small sample of observed idling data, with the ultimate goal of improving emission inventories and better constraining the impacts of CT idling emissions.

Table of Contents

List of Figures

List of Tables

1. Introduction

Combustion engines are a major source of greenhouse gases and air pollutants. One of the highest-polluting vehicle classes is the **long-haul combination truck (CT)**. CTs move freight nationwide, but their idling activity remains understudied compared to traditional on-road emissions. Fueling the goods-transport sector's status as the fastest-growing contributor to greenhouse gas emissions (Illinois Warehouse Boom), it is vital to understand the landscape of combination truck idling, as it is a direct player in these environment-affecting developments. We propose that there may be a gap in the current standard for evaluating such idling and thus their consequent emissions: the EPA's regulatory-grade model SMOKE-MOVES.

This model consists of two parts: MOVES (Motor Vehicle Emission Simulator) (Overview of EPA's Motor Vehicle Emission Simulator)—the model that estimates emissions from motor vehicles based on vehicle types, operating conditions (e.g., cold start, idle), road types, time periods, and pollutants—and SMOKE (Sparse Matrix Operator Kernel Emissions) (SMOKE)—the system that distributes emissions inventories (including the estimates from MOVES in the absence of point data) into gridded, speciated, hourly inputs for air quality models. To distribute the emissions inventories and MOVES predictions across a gridded area (Beidler), SMOKE employs a land cover development intensity spatial surrogate (Eyth), using land use development intensity as a proxy for where trucks are likely to idle, since the EPA does not have access to the data for every warehouse and loading dock in the National Emissions Inventory. This proxy comes from the National Land Cover Database, which assigns development intensity values such as low, medium, high, and combinations of these (National Land Cover Database) to grid cells as a way to weight the idling predictions made there.

However, SMOKE-MOVES may be missing and underestimating emissions from idling hotspots in less developed areas such as rural truck stops or smaller warehouses outside of urbanized areas. Comparing observed daily idling hours in the **seven-county Chicago metropolitan area (7-CMA)** (Cook, DuPage, Will, Lake, Kane, McHenry, and Kendall counties) from our telemetry dataset to those predicted by the 2016v2 platform running MOVESv3, external collaborator Victoria Lang, DEEPS PhD candidate, found 626,216 versus 44,937 hours, respectively, a staggering difference of a 13.9-times-greater magnitude of idling, and hotspot daily maxes of 16089 versus 4604 hours, respectively (see below).

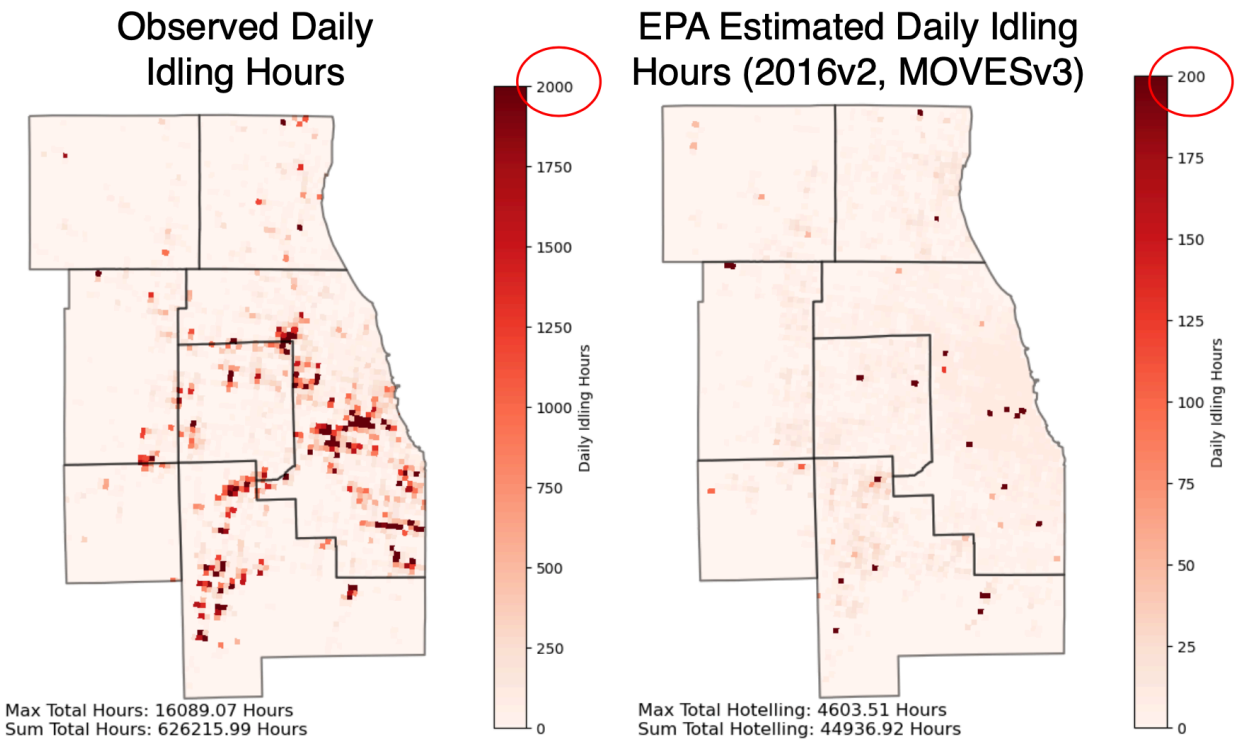

Figure 1.1. Potential EPA Idling-Estimation Gap (Victoria Lang)

To overcome this gap, why not just collect directly-observed idling data from combination trucks, as those owned by trucking companies are mostly already equipped with the

monitoring systems necessary to collect this information (Improving Vehicle Fleet)? Because there is a great obstacle here: obtaining the proprietary trucking logistics data is incredibly expensive—our dataset of telemetry activity in the 7-CMA alone is worth six figures.

The goal of our research is to overcome this obstacle and bridge the potential emissions gap by developing a specialized machine learning framework to learn patterns from a small subset of national idling telemetry data in order to test the potential for extrapolation of learned idling behavior and infrastructure relationships to a wider domain. Thus, our work could provide a systematic way for industry scientists and academic researchers to more accurately and more efficiently predict long-haul combination truck idling emissions for the purpose of improving emissions inventories. Either they could collect or obtain a small sample of regional idling, as we did, and apply our same methodology to predict idling throughout the wider area, or, if our model is already able to accurately predict idling far outside our study area, i.e., in different states or countries than Illinois, US, then they could simply input their own area's features in our model and receive a prediction. However, this latter approach seems improbable, as these areas most likely have very different infrastructure and data features than ours, i.e., the learned patterns our model discovered are probably not universally applicable.

2. Related Work

2.1. Two-Part (Hurdle) Model

In “Hybrid machine learning approach to zero-inflated data improves accuracy of dengue prediction” by Francisco et al., a similar problem to our project’s was present in the data which necessitated a specific modeling approach to improve regression accuracy. This problem was zero-inflation, as 70% of their data consists of zeros (comparable to the ~60% that our data has). They found that standard regression models trained directly on zero-inflated continuous data struggle to accurately predict extreme values, because of the biasing towards zero that occurs as a result. To overcome this, Francisco et al., implemented a two-step hybrid approach. First, a qualitative classification model transforms the data into a binary variable to predict the presence or absence of a case. If the qualitative (classification) model predicts the presence of such a case, then a second-stage quantitative (regression) model is used to estimate the magnitude of the case. By separating these tasks, the qualitative model effectively handles the data with numerous zeros, while the quantitative model can become specialized for estimating the continuous outcomes. Their experiments found that filtering out the zero predictions in the first step with such an approach improved the accuracy of the second-stage quantitative model. We implemented a similar two-step (hurdle) model to address the zero-inflation of our own data, based on the efficacy of this framework that this paper demonstrates.

2.2. XGBoost

Tianqi Chen and Carlos Guestrin, in their proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), “XGBoost: a scalable tree boosting system,” discuss the application of the XGBoost model. By building

shallow trees sequentially, where each tree is trained to predict the error (residual) of the previous trees, XGBoost has a greater capability to reach predictions towards extreme values. If the model severely under-predicts an event, the next tree focuses aggressively on that specific massive error. Because it then adds the predictions together, rather than averaging them like Random Forest (RF) and other popular models do, it has the possibility of "boosting" its way up to the exact extreme value. Furthermore, unlike older gradient boosting methods that only look at the direction of the error, XGBoost utilizes second-order approximation (incorporating both the gradient and the Hessian) to optimize the objective function. This allows the model to take much more precise steps toward extreme outlier values without overshooting or ceasing to make progress. Additionally, because outliers are inherently rare, the path down a decision tree to find an outlier is often sparse. To combat this, XGBoost uses a "sparsity-aware" algorithm that automatically learns optimal default directions to handle missing or sparse data (thus also working well as a classifier with our zero-inflated data). Because our data has extreme outliers and predicting them is one of our project's main focuses, XGBoost's ability to reach predictions up to the highest end of data values led us to implement it in our procedures.

2.3. Tweedie Loss

In the work of Anyidoho et al., "A machine learning approach for predicting hurricane evacuee destination location using smartphone location data," to improve regression prediction accuracy for their zero-inflated and highly right-skewed real-world spatial flow data on disaster evacuation patterns, they utilized a Tweedie distribution objective (loss) function, which is a versatile family of exponential dispersion models. It establishes a power-law relationship between the mean and the variance, $Var(Y) = \phi\mu^p$ (where ϕ is just a scaling/dispersion factor).

This means that the variance of the predicted target (Y) is expected to increase alongside the magnitude of the prediction, so the greater the magnitude of a data point, the greater the variance is expected to be (meaning the greater the error)—the trait of data having this property is called heteroscedastic. Here, p is a special "variance power" parameter: by setting different values for p, the Tweedie distribution morphs into familiar standard distributions—p = 0: normal (Gaussian) distribution (bell curve, allows negative numbers); p = 1: Poisson distribution (counts like 0, 1, 2, 3... but no decimals); p = 2: Gamma distribution (strictly positive continuous numbers, right-skewed, but cannot be exactly zero); p = 3: Inverse Gaussian (even more heavily right-skewed). Our implementation used the default p =1.5 for a distribution that is a mix of exact zeros and continuous positive numbers, i.e., a Compound Poisson-Gamma distribution. Because our hard-gating approach (see *Section 4.3.*) did not feed any zeros to the second-stage regressor that used Tweedie, the model behaved exactly like a Gamma regressor. However, when used with the soft-gating approaches (see *Section 4.3.*) that did train the regressor on zero-values and allowed for idling events with 0% expected-probability to be passed to the regressor, the Poisson part (p=1) of this compound distribution allowed for zero-handling that would not have worked with a pure Gamma distribution (because it doesn't allow for exact zeros), while simultaneously drawing a curved, right-skewed tail for the high-magnitude idling events. By way of this unique mathematical capability, Tweedie distribution is specially engineered to handle such heavy-tailed data with a large mass at zero as the data of both this research and our own project. Compared to standard parametric models, which usually rely on least squares or standard Poisson regression, they are not flexible enough to capture such severe nonlinearities as were present in our data because they assume a more normal distribution.

2.4. K-Means

The research done by Ren et al., "Prediction of Distribution Network Line Loss Rate Based on Ensemble Learning," provides an argument that traditional analytical methods often miss deeper data patterns, and supports an approach to mitigating this limitation through feature engineering by advanced preprocessing. The researchers used K-means clustering to find similar distribution patterns among different samples, simplifying the calculation process and increasing efficiency. Once the optimal number of clusters were formed, these classification labels were added as entirely new features to the sample data. Then using these clustering results, separate linear regression models were trained for each cluster to fit the specific features of that data class. In our future work, we plan to implement a similar K-means clustering approach for feature engineering to increase the model's understanding of the data landscape by supplementing features based on arbitrary county boundaries with custom groupings based on idling profiles. In finding these homogenous groupings that can better convey the reality that the data is drawn from, similarly to Ren et al., who state that adding cluster labels as new features provides the model with richer feature sample data, our idling groups may provide an improved contextual understanding of the region as a supplement for data sparsity. By identifying some amount of idling categories based on the optimal tradeoff between intergroup variance and computational cost, we would follow their systematic approach of optimizing hyperparameters to achieve this balance. Their paper found that this integrated methodology surpassed previous predictions; if our integration of this approach found that idling groups became more important than county as a feature, it would align with their conclusion that data-driven patterns are superior to traditional indicators.

3. Data Analysis:

3.1. Dataset Context

Our project is enabled by a proprietary trucking logistics dataset with a six-figure price tag acquired by Northwestern's Climate Change Research Group. This spatiotemporal data of combination truck idling collected from August 2018 to April 2019 in the 7-CMA has two parts that act as our dependent variables and is the target of our prediction models. Including both **hot idling**, defined as occurring for less than thirty minutes (while the engine is still hot), and **cold idling**, which includes anything longer than thirty minutes, we aim to predict the location and magnitude (duration) of these two forms of idling. The EPA's SMOKE-MOVES uses **ONI (off-network idling)** to refer to hot idling and **hotelling** to refer to cold idling. Cold idling is much more prevalent in our study area, and also more harmful. Since it occurs after thirty minutes of idling, when the engine has already cooled off, this cold engine becomes much less efficient at burning fuel as well as powering the various "aftertreatment systems" (which minimize certain pollutants) installed in the truck. Thus, cold idling results in greater emissions and higher emission rates than hot idling (Martin).

3.2. Preliminary Sensitivity Simulation

This prevalence of hotelling (cold idling) is shown by a preliminary sensitivity simulation that Victoria Lang conducted by updating MOVES-estimated idling hours via interpreted telemetry data and revising the spatial distribution to reflect the telemetry-based spatial footprint. Using the 2016v2 platform, which incorporates MOVESv3, an annualized comparison was performed (August/October 2018, January/April 2019) which shows a change from 22.15 to 360.22 metric tons/yr of fine particulate matter with diameters of 2.5 micrometers or smaller

(PM2.5) when comparing expected air-pollutant emissions from MOVES-predicted versus observed annual *hotelling* (cold) idling hours, an increase of 1626.28%. When examining expected PM2.5 emissions from MOVES-predicted versus observed annual *ONI* (hot) idling hours, a change from 66.33 to 114.25 metric tons/yr of PM2.5 appears, or an increase of 172.24%. Whereas baseline annual MOVES-predicted hotelling and ONI idling are expected to result in 22.15 and 66.33 metric tons of PM2.5/yr, respectively (essentially a third as much from hotelling than ONI), these expected metric tons of PM2.5/yr for hotelling and ONI idling change to 360.22 and 114.25 when being drawn from observed idling, respectively (more than three times as much from hotelling than ONI, completely flipping the relation), indicating hotelling's greater expected impact on total air-pollutant emissions and greater change in its own contribution thereto, when estimated through imputation from our observed-idling data. Coupling this with the vastly greater amount of general idling hours contributed by hotelling compared to ONI: 540,411 vs. 85,805 hours, respectively (see below), it shows the importance of investigating hotelling's specific idling hours (and thus emissions) contribution, especially as its expected PM2.5 contributions had such a drastic, sixteen-fold increase from those using baseline EPA model idling estimates to those using observed idling.

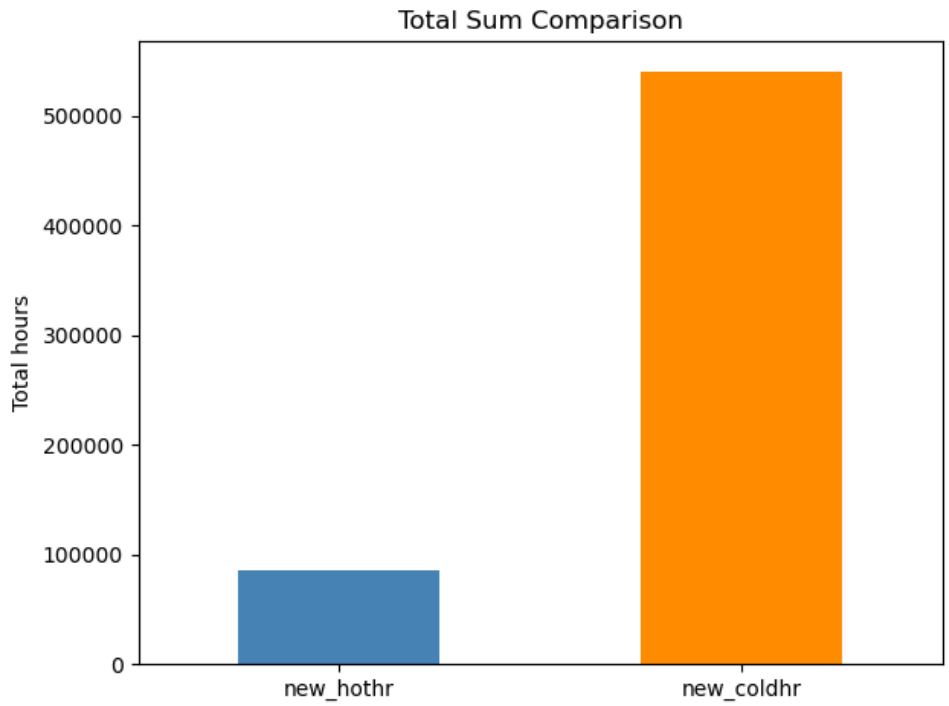

Figure 3.1. Hot vs. Cold Idling Prevalence

3.3. Targets Statistical Analysis

Our dependent variables (targets) of hot and cold idling are very zero-inflated, consisting of ~60% zeros, and highly right-skewed, with skew values of 9.5 for hot and 13.5 for cold—for context, statisticians consider any value greater than 1.0 to be "highly skewed," so these values are *extremely* skewed (see *Figure 3.3.*). The observed hot and cold idling data from our 7-CMA study region has 6,387 “instances” of idling (although hot idling was 58.32% zeros and cold idling was 59.09% zeros). The average length of time of an instance of hot idling is 14 hours with a standard deviation of 55; comparatively, a cold idling event has a mean of 91 hours with a standard deviation of 461. The respective maximum hot and cold idling events are 1,334 hours and 14,755 hours (see below).

	new_hothr	new_coldhr
count	6387.000000	6387.000000
mean	14.364078	90.645235
std	55.445835	461.264761
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	5.422600	13.807271
max	1334.122079	14754.951763

Figure 3.2. Targets Descriptive Statistics (Hot/Cold Idling)

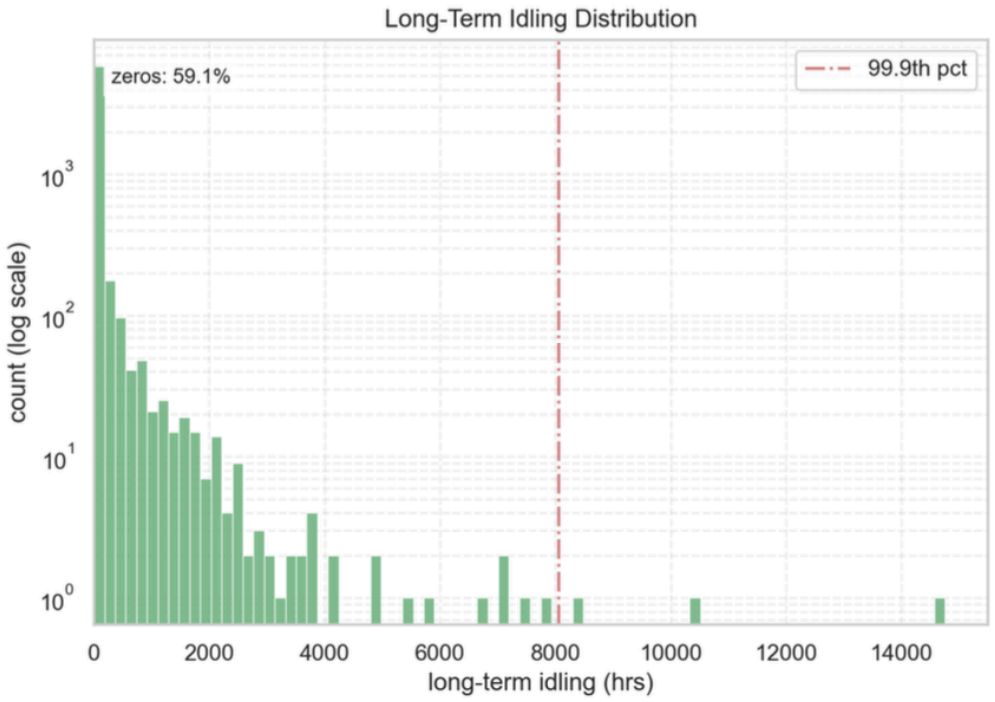

Figure 3.3. Distribution of cold (long-term) idling hours

3.4. Public Features Data Sources Overview

As independent variables (features) to be input into our model so it can identify patterns between these and the targets (idling) that it outputs predictions for, we use publicly-accessible

government data on warehouses, truck stops, intermodal terminals, traffic volume, and land use. These specific data sources were chosen because of their general availability and because they represent relevant infrastructure for the freight-trucking industry, like where loading, resting, and other idling-causing activities occur. The following table presents a brief overview of each of these data sources based on the category of data they represent:

Data Category	**Dataset Name**	**Primary Source**	**Compilation/Update Date**	**Key Features Used**
Logistics Infrastructure	Truck Stops	FHWA / CMAP	2019 / 2024	Parking capacity (*number_of_spots*)
	Intermodal Terminals	CMAP	2021 / 2024	Pipeline/Rail capacity (*total_rail_terminal_capacity*)
	Freight Rail	CMAP	2003 / 2025	TOFC/COFC facility types and transport modes
Built Environment	Land Use Inventory	CMAP	2020 / 2025	57 land-use categories for northeastern Illinois
Road Connectivity	HPMS Highways	FHWA / CMAP	2020	Road functional classes (1–7) and urban codes
Demographics	Gridded Population	NASA / CMAP	2020	Population counts and density distributions

Table 3.1. Public Data Sources Overview

Logistics Infrastructure: Datasets for truck stops, intermodal terminals, and freight rail were used to identify "idling affordances"—locations where trucks are physically likely to idle.

Specifically, the inclusion of Trailer-on-Flatcar (TOFC) and Container-on-Flatcar (COFC) data allows the model to differentiate between different types of freight transfer intensities.

Built Environment: CMAP Land Use Inventory was the foundation for spatial context. With 57 distinct categories, this dataset allows the model to distinguish between "Industrial/Warehouse" zones (higher idling potential) and "Residential" zones (lower idling potential).

Road Connectivity: the Highway Performance Monitoring System (HPMS) provides the "circulatory system" of the model. By including functional classifications—ranging from Interstates (1) to Local roads (7)—the model can weight idling probability based on the road's intended volume and connectivity.

Demographics: the NASA Gridded Population data was used to incorporate the human element. High population density cells might represent areas with higher delivery demand but also areas where idling emissions have a more significant public health impact regarding PM2.5 exposure (see *Section 3.2.*).

Finally, these raw shapefiles were spatially joined to our 6,374 grid cells. Raw features from the CMAP and NASA datasets were aggregated into a unified grid structure, where each cell was assigned a 124-dimensional feature vector. This process involved converting string-based capacities into numeric values and cleaning data such as negative shapefile polygons through zero-buffering to ensure statistical validity.

3.5. Features Statistical Analysis

This public government data used to provide infrastructural, environmental, and demographic context and use as features to try to predict our idling hours targets was even more

zero-inflated and right-skewed than the target data, with 76% of feature rows consisting of more than 90% zeros and having an average skew of 24.53 and median skew of 19.67 (compared to the hot and cold idling target's skew of 9.5 and 13.5, respectively) and 96.2% of features having a skew above 3.0 (see below).

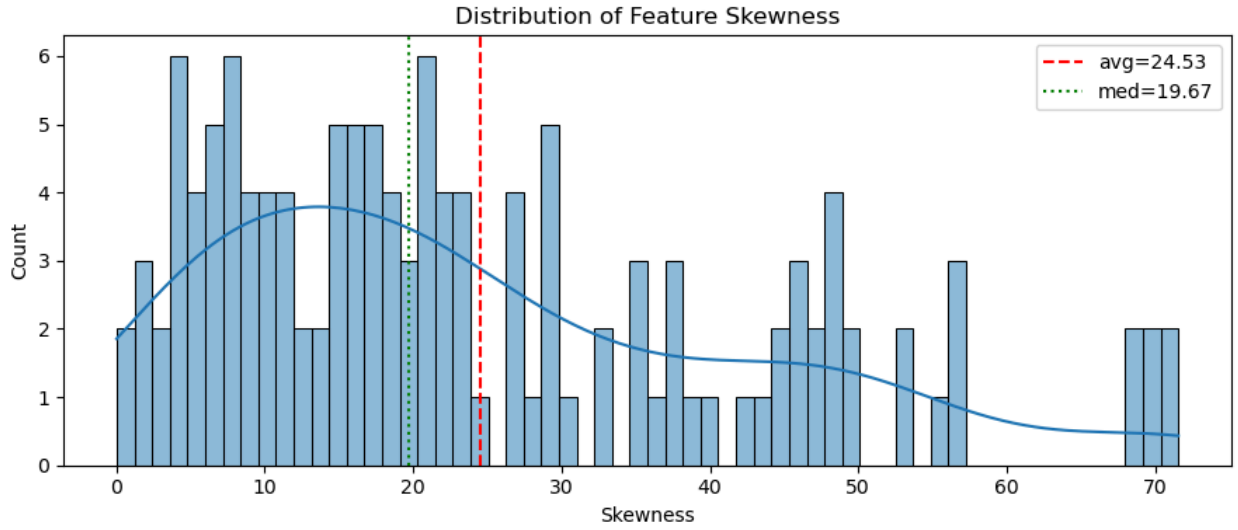

Figure 3.4. Feature Skewness

A preliminary Pearson correlation analysis (see below) showed that the features with the highest correlation with both idling targets hot idling (ONI) *and* cold idling (hotelling) were *total_RBA* (total Rentable Building Area, essentially square footage, of warehouse), *total_loading_docks* (of warehouse), and ***1432_area*** **(designating warehouses with area over 100,000 ft^2)**.

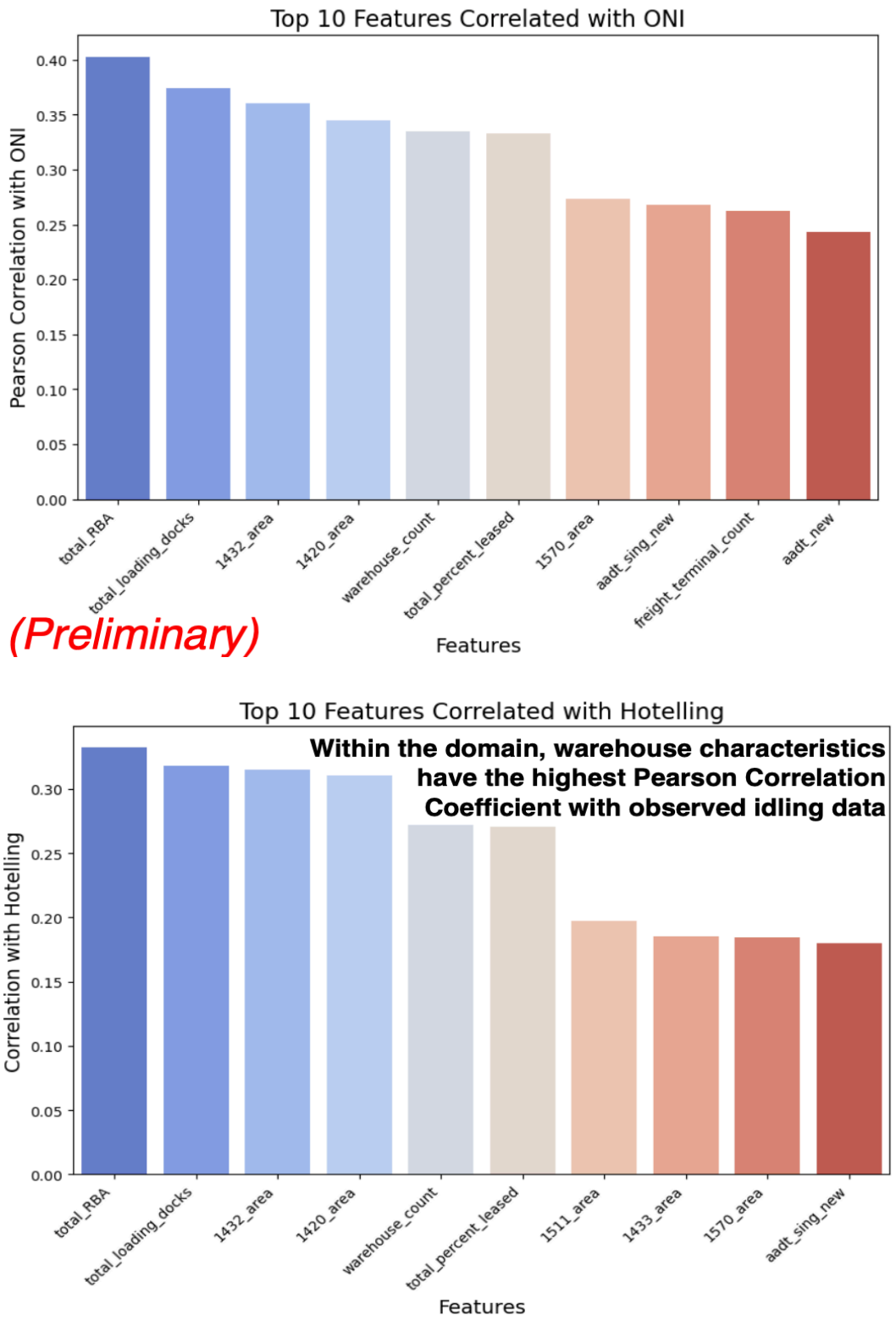

Figure 3.5. Top Idling-Correlated Features (Victoria Lang)

Comparing the seven counties of the 7-CMA by computing Euclidean distance of key predictors of idling, discovered through feature importance (see *Section 5.3.*), as well as idling magnitudes (targets), all stratified by county lines, are summarized in the chart below. What we gather from this exploration is that there is a great difference between counties, according to key predictors of idling and idling quantities themselves, with some, such as Kane vs. McHenry and Kane vs. Kendall, having small Euclidean distances near 1.0, while most pairings of counties are comparatively very different according to these metrics: 15/21 pairings have a Euclidean distance greater than 4.3, indicating that their idling landscapes differ.

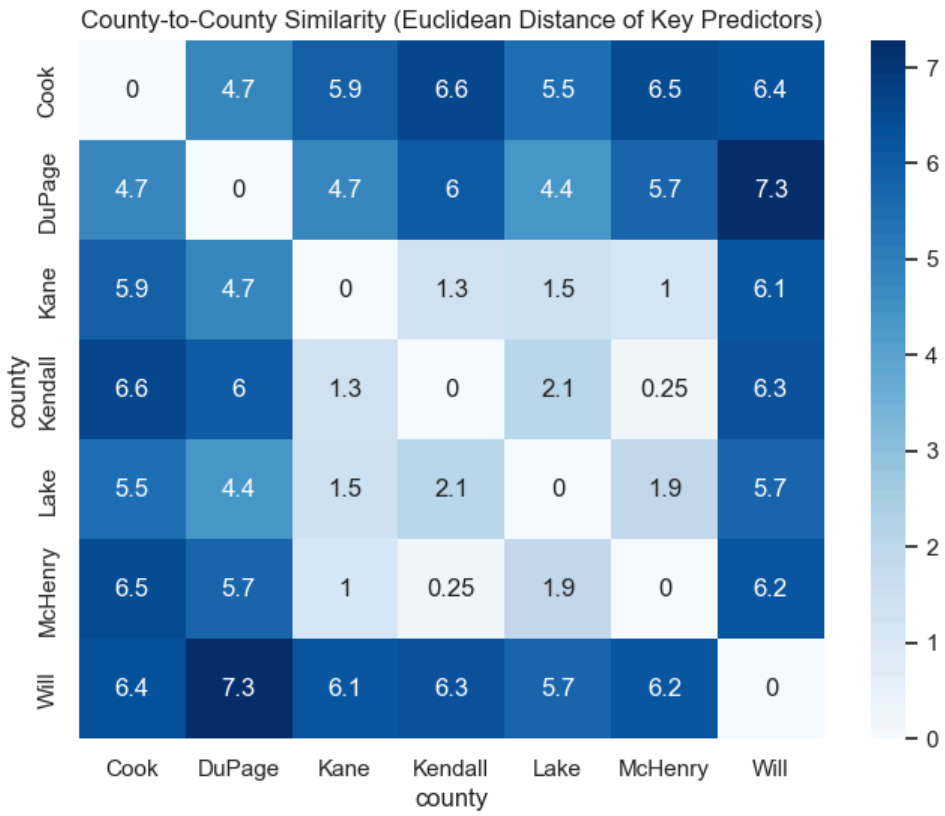

Figure 3.6. County-to-County Similarity

4. Methodology:

4.1. Data Preprocessing

To establish a robust foundation for modeling, raw and standardized features were combined with target variables into a centralized dataset, while ensuring grid-cell alignment. Initial data cleaning involved a systematic analysis of missing data: specifically, in comparing the percentages of zeros and NaN in each row, since both were present. We identified that rows had either only zeros (and no NaNs) or only NaNs, leading us to believe that the NaNs in those rows likely represented zeros rather than missing entries, which would have been a concern, especially given the high percentage of missing values, as it would have indicated that NaNs were potentially missing values that would have to be removed or imputed. Because this was not the case, we decided to convert NaN entries to zeros for feature variables to ensure compatibility with machine learning approaches. This was followed by the removal of zero-variance features to eliminate variables that provided no predictive information gain. To address potential multicollinearity and preserve the integrity of subsequent feature importance analysis, Variance Inflation Factor (VIF) filtering was applied to remove redundant predictors. However, final aggregated files of both VIF-filtered and non-VIF-filtered data were saved to compare which approach resulted in better results: the VIF-filtered version resulted in an early jump of r^2 from 0.05 to 0.11. The final stage of preprocessing involved managing the dataset's distribution to ensure consistency between subsets. A stratified train/test split was implemented to preserve the ~60% zero-inflation characteristic inherent in the idling data. Additionally, skewness drift was monitored throughout the process to verify that the training and testing distributions remained representative, thereby mitigating the risk of biased model predictions.

4.2. Jumping Hurdles

Our machine learning architecture consists of a two-part (hurdle) model designed to address the challenges of high degrees of zero-inflation and right-skew in the target and feature data. Motivated by the hybrid model approach employed by Francisco et al., by first using a classifier to predict whether or not an idling event is likely to occur, then using the second-stage regressor to predict the duration of this excepted idling event, we are overcoming the hurdle of having ~60% of the target data at zero. Without this split approach, i.e., using only a regression model, there would be a great influence from the mass at zero which would significantly drive down the magnitude predictions. Through this hurdle architecture, we are instead able to train the regressor exclusively on non-zero ground-truth values, maintaining a system where it can solely focus on identifying the extent of idling magnitude only when idling *does* occur.

4.3. Entering Soft/Hard Gates

Our final implementation utilizes this hard(-inference)-gating approach just described, which was reached after also trying a soft-gating ("probabilistic weighting") approach where the regressor is trained on all data (including zeros) and the classifier predicts a percentage of idling likelihood (rather than a 0/1 binary) which the regressor's magnitude predictions are then weighted by to dampen durations for idling events whose occurrence the classifier doesn't expect to be very likely (Sieminski). Soft-gating trades "safer" predictions (by not assigning the same weight to all places of expected idling, since our classifier does not have 100% accuracy) in favor of less accurate regression estimates, because the regressor is trained on all zeros as well.

In addition, a combination of the two in a threshold-based ("curriculum learning") approach was considered, in which only those idling events beyond a certain threshold of

classifier-predicted idling likelihood percentage are passed to the regressor for its training (Bengio). This preserves a greater level of "safety" of predictions by adjusting them based on the expected chance of idling, while balancing this with a reduction in the amount of zeros that the regression is trained on, including only those observations where the classifier identifies a likely case of idling. This effectively gates the learning phase, allowing the regressor to specialize on high-magnitude features without the statistical bias of the zero-inflated majority.

Ultimately, we settled on the hard-gating approach because of our project's focus on extreme outlier detection. Soft-gating suffers from a risk we term *double damping*, which is the twice-occurring reduction of regression-predicted values: firstly by the regressor's training on all values (~60% zeros) which would result in lower estimations and secondly by the weighting of these already damped predictions by the classifier-predicted probability [0.0, 1.0] that further brings down the final idling magnitude estimates. A threshold-based soft-gating would still suffer from damping, because the final output is still scaled by the classifier's probability at inference (providing a "safety" damping during inference), but with the big advantage over soft-gating that the regressor itself is no longer "learning-damped" by being trained on all the zero-idling data points. Conversely, using hard-gating to ensure that the regressor will be trained on purely non-zero ground-truth values, and not weighted by any classification probability (because the hard-gated classifier outputs either 0 or 1), we enable it to have a greater chance to predict extreme idling magnitudes, as it is no longer influenced by the majority mass at zero, nor dampened by "classification-safety," and can instead learn the patterns solely applicable to positive idling instances. While this results in our regressor being trained on less overall data, something that is not ideal, we decided that for the purposes of our study in trying to estimate the extent of CT idling hotspots and where they occur, this was a worthy trade-off. So we finally

settled on a hard-gating approach due to its theoretical advantage for our aims, hoping that future work might take our insights further to consider the potential importance of gating approach in mixed classification/regression hurdle models.

4.4. XGBoost-ing

To address the high right-skew of our target and feature data, we select XGBoost (Chen) as our model for both classification and regression and tune it with the Tweedie loss function (Anyidoho). When an initial Random Forest regression model was implemented, total idling hours across our 7-CMA study region were predicted accurately (total idling—RF vs. observed: 86,768 vs. 85,805 hrs), however, the max idling hours within a single grid cell was below observed values by a factor of 4 (max idling—RF vs. observed: 334 vs. 1,334 hrs). Because of RF bagging data to train all trees in parallel and then averaging the results, it spread the predicted idling across the region of study, failing to identify the magnitude of specific extreme idling hotspots.

For our purposes, identifying where the highest idling hours were occurring in these specific hotspots was a priority of the study. Since idling is a localized event that happens in a stationary position, the emissions from it are also localized, thus highly affecting local populations where it occurs. By predicting the extreme outliers of our target telemetry data and where they occur, we would improve emissions inventories that currently lack such a high-resolution picture of idling emissions. The gradient-boosting that is central to the XGBoost model trains smaller models iteratively, with each successive one learning off of the errors of the previous epoch. In doing so, it is possible for the model to extrapolate from the data and predict

values at the high end of the range that it has previously seen, compared to RF's averaging that cannot possibly achieve this.

4.5. Tweedie Bird

Using the Tweedie objective, it tweaks the loss function to deal with data that is specifically heavily right-tailed with a large mass at zero, rather than the normal distribution assumed by most machine learning models (Anyidoho). Combining the hurdle framework using hard-gating with XGBoost using Tweedie loss, we were able to design a modeling approach that met the needs of our project and the data we had available to us. Then, the models were run with an 80/20 train-test split and hyperparameter-tuned using five-fold cross-validation (with RandomizedSearchCV) with twenty iterations. However, just because all of these tools improved the modeling prospects doesn't mean we were able to output perfect regression prediction on data with average skew values of 9.5/13.5 and 20 for targets (hot/cold) and features, respectively.

4.6. Interaction Features

By creating new features that consisted of the interaction between a combination of top-contributors in the feature importance analysis, we were able to again provide the model with more information, since the data was not as abundant as would have been ideal (Thakur). While interaction features can give better predictive power to a model, they naturally restrict real-world interpretation of their effects, because while the algorithm understands what *1432_area* X *1650_area* means, for the purposes of our post-predictive analysis, it is not clear how to extract our own understanding from this feature. Multiple interaction features ended up as top important features in our later analysis, indicating that their creation improved our model (see *Table 5.6.*).

4.7. Love Thy K-Nearest Neighbors

To improve our prospects further and deal with the limited amount of data, specifically in the smaller counties of our seven-county area, we engaged in feature engineering to create spatial neighbor context features and interaction features. To provide more spatial context to the model in dealing with our sparse input data, we used K-Nearest Neighbors to generate additional data features that were informed on what was happening in the cells neighboring each grid cell (Wang). We were able to calculate spatial lag features—such as the average idling duration of the immediate neighborhood—to directly capture the local spatial autocorrelation present in the dataset, and made them into features that were added to these grid cells, so that the model gained a wider picture of the area just from looking at a single grid cell. This gave it greater insight into the surrounding idling (target) and infrastructure/traffic (features) landscape, like playing sudoku starting with more numbers filled in.

4.8. Single vs. Per-County Modeling

The extensive exploratory data analysis that was done to learn about the data—its qualities, its limitations, and the approaches to modelling and specific models yielding accurate predictions that it would be most conducive to—revealed that there was significant spatial heterogeneity, specifically between the seven counties of the study area, as computed based on identified key predictors (features) and magnitudes (targets) of idling (see *Figure 3.6.*). A per-county modeling approach was attempted, training and testing a separate model for each county in an attempt to best capture the landscape of each county's respective idling. Initially, it seemed these models performed poorly when assessed by the r^2 correlation coefficient because

each was limited to only the data of each's county (see *Table 5.4.*). However, a later analysis revealed that by comparing them to the single model's results stratified by county, rather than the single model's global metric, the per-county models actually seem to outperform the single model (see *Table 5.5.*).

More broadly, in carrying out this modeling experiment, the importance of spatial context was seen to be unignorable. With vastly different warehouse infrastructure, land use, and other situations measured by the public government data making up each county's independent feature variables, it remains a question whether the sparsity of model data reached by assigning a model to each individual county is a greater impediment to prediction accuracy than the spatial heterogeneity that results in difficulties applying global (total) area trends to specific counties, especially those with less data (and thus a worse understanding by the model of each county's intricacies). By continuing with the single model—that was at the time seen as obviously outperforming the per-county approach—but also wishing to improve the model's vision of the per-county idling situation, our kNN feature creation (see *Section 4.7.*) attempted to address this spatial heterogeneity issue by providing additional spatial contextualization of the landscape of idling magnitude and contributors.

5. Results:

The top performing model iteration was the single MultiOutputRegressor (one model for hot idling and one for cold idling, metrics averaged), XGBoost hard-gated hurdle model with Tweedie objective function, using the raw, cleaned, VIF-filtered dataset with *county* feature added, spatial feature engineering from kNN, interaction features (from most important features according to previous XGBoost models), non-important features removed (dimensionality reduction), 80/20 training/test split, 20 iterations of 3 cross-validations with RandomizedSearchCV, and the following hyperparameter options, of which the bolded values were ultimately selected: *estimator__n_estimators*': [100, **200**, 300], *estimator__max_depth*': [**3**, 5, 7], *estimator__learning_rate*': [0.01, **0.03**, 0.05, 0.1], *estimator__subsample*': [0.7, **0.8**, 0.9, 1.0], *estimator__colsample_bytree*': [0.7, **0.8**, 0.9, 1.0], *estimator__reg_alpha*': [0, 0.1, **1.0**], *estimator__reg_lambda*': [0.5, **1.0**, 2.0].

5.1. Classification

Classification (stage one) results of the just-described model:

	precision	**recall**	**f1-score**
0 (no idling)	0.86	0.87	0.87
1 (idling)	0.81	0.80	0.82

Table 5.1. Classification Metrics

(Precision refers to what percentage of classifications that it did make were correct (success at avoiding false positives), and recall refers to what percentage of total observed no-idling or idling events (all those it should ideally have recognized) were actually correctly identified

(success at avoiding false negatives). F1-score is the harmonic mean of precision and recall scores).

These results indicate that the model is correct that there is no idling 86% of the time (precision), and recognizes 87% of actual no-idling events as having no idling (recall). Whenever the model says there is idling it's correct 81% of the time (precision), and it recognizes 80% of actual idling events as having idling occurring.

5.2. Regression

While classification was very successful, predicting how long these idling events would last for data as extremely right-skewed and zero-inflated as ours proved difficult. The same model got the following regression (stage two) results with averaged metrics of hot and cold idling models (by maintaining separate regression pathways for hot and cold idling with MultiOutputRegressor and reporting the final performance as composite scores):

r^2	**Spearman**	**wMAPE** (weighted Mean Absolute Percentage Error)	**MAE** (mean absolute error)	**MedAE**
0.38	0.64	177.6%	93.28	25.35

Table 5.2. Single Model Global Regression Metrics

While the r^2 correlation isn't very high, with 38% of the variance of the target idling hours being explained by the features, the Spearman correlation is quite high, indicating a strong,

positive monotonic relationship between idling hours and our public data features—i.e., the model is able to rank the ordering of idling events according to magnitude quite well. Additionally, the difference between MAE (93.28) and MedAE highlights the extreme heterogeneity of the dataset. While the wMAPE of 177.6% appears high, it is a direct mathematical consequence of the extreme right-skewness—where a single outlier in the 14,000-hour range carries more weight than thousands of zero-value cells. By presenting these metrics, we demonstrate that the model maintains high fidelity for typical urban cells, especially in classification, while struggling with the unpredictable magnitudes of rare, ultra-high-idling hotspots. We observe this high fidelity for typical urban cells by MedAE remaining significantly lower than MAE and being relatively close to zero, indicating that the model more accurately characterizes the zero-inflated majority of the study area (which is also shown by its classification prowess). Furthermore, a global Spearman correlation coefficient of 0.64 and county-level Spearman correlations of up to 0.78 (see next chart) demonstrate that the model effectively ranks geographic hotspots, even when the absolute magnitude of extreme outliers proves difficult to quantify precisely.

Analyzing this model's regression results by county (and stratifying them by hot and cold idling) paints a much different picture (only r^2 and Spearman are shown here for simplicity):

County	# of Events	Cold r^2	Cold Spearman	Hot r^2	Hot Spearman
Cook	232	0.29	0.66	0.24	0.66
DuPage	82	0.34	0.78	0.47	0.75
Will	76	0.44	0.54	0.4	0.59
Lake	53	0.08	0.33	-0.32	0.29
Kane	49	0.38	0.57	0.45	0.6
McHenry	25	-0.06	0.34	0.06	0.3
Kendall	15	-0.23	0.48	-0.12	0.68

Table 5.3. Single Model County-Stratified Regression Metrics

This closer look indicates that the model performance is much more variable depending on which county the idling event whose magnitude was predicted occurred in. We see r^2 values range all the way from -0.32 to 0.47 for hot idling, and Spearman values range from 0.34 to 0.78 for cold idling. It seems that the data sparsity may be a major factor limiting the efficacy of modeling, as in this cleaned, VIF-filtered dataset with only the most important features left in, the lowest counties have only 15 and 25 idling events. The table is ordered by number of idling events to show how it may be related to the poor performance in certain counties, however, Lake's hot r^2 of -0.32 stands out as much lower than the two counties with the closest number of idling events to it, Will and Kane, which have hot r^2's of 0.4 and 0.45, respectively, meaning it is not simply an issue of data sparsity.

County	# of Events	Cold r^2	Cold Spearman	Hot r^2	Hot Spearman
Cook	1125	0.04	0.65	0.06	0.71
Kendall	433	0.17	0.62	0.26	0.65
DuPage	386	0.4	0.62	0.03	0.64
Kane	290	0.05	0.55	0.02	0.66
Will	232	0.24	0.67	0.49	0.65
McHenry	123	0.21	0.32	0.19	0.45
Lake	73	0.25	0.46	-0.01	0.44

Table 5.4. Per-County Model Regression Metrics

The above table shows the regression metrics of the earlier per-county modeling attempt (see *Section 4.8.*), instead of the single model regression results shown stratified by county, as in the previous table. This attempt also utilized soft-gating (probabilistic weighting) (see *Section 4.3.*). Notice that this earlier experiment has far greater amounts of idling events for each county, because the dimensionality reduction from feature importance results had not yet been implemented. The below table shows a comparison of the means, standard deviations, and coefficients of variance (which is a measure of the amount of "noise" relative to "signal") between the regression metrics for the single (1) and per-county (2) models. What's significant is the difference in means coupled with the correlation of variation (CV) % change: as more data points were available to model 2 (per-county)—evident by the far greater mean number of events in model 2 than model 1—there was a 54.53% *stability gain* in Cold r^2 and 43.98% stability gain in Hot Spearman. Stability gain refers to a decrease in CV, meaning a decrease in noise over

signal (σ/μ), therefore a reduction in uncertainty. This seems to actually be an indication that while the single model's performance appeared to be much stronger than the per-county models' when looking at the single r^2 and Spearman in *Table 5.2.* (because of the global metrics' reliance on high-event counties), when stratifying the single model's results by county and target (*Table 5.3.*) and comparing these to the per-county results (via the table below), it seems the earlier per-county model actually performed better, according to the stability gain (CV % change). However, since many factors changed between the single model and this earlier per-county model, further work is needed to reexamine this modeling approach and conduct more experiments to isolate the factor of model architecture to find out if it is the deciding factor in this "improvement."

Model	**Single (1)**			**Per-County (2)**			
Variable	**Mean 1 (μ)**	**StdDev 1 (σ)**	**Coeff of Variat 1 (CV)** (σ/μ)	**Mean 2 (μ)**	**StdDev 2 (σ)**	**Coeff of Variat 2 (CV)** (σ/μ)	**CV % Change**
# of Events	76	72.97	0.96	380.29	353.12	0.93	-3.29%
Cold r^2	0.18	0.25	1.39	0.19	0.12	0.63	-54.53%
Cold Spearman	0.53	0.16	0.3	0.56	0.13	0.23	-23.10%
Hot r^2	0.17	0.31	1.82	0.15	0.18	1.2	-34.19%
Hot Spearman	0.55	0.18	0.33	0.6	0.11	0.18	-43.98%

Table 5.5. Single vs. Per-County Model Regression Metrics, Statistical Comparison

Ultimately, while the single model seems to outperform the per-county model when comparing its global metric (*Table 5.2.*) to the per-county model's mean metrics (above)

(especially global r^2 of 0.38 vs. mean cold r^2 of 0.19), at the same time, the single model seems to perform worse than the per-county models when comparing both mean metrics within counties (see CV % changes in table above). However, this is a result of the global metric being heavily weighted by high-volume counties like Cook (n=232) and Will (n=76), where the model demonstrates high fidelity. Conversely, the lower unweighted (not global) means of the county-stratified views compared above reflects the inherent challenges of predicting idling magnitudes in data-sparse counties. While the global score of the single model highlights its effectiveness as a regional tool, the per-county analysis provides a more transparent view of its own better spatial consistency across the heterogeneous 7-CMA geography. Essentially, the per-county model identifies idling with more consistent accuracy across all seven counties, rather than having the dominant signal of the urban center overpower the unique idling patterns of the smaller, collar counties.

There remains one glaring disparity between the single model's county-stratified results and the per-county model's results: number of events per county, specifically, the relative order and amount of each county between the two models. Most evidently, Kendall goes from having the least amount of idling events in the single model to second-most amount in the per-county model. 5/7 counties, except for Cook and McHenry, changed their index in this ranking. This may be the result of the single model's hard-gating compared to the per-county model's soft-gating. With the hard-gated baseline, event counts were suppressed in data-sparse counties like Kendall, where lower classifier confidence resulted in a high rate of rejected observations and thus low numbers of idling events. In the soft-gating approach, preserving the probabilistic spectrum of the classifier allowed for a more representative accounting of regional idling activity, revealing that the lower number of idling events in collar counties may be a byproduct

of classifier uncertainty rather than a physical absence of idling. Therefore, experiments controlling these gating approaches should also be added to further work in order to try to make definitive conclusions about which yields better results, alongside the model architecture factor controls mentioned earlier in this subsection.

5.3. Feature Importance

To interpret the decision-making logic of the single XGBoost model, we utilized two distinct feature importance metrics: Information Gain and SHAP (SHapley Additive exPlanations) (Koç). The Gain scores represent the actual improvement in accuracy (the reduction in loss/error) contributed by each feature, identifying the variables most critical for model accuracy as a global, model-centric metric. Alternatively, the SHAP summary offers an instance-centric, game-theoretic attribution of how each feature shifts individual predictions from the baseline. The consistent ranking of *1432_area* (warehouses with area over 100,000 ft^2) as the primary predictor across both metrics supports its role as the dominant driver of idling activity within the 7-CMA. Generally, these land use categories (of which *1432_area* is one)—which are designators of land types based on what can be built on the land and what the land can be used for (see *Section 3.4.*)—had the greatest contribution to idling magnitude, making up all but one (the second-highest SHAP value) spot throughout both Gain and SHAP results. Their prevalence may be the result of VIF-filtering or other feature importance cleaning which kept only the most explanatory features (see *Section 4.1.*). Compared with Lang's preliminary Pearson correlation analysis (see *Figure 3.5.*), the top two correlated features *total_RBA* and *total_loading_docks* may have been "explained away" by these area codes, so the model would have cut them from the final feature vector for dimensionality reduction. In the

same figure, we see many area codes as the highest-correlated features, so they were expected to contribute greatly to feature importance in any case.

In the Information Gain feature importance table below (which is normalized so that the summed gain of all features adds up to 1.0), we see that similar features had the greatest weight on the model output between hot and cold idling, with the same two being the first- and second-most influential.

Hot Idling		**Cold Idling**	
Feature	**Gain** (normalized)	**Feature**	**Gain** (normalized)
1432_area	0.08	*1432_area*	0.07
1432_x_1420	0.07	*1432_x_1420*	0.05
4130_area	0.05	*1420_area*	0.04
sqrt_1570	0.04	*1432_div_1570*	0.03
1420_area	0.03	*4130_area*	0.03

Table 5.6. Feature Importance by Information Gain

The following SHAP summary, which provides a comprehensive global view of feature importance and the distribution of feature effects across a dataset, shows which features have the greatest effect on final target prediction (in this case cold (long-term) idling), and also how their values impact predictions. *1432_area* is also the top feature by SHAP value. However, the second most-important feature (*lat*, which represents latitude) demonstrates a major difference in methodology between Information Gain and SHAP and represents a classic machine learning phenomenon: *lat* might not solve the whole model's error (because it's a continuous number, so the model has to split it many times in tiny increments to get results, and since no single split on

latitude fixes a huge chunk of the error at once: low Gain), but for individual cells, it is a critical feature acting as a modifier that determines if that cell is in a high-idling area (thus high SHAP). This high-SHAP-value *lat* feature also shows that spatial position of idling has a great influence on final idling magnitude, meaning spatial heterogeneity is important.

Features are listed vertically on the y-axis, ordered by their global importance (mean absolute SHAP value) from top to bottom. The x-axis represents the SHAP value, where points to the right indicate a positive impact (increasing the prediction) and points to the left indicate a negative impact (decreasing the prediction). The color coding reflects the original value of the feature for each instance (blue for low values and red for high values), showing correlations between feature magnitude and impact direction. Each point on the plot corresponds to a specific instance in the dataset, with points stacked vertically when they overlap to show the density of the data at that SHAP value, revealing whether high or low feature values consistently drive the model's output up or down.

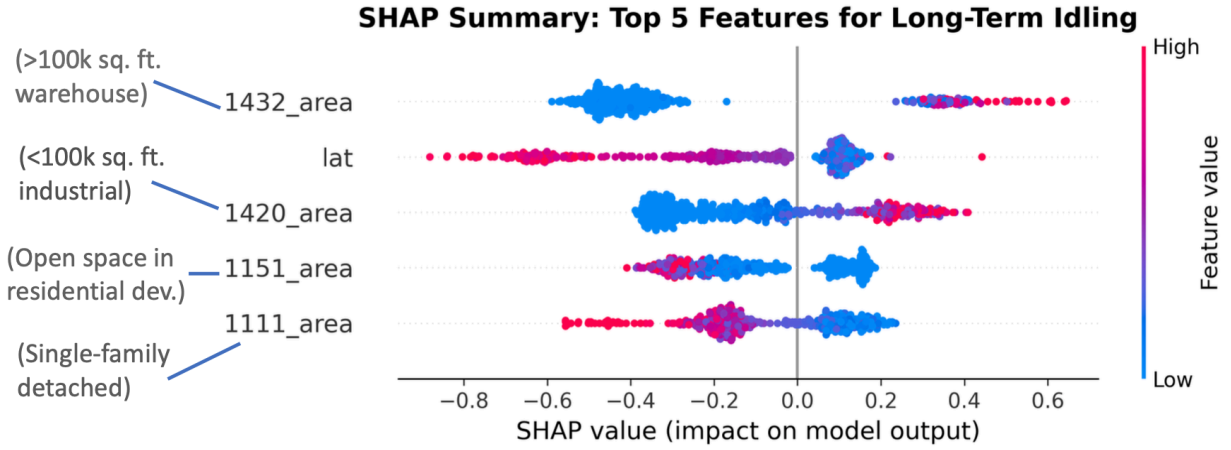

Figure 5.1. Feature Importance by SHAP

For instance, for the first and most important feature, *1432_area* we see red points only having positive SHAP values (on right side of plot). This means that when the value of instances of the *1432_area* feature are high (points are red), i.e., there are a lot of very large warehouses in an area, the points have positive SHAP values (on right side of plot), i.e., lots of very large warehouses in an area impacts idling positively, i.e., is an indicator of high idling. Conversely, we see mostly blue points with positive SHAP values for *1111_area*, which indicates single-family detached housing, meaning that low (blue) values of *1111_area* (low amounts of single-family detached homes in an area) contributes to higher idling. Both these findings are expected, as the presence of large warehouses would be expected to have a positive correlation with idling, and the absence of single-family detached houses indicates the possibility of an industrial area (as it wouldn't likely be residential) which is expected to have greater idling activity.

6. Future Work:

6.1. Upsampling

Next steps in the research would include examining upsampling as a potential for model improvement, specifically as it relates to outlier capture (prediction). By artificially creating more target data values on the high end of idling distribution, we ensure that the model "sees" more of these extreme data points and thus weights its predictions higher because it acknowledges that there are substantial cases of high idling duration—not just a few exceptions, but a significant contribution to the trend. Despite inputting these data points ourselves, by forcing the model to come across more of them, we are conveying our greater emphasis on predicting these idling hotspots in a statistical way through the upsampling process. The efficacy of upsampling with a different model were already seen—during the initial RF modelling attempts (see *Section 4.4.*), upsampling led to a more accurate cell-specific max idling hours estimate (max idling—initial RF vs. upsampled RF vs. observed: 334 vs. 661 vs. 1,334 hrs). But, because these high values were intrinsically averaged across all cells by the nature of RF's architecture, this caused the total idling hours prediction to overshoot considerably (total idling—initial RF vs. upsampled RF vs. observed: 86,768 vs. 139,950 vs. 85,805 hrs). However, implementing upsampling would require careful imputation of data points throughout the upper range of idling, based on the actual distribution observed, to avoid losing an understanding of predictive trends in the other ranges of the distribution. But, since the regression results performed better according to Spearman correlation than r^2 (see *Table 5.2.*), the model ranks idling duration better than it is able to predict the exact hours, meaning that even with an artificially-increased idling outlook on the upper end, the model should be able to still identify where idling is likely going to be higher than lower, i.e., the monotonic relationship between

features and idling magnitude. So, if the predictions of idling magnitude end up being higher as a result of upsampling (which is the expected result), the total ranking of all predictions should remain similar, just with the average prediction being pulled higher, towards the right end, due to the imputation of more upper-end idling values before training, allowing for better outlier capture. Thus, the existing competence of the model in capturing the monotonic relationship, as evident by the higher Spearman correlation indicates that there's a type of safety net: the model already has a good understanding of where the hotspots are (since the magnitudes are ranked well), just not the full extent of the hotspot magnitudes, because the average idling magnitude is so comparatively low. Tools like SMOTE (Synthetic Minority Over-sampling Technique)—which is a key method for addressing imbalanced data by generating synthetic examples of the minority class, in our case the extreme idling outliers—can provide a means of achieving this (Imani).

6.2. K-Means

Another important technique to experiment with would be additional feature engineering and model context improvement through K-means clustering (Ren). Rather than relying solely on the arbitrary county lines to inform the model's understanding of the differences in idling opportunity across the study area, we could generate our own "counties" as added features. First, we would have to identify an optimal number of groups to maintain the greatest variance while being computationally efficient, and cluster the grid cells into this number of different categories according to idling duration intensity and feature data context that promotes idling (key predictors). In adding these group indices as features to the grid cells, we would create further contextual understanding of the region as a supplement for data sparsity and provide it to the

model as a direct response to significant county differences in idling duration and idling infrastructure/contributors (see *Figure 3.6.*). By providing much more homogenous groupings of grid cells compared to county lines, these idling groups would ideally become a more important feature when assessing final regression prediction than county, which would validate their inclusion as beneficial to the ultimate goal of more accurate idling magnitude prediction.

7. Conclusion:

In conclusion, while there was previously thought to be a clear winner in modeling approach, the later analysis shown in *Table 5.5.* reveals that there remain many questions about the optimal way to set up the modeling architecture for this problem, most notably: single vs. per-county model (*Section 4.8.*) and hard- vs. soft- vs. threshold-based gating (*Section 4.3.*). Ultimately, nothing is ever perfect, so it is most likely a case of drawbacks and advantages. By using per-county models over a single model, you trade (give up) greater data abundance in favor of a specialized understanding of region-specific targets/features relationship; by using soft-gating over hard-gating, you trade (give up) boosted regression predictions able to reach farther toward outliers in favor of greater data abundance and "safer" overall regression predictions. This work demonstrates the great importance of initial planning, theoretical understanding, and practical understanding of the data and the needs of the project—through deep exploratory data analysis, research on methods/tools/technologies, a curiosity to discover, and an open-minded, willingness to experiment.

Our results show that it is possible to construct machine learning models that take in highly zero-inflated and extremely right-skewed target data (see *Section 3.3.*), combine it with even more highly zero-inflated and even more extremely right-skewed feature data (see *Section 3.5.*)—the traits of which go entirely against the original assumptions of these mathematical tools—and still generate insights into a problem. In our case, our high classification metrics (*Table 5.1.*) and Spearman correlations (*Table 5.4.*) show that our models can identify well where certain spatial occurrences will take place and where they won't, as well as capture a strong, monotonic relationship between these difficult targets and difficult features. For the broader implications of this research, this means that our approaches can be used to mitigate the

difficulties inherent in machine learning with such difficult data, and can provide specific insights into how a better understanding of long-haul combination truck idling location and magnitude prediction might be reached.

https://github.com/damiandobrowolski13/idlingML

References:

Context

Beidler, James. *Geospatial Gridding Surrogate Updates for the 2020 Emissions Modeling Platform*, Environmental Protection Agency, www.epa.gov/system/files/documents/2023-11/1120am_beidler.pdf.

Eyth, Alison. *Technical Support Document (TSD): Preparation of Emissions Inventories for the 2021 North American Emissions Modeling Platform*, U.S. Environmental Protection Agency Office of Air Quality Planning and Standards, www.epa.gov/system/files/documents/2023-12/2020_emismod_tsd_dec2023_4.pdf.

Illinois Warehouse Boom: Tracing the Growth of Mega-Warehouses and Their Health Impacts, Environmental Defense Fund, news.wttw.com/sites/default/files/article/file-attachments/IL_Warehouse_Boom_Report_EDF_4-24-24.pdf.

Improving Vehicle Fleet, Activity, and Emissions Data for on-Road Mobile Sources Emissions Inventories, Federal Highway Administration, www.fhwa.dot.gov/Environment/air_quality/conformity/research/improving_data/taqs04.cfm.

Martin, Randal S., et al. *Assessment of Automobile Start and Idling Emissions Under Utah Specific Conditions: Cold Start, Hot Start, and Idle Emissions as Measured on Northern Utah Vehicles*, Utah Division of Air Quality, apps.weber.edu/wsuimages/ncast/projects/Cold_Hot_Start_Idle_Emissions_Final_Report.pdf.

National Land Cover Database (NLCD) 2021 Land Cover Conterminous United States, U.S. Geological Survey, mnatlas.org/metadata/nlcd_2021_landcover_UTM.html#:~:text=Developed%2C%20Low%20Intensity%20%2DIncludes%20areas,23.

Overview of EPA's Motor Vehicle Emission Simulator (MOVES3), Environmental Protection Agency, www.epa.gov/sites/default/files/2021-03/documents/420r21004.pdf.

"SMOKE (Sparse Matrix Operator Kernel Emissions) Processing System." *Community Modeling and Analysis System (CMAS)*, Institute for the Environment - UNC-Chapel Hill, www.cmascenter.org/smoke/.

Hurdle Model

Francisco, Micanaldo Ernesto et al. "Hybrid Machine Learning Approach to Zero-Inflated Data Improves Accuracy of Dengue Prediction." PLOS Neglected Tropical Diseases 18 (2024): n. pag. https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0012599.

Gating Approaches

Bengio, Yoshua. *Curriculum Learning in Machine Learning*, University of Montreal, www.scribd.com/document/425393423/Bengio-2009-Curriculum-Learning-pdf.

Sieminski, Antoni M. et al. "Forecasting overhead distribution line failures using weather data and gradient-boosted location, scale, and shape models." (2022). https://www.semanticscholar.org/paper/Forecasting-overhead-distribution-line-failures-and-Sieminski-Andrews/364aedee6e7ac1ba77fc7c1f92716349ebdca7f7.

XGBoost

Chen, Tianqi, and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining." *ACM Digital Library*, Association for Computer Machinery, dl.acm.org/doi/10.1145/2939672.2939785.

Tweedie

Anyidoho, Prosper K. et al. "A machine learning approach for predicting hurricane evacuee destination location using smartphone location data." Computational Urban Science 3 (2023): n. pag. https://link.springer.com/article/10.1007/s43762-023-00102-0.

Interaction Features

Thakur, Deepak et al. "An enhanced diabetes prediction amidst COVID-19 using ensemble models." Frontiers in Public Health 11 (2023): n. pag. https://www.semanticscholar.org/paper/An-enhanced-diabetes-prediction-amidst-COVID-19-Thakur-Gera/7bf90250e428c7ceb42de9e275e59d7972cb98ee.

K-Nearest Neighbors

Wang, Haoyu et al. "Nearest-Neighbor Neural Networks for Geostatistics." 2019 International Conference on Data Mining Workshops (ICDMW) (2019): 196-205. https://www.semanticscholar.org/paper/Nearest-Neighbor-Neural-Networks-for-Geostatistics-Wang-Guan/1113972f05e299e680547fbfe370c0217753f6b8.

Feature Importance

Koç, Neriman Sıla et al. "Machine Learning-Based Prediction and Feature Attribution Analysis of Contrast-Associated Acute Kidney Injury in Patients with Acute Myocardial Infarction." Medicina 62 (2026): n. pag. https://www.semanticscholar.org/paper/Machine-Learning-Based-Prediction-and-Feature-of-in-Ko%C3%A7-Ulusoy/fda98577a48299f0864ecdf93265171bbcaa04c1.

Upsampling (SMOTE)

Imani, Mehdi et al. “Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels.” Technologies (2025): n. pag. https://www.semanticscholar.org/paper/Comprehensive-Analysis-of-Random-Forest-and-XGBoost-Imani-Beikmohammadi/0799bfdeebabc6b8322df5b2d83a2760617c69f3.

K-Means

Ren, Jian-Yu et al. “Prediction of Distribution Network Line Loss Rate Based on Ensemble Learning”. International Journal of Engineering and Technology Innovation, vol. 14, no. 1, Jan. 2024, pp. 103-14, https://doi.org/10.46604/ijeti.2023.12869.