A Machine Learning Approach to Methane Emissions Mitigation in the Oil and Gas Industry

Jiayang Wang
Department of Data Science
Harrisburg University
Harrisburg, PA 17101
jiawang@my.harrisburgu.edu

Selvaprabu Nadarajah
Department of Information and Decision Sciences
University of Illinois at Chicago
Chicago, IL60607
selvan@uic.edu

Jingfan Wang
Stanford University
Stanford, CA94305
jingfan@stanford.edu

Arvind P. Ravikumar
Department of Systems Engineering
Harrisburg University
Harrisburg, PA 17101
aravikumar@harrisburgu.edu

Abstract

Reducing methane emissions from the oil and gas sector is a key component of climate policy in the United States. Methane leaks across the supply chain are stochastic and intermittent, with a small number of sites ('super-emitters') responsible for a majority of emissions. Thus, cost-effective emissions reduction critically relies on effectively identifying the super-emitters from thousands of well-sites and millions of miles of pipelines. Conventional approaches such as walking surveys using optical gas imaging technology are slow and time-consuming. In addition, several variables contribute to the formation of leaks such as infrastructure age, production, weather conditions, and maintenance practices. Here, we develop a machine learning algorithm to predict high-emitting sites that can be prioritized for follow-up repair. Such prioritization can significantly reduce the cost of surveys and increase emissions reductions compared to conventional approaches. Our results show that the algorithm using logistic regression performs the best out of several algorithms. The model achieved a 70% accuracy rate with a 57% recall and a 66% balanced accuracy rate. Compared to the conventional approach, the machine learning model reduced the time to achieve a 50% emissions mitigation target by 42%. Correspondingly, the mitigation cost reduced from $85/t CO2e to $49/t CO2e.

1 Introduction

As the main component of natural gas, methane (CH_4) is a potent greenhouse gas whose global warming potential (GWP) is 25 times that of carbon dioxide (CO_2) over 100 years [1]. Therefore, mitigating methane leaks from the oil and gas sector is a critical component of climate action in the United States and Canada.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

There are three major challenges to effective methane emissions reductions. First, the stochastic and time-variant nature of methane emissions across geographically sparse sites makes leak detection challenging. Second, the occurrence of large leaks or ‘super-emitters’ is highly complex and not easily predicted. It depends on several variables such as infrastructure age, production volumes, geologic characteristics, maintenance procedures, and local weather conditions. Third, conventional approaches to finding these super-emitters take a brute-force approach (e.g., survey all sites) that are time consuming and expensive. Thus, an effective solution that can identify ‘super-emitter’ sites would reduce methane emissions, improve cost-effectiveness of methane regulations, and potentially help operators develop predictive capabilities and improve maintenance procedures.

Regulations typically require operators to conduct leak detection surveys at all sites, followed by repairs – these are collectively called leak detection and repair or LDAR programs. Most regulatory agencies currently require the use of optical gas imaging (OGI) technology to conduct leak detection surveys. OGI-based LDAR surveys are walking surveys typically performed by a crew of 1 – 2 people at a survey rate of 3 – 5 sites per day. Leaks detected during this survey are flagged for subsequent repairs. However, conducting these surveys is time-consuming and costly.

Modeling efforts have focused on understanding the relationship between emissions and other site characters. Several studies have investigated the association between emissions and attributes like production and site age [2, 3, 4, 5, 6]. The common approach is to use multiple linear regressions to analyze the percentage of variance in emissions explained by production volume. A recent study [7] used a logistic regression model to predict occurrence of site-level emissions, which found well age, oil production, and energy content from oil to have stronger predictive power compared to other parameters. However, the stochasticity in the occurrence of leaks and the influence of several known and unknown variables make deterministic approaches such as linear regressions over a small number of variables generally ineffective.

In this work, we explore a machine learning approach to estimate the probability of a site being a ‘super-emitter’. Compared to previous approaches, machine learning models are better suited to the task because they do not require explicit delineation of the relationship between exogenous variables and methane emissions. Using data collected from recent field studies, we will train the model to estimate the probability of super-emitting sites without the need to explicitly define model parameters that affect emissions. Moreover, machine learning techniques can address the imbalanced dataset problem typically seen in oil and gas methane emissions distributions as a significant amount of emissions come from only a small number of sites [8]. The outcome of the model can guide operators to prioritize these high-emitting sites for repair.

2 Methods

The prediction problem is posed as a probabilistic classification problem – the outcome indicates a site’s likelihood to be high-emitting. We combined field measurement data on emissions and public data on site demographics to produce a modeling dataset. Emissions data were collected in a field campaign that took place at oil and gas production sites in Canada from August 2018 to September 2019. The field crew used an optical gas imaging camera to detect leaks, paired with a handheld instrument to quantify emission rates. Sites were randomly selected and representative of the oil and gas production distribution in the region. Over the course of 12 months, emissions data were collected from 436 sites that included 207 gas production sites and 229 oil production sites. A large sample size was necessary to sample a representative number of high-emitting sites in the full dataset. Publicly available site demographic data was collected from Petrinex.

The super-emitter pattern was observed in our data, with the top 20% of high-emitting sites contributing to 77% of total emissions. These high-emitting sites are not necessarily high-producing sites. Among the top 20% of high-producing sites (n=87), only 30 sites are also in the top 20% high-emitting group. Therefore, simple regression analysis over production volumes as undertaken in prior studies would not have high predictive power.

The last step in data collection is to find an emission size to define high-emitting sites (positive class). Common practice in emission studies defines high-emitting sites as a certain percentage of total sites, sorted in descending order, typically the top 5% – 20%. However, such a definition is constrained by individual field campaigns and thus, creates a large range of emission cutoff sizes from various studies. Here we took a different approach and used marginal return of emission coverage to find the cutoff emission size. We compared the marginal return (the increase in the cumulative percentage of emissions coverage) of surveying one additional site. As the law of diminishing return indicates, the marginal return will keep increasing until it reaches the optimal point, after which it starts to decrease. In our dataset, the optimal point is 212 CH_4 kg/day – within the top 5% – 20% cutoff range from a recent study in the same region [2]. Thus, we use 200 CH_4 kg/day as the cutoff value. Sites with emissions higher than 200 CH_4 kg/day (25% of total sites, 86% of total emissions) should be prioritized for mitigation efforts.

3 Predictive models and performance

Training and testing sets were created using a 75% and 25% split. The training dataset was used to develop predictive models and the testing dataset was used to evaluate model performance. Both the training and testing datasets were imbalanced because only 25% of sites are high-emitting. For the training dataset, we addressed this issue by bootstrapping from the high-emitting sites until the dataset is balanced. Because the testing dataset was still imbalanced, we also considered balanced accuracy rate – the average accuracy per class – as an evaluation metric, in addition to the accuracy and recall/sensitivity rate. We selected models that require minimal feature engineering for our dataset (Logistic Regression, Decision Trees, Random Forest, and AdaBoost) and trained them on the bootstrapped training dataset.

Table 1: Comparison of model performances.

Model	Accuracy	Recall/Sensitivity	Balanced Accuracy
Logistic Regression	70%	57%	66%
Decision Trees	72%	46%	64%
Random Forests	73%	20%	56%
AdaBoost	72%	32%	59%

The best performing model is logistic regression (Table 1) with the highest recall/sensitivity and balanced accuracy rate of 57% and 66%, respectively. Even though decision trees, random forests, and AdaBoost all have higher accuracy rate, they performed poorly on correctly predicting the positive class – low recall/sensitivity rate.

4 Results

To assess the effectiveness of the machine learning model over conventional approaches, we compared emissions mitigation and cost-effectiveness of three scenarios.

Scenario 1 (Baseline) : The baseline scenario involves survey of all production sites in the dataset in a random order. This scenario is currently used in LDAR methane regulations in the United States and Canada. Monte-Carlo simulations are used to derive confidence intervals around the efficacy of the baseline survey method.

Scenario 2 : We derived probability outcomes for high-emitting sites from our model and ranked them from highest to lowest. The order from the machine learning model is then used as the survey order for operators.

Scenario 3 : The third model for comparison comes from the gas production method where sites are ranked by gas production and operators start with the highest producing sites. Large production sites often have a significant number of equipment on site that are prone to leaking.

While assessing model impact on time and financial investment, we assume an average OGI-based LDAR survey speed of 5 sites/day at a cost of $3000/day, observed in typical field conditions and follows EPA guidelines [9, 10]. Figure 1 shows the number of super-emitter sites surveyed across the three survey scenarios. In Figure 1(a), by following the order predicted by logistic regression model, 18 super-emitter sites were surveyed by the end of week 1, covering 51% of total super-emitter sites. Compared to the gas production and the baseline scenarios, the order predicted by machine learning surveyed up to twice the amount of super-emitter sites in the first week. Figure 1(b) presents the results from random forest model, which surveyed 17 super-emitter sites in week 1.

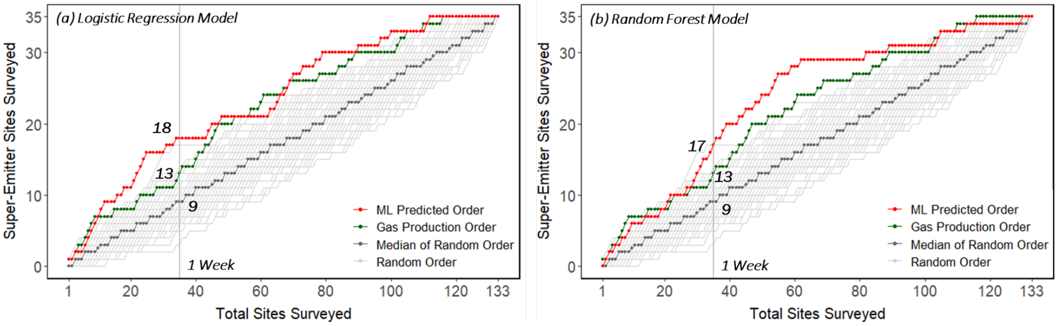

Figure 1: Number of super-emitter sites surveyed across the three survey scenarios. The x-axis shows the number of total sites surveyed and the y-axis shows the cumulative number of super-emitter sites surveyed.

Figure 2 shows the time evolution of cumulative fraction of emissions detected across the three survey scenarios. Figure 2(a) presents results from the logistic regression model. Using the survey order predicted by this model, operators detected 50% of total emissions by day 7. In comparison, the gas production and baseline survey scenarios detected 50% of total emissions by day 10 and day 12, respectively. By leveraging machine learning, we were able to reduce the time needed to detect 50% of total emissions by up to 42%. The average cost per site in reaching the 50% mitigation target is \$158 in the machine learning model, which is only 26% of the estimated \$600 estimated by EPA [10]. The mitigation cost decreased from \$85/t CO2e in the baseline approach to \$49/ton from the machine learning model.

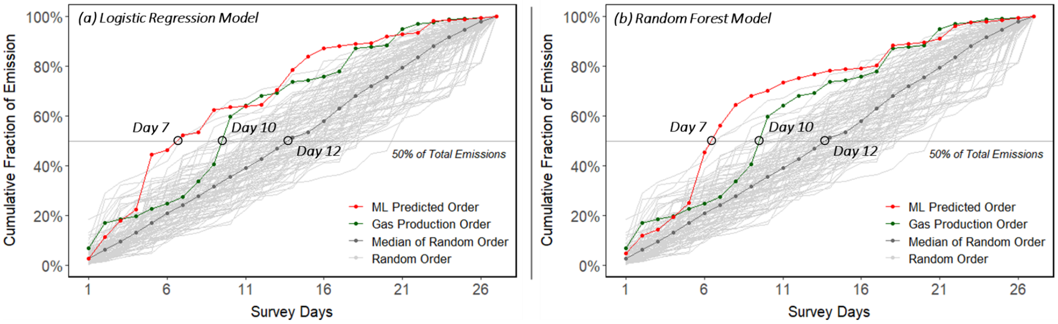

Figure 2: Time evolution of cumulative fraction of emissions detected across the three survey scenarios. The x-axis shows number of days and the y-axis shows the cumulative fraction of emissions covered.

Model performance does not necessarily translate into improvements in reaching the mitigation target. Figure 2(b) shows the survey results from the random forest model. Even though the random forest model performed worse compared to the logistic regression model, it still managed to reach the mitigation goal by day 7. The disconnect between model performance and speed in reaching mitigation target is likely from the high variance within high-emitting sites. Within the high-emitting sites, the highest emitting site emits 30 times more than the lowest high-emitting site. Because the machine learning models we used evaluated the probability of a site emitting more than 200 kg/day, they do not differentiate how much a site has exceeded this threshold. Such disconnect may be improved by dividing the sites into more emissions size groups instead of binary groups.

5 Model limitations and future work

This research is intended to be a proof-of-concept to illustrate the potential of machine learning algorithms to cost-effectively address methane emission mitigation in the oil and gas sector. By following the model predicted order, we managed to reduce the cost per site to reach a 50% mitigation target by 76% from $600 to $158, and decrease the mitigation cost of CO2e by 42% from $82/t CO2e to $49/t CO2e. Our preliminary findings show significant promise and motivate a more extensive study of machine learning methods for emissions mitigation. An improved version of this model will include more site level emissions data from basins across North America. Larger datasets that are representative of production characteristics across basins will improve both the robustness and the generalizability of our model. The site demographics variables used in this model are readily publicly available and require minimal feature engineering. We intend to incorporate more attributes including site equipment counts, geologic features, time since the last LDAR survey, and characteristics of nearby sites into the model. While the current model is designed as a classification model that uses probability as the proxy for ranking, we plan to explore the use of ranking models to directly predict the probability of a certain site ranking higher than another one. A robust predictive model will efficiently identify high-emitting sites, thus not only reduce methane emissions but also optimize the return on investments in methane mitigation.

References

[1] Environment and Climate Change Canada. Technical backgrounder: Federal methane regulations for the upstream oil and gas sector.

[2] D. Zavala-Araiza, S.C. Herndon, J.R. Roscioli, Yacovitch T.I., M R. Johnson, D.R. Tyner, M. Omara, and B. Knighton. Methane emissions from oil and gas production sites in alberta, canada. *Elem Sci Anth*, 6(1):27, Mar. 2018.

[3] M. Omara, M.R. Sullivan, X. Li, R. Subramanian, A.L. Robinson, and A.A. Presto. Methane emissions from conventional and unconventional natural gas production sites in the marcellus shale basin. *Environmental Science and Technology*, 50(4):2099 – 2107, Feb. 2016.

[4] M. Omara, N. Zimmerman, M.R. Sullivan, A. Li, X. adn Ellis, R. Cesa, R. Subramanian, A.A. Presto, and A.L. Robinson. Methane emissions from conventional and unconventional natural gas production sites in the marcellus shale basin. *Environmental Science and Technology*, 50(4):2099 – 2107, Feb. 2016.

[5] A.L. Mitchell, D.S. Tkacik, J.R. Roscioli, S.C. Herndon, T.I. Yacovitch, D.M. Martinez, T.L. Vaughn, M.R. Sullivan, C. Floerchinger, M. Omara, R. Subramanian, Daniel Zimmerle, A.J. Marchese, and A.L. Robinson. Measurements of methane emissions from natural gas gathering facilities and processing plants: measurement results. *Environmental Science and Technology*, 49(5):3219 – 3227, Mar. 2015.

[6] E.D. Brantley, H.L.and Thoma, W.C. Squier, B.B. Guven, and D. Lyon. Assessment of methane emissions from oil and gas production pads using mobile measurements. *Environmental Science and Technology*, 48(24):14508 – 14515, Dec. 2014.

[7] D.R. Lyon, R.A. Alvarez, D. Zavala-Araiza, A.R. Brandt, R.B. Jackson, and S.P. Hamburg. Aerial surveys of elevated hydrocarbon emissions from oil and gas production sites. *Environmental Science and Technology*, 50(9):4877 – 4886, May 2016.

[8] A.R. Brandt, G.A. Heath, and D. Cooley. Methane leaks from natural gas systems follow extreme distributions. *Environmental Science and Technology*, 50(22):12512–12520, Nov. 2016.

[9] A.R. Ravikumar, D. Roda-Stuart, R. Liu, A. Bradley, J. Bergerson, Y. Nie, S. Zhang, X. Bi, and A.R. Brandt. Repeated leak detection and repair surveys reduce methane emissions over scale of years. *Environmental Research Letters*, Jan. 2020.

[10] Environmental Protection Agency. Oil and natural gas sector: Emission standards for new, reconstructed, and modified sources reconsideration, Sep. 2020.