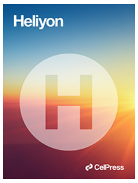
Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Machine learning techniques to increase the performance of indirect methane quantification from a single, stationary sensor

Robert S. Heltzel [*], Derek R. Johnson, Mohammed T. Zaki, Aron K. Gebreslase, Omar I. Abdul-Aziz

West Virginia University, Mechanical and Aerospace Engineering Department, Center for Alternative Fuels, Engines, and Emissions, 263 Engineering Sciences Building, Morgantown, WV 26506, United States

ARTICLE INFO

Keywords:
OTM 33A
Eddy covariance
Indirect quantification
Controlled releases
Methane emissions
Machine learning
Random forest
Neural network

ABSTRACT

Researchers are searching for ways to better quantify methane emissions from natural gas infrastructure. Current indirect quantification techniques (IQTs) allow for more frequent or continuous measurements with fewer personnel resources than direct methods but lack accuracy and repeatability. Two IQTs are Other Test Method (OTM) 33A and Eddy Covariance (EC). We examined a novel approach to improve the accuracy of single sensor IQT whereby the results from both OTM and EC were combined with two machine learning (ML) models, a random forest (RF) and a neural network (NN). Then, models were enhanced with feature reduction and hyperparameter tuning and compared to traditional quantification methods. The NN and RF improved upon the default OTM by an average of 44% and 78%, respectively. When compared to traditional OTM estimates with low Data Quality Indicators (DQIs), RF and NN models reduced 1σ errors from ±66% to ±13% and ±34%, respectively. Models also reduced the standard deviation of estimates with 93% and 85% of estimates falling within ±50% of the known release rate. This approach can be deployed with single sensor systems at well sites to improve confidence in reported emissions, reducing the number of anomalous overestimates that would trigger unnecessary site evaluations. Additional improvements could be realized by expanding training datasets with more methane release rates. Further, deployment of such models in a variety of situations could enhance their ability help close the gap between bottom-up inventory and top-down studies by enabling continuous monitoring of temporal emissions that could identify with improved confidence, atypically higher emissions. Accurate remote single sensor systems are key in developing an improved understanding of methane emissions to enable industry to identify and reduce methane emissions.

1. Introduction

Over the past two decades the United States has experienced a natural gas (NG) boom due to advances in exploration and production operations. This growth has led to a rapid increase in the number of NG producing wells, production sites, and the complexity and equipment on any given site. The increase in NG production has resulted in a concern about leaked methane emissions, as it is a potent greenhouse gas (GHG). Estimating the methane emissions from NG production has been difficult to achieve with a reliable level of accuracy. Many researchers and government agencies have produced estimates that differ dramatically [1, 2, 3]. For example, Zavala-Araiza et al. estimated emissions rates that were 1.9, 3.5, and 5.5 times higher than the EPAs Greenhouse Gas Inventory, EPAs Greenhouse Gas Reporting Program, and the Emissions Database for Global Atmospheric Research (EDGAR), respectively. To understand the environmental impact of the NG production industry, uncertainty needs to be reduced and confidence in indirect quantification techniques (IQTs) improved. Traditionally, methane emissions have been quantified through direct measurements. However, as the number of production sites increases, this time-intensive method may no longer be economically viable for industry or researchers. Also, methane emissions from the NG infrastructure are now known to be highly temporal [4, 5, 6]. Direct measurements campaigns represent “snapshots” in time and may further obfuscate total average emissions due to temporal variability. These time variant emissions can only be understood with more frequent or continuous measurements. In response, a few novel IQTs have emerged. IQTs allow for increased measurement frequency, though often at the

[*] Corresponding author.
E-mail address: rheltzel@mix.wvu.edu (R.S. Heltzel).

https://doi.org/10.1016/j.heliyon.2022.e11962
Received 1 February 2022; Received in revised form 12 April 2022; Accepted 21 November 2022

Synopsis

Machine learning methods were applied to indirect methane quantification techniques to improve accuracy and reduce uncertainty.

expense of accuracy. These methods involve different scales (local and regional) and techniques (point measurement, mobile vehicle, aerial flux).

Two currently utilized stationary single-sensor IQTs, Other Test Method (OTM) 33A and Eddy Covariance (EC), are both capable of measuring methane fluxes from local sources. These techniques differ in their principle of measurement but use similar data acquisition (DAQ) equipment. OTM attempts to measure a horizontal flux from a point source and was developed based on point source gaussian dispersion principles [7]. Several researchers have utilized the OTM method to quantify methane emissions from known controlled releases [8, 9, 10]. Comparing estimates to known release rates has often resulted in large uncertainties and highly variable estimates. Robertson et al. estimated that 2σ errors of OTM were ±56%, even when eliminating data of poor quality [10]. Edie et al. used OTM to quantify controlled releases from a simulated NG production site and estimate errors ranged from −60% to +170% [9].

EC is a method for measuring vertical fluxes. The general theory of EC has been outlined in detail by Burba et al. [11]. Traditionally, EC has been used to measure vertical fluxes from homogeneous area sources, however, some new research has utilized it to target more heterogenous sources such as livestock [12, 13]. EC estimates are often combined with "flux footprint" models to attribute emissions to sources. These models have traditionally been based on advection-diffusion or Lagrangian dispersion models [14, 15, 16]. Models of this kind are complex and utilize many simplifying assumptions – which results in large estimate and source attribution uncertainties. Researchers have stated that "additional tracer release studies are necessary to improve the confidence of EC measurements and validate footprint model estimates" [13].

More reliable methods must be developed before IQTs can be relied upon by researchers, industries, and governmental policymakers to replace accurate direct measurements. The hypothesis of this research was that a combination of two current methods could help to enhance the estimates of methane mass emissions from the NG infrastructure. We believed that combining the outputs of currently used methods (EC and OTM) with machine learning (ML) algorithms would enhance the accuracy of single sensor techniques. Such a method could eliminate the need for complex models and a multitude of assumptions. ML methods can interpret diverse data and recognize complex patterns. Such attributes are inherent in problems involving the stochastic nature of micrometeorological measurements.

Two machine learning methods examined included Random Forests (RF) and Neural Networks (NN). RF were developed by Breiman in 2001 and have been applied in various scientific fields including ecology, medicine, astronomy, traffic and transportation planning, and agriculture [17]. RF application is beneficial to large-scale data, provides resistance to overfitting, and has recently been used to study the connection between various factors and carbon emissions [18]. RF has also been used to investigate carbon flux emissions from soils and forests [19, 20]. Mascaro et al. showed that RF aided in carbon mapping applications using remote data and models reduced RMSE by over 20% [19]. Philibert et al. also applied RF to predict greenhouse gases (N_2O) and showed RF performed better than standard regressions models [21].

NN have a longer history and have been developed over decades and have been applied broadly across the sciences [22]. NN have been used to predict GHG from transportation systems [23], predict methane emissions from biologic sources [24], identify downwind methane spikes associated with stages of well site development [25], and to locate and quantify fugitive natural gas leaks [26]. Travis et al. utilized NN along with new leak detection sensors to detect and estimate methane emissions from mock natural gas sites, but we note their method overestimated methane emissions by a factor of 1.77–1.83 depending on pad size [26]. In addition, Wang et al. have incorporated NN with machine vision to help detect methane emissions from infrared cameras [27]. Based on these recent research results, both RF and NN will be examined as machine learning techniques to increase the performance of indirect methane quantification from a single, stationary sensor.

2. Methodology

The hypothesis was tested through a series of controlled methane releases, allowing for comparison of traditional methods to the novel ML models. The DAQ system utilized was a small, mobile EC tower (MECT). The tower was outfitted with the instrumentation required for collecting both OTM and EC data. These data were collected at a rate of 10 Hz. The primary instruments were a LICOR LI-7700 (Lincoln, NE, USA) for methane concentration measurement and a Gill® Windmaster 3-D sonic anemometer (Lymington, UK) for wind speed [28, 29]. A more detailed description of the DAQ system can be found in concurrent literature [30, 31].

The controlled releases of methane occurred from a bluff body in an open, grass field at the JW Ruby Research Farm located in Reedsville, WV. The MECT was stationed downwind of the releases based on the prevailing wind direction. Three release rates (0.04, 0.12, and 0.24 g/s) were measured at varying distances between 40 and 120 m. The release matrix is presented in Table S1. Data for controlled releases and background periods occurred between May 21st and September 11th, 2019.

Measurements from the tower were collected continuously and EC and OTM estimates were calculated based on the standard methods. OTM calculations were performed in Python with scripts based on those published by the EPA [32]. EC results were determined using EddyPro® software [33]. More extensive descriptions of the calculations of both OTM and EC are presented elsewhere [7, 11, 30, 31]. Outputs from OTM and EC were determined using 15-minute averaging periods.

Averaging periods were eliminated from consideration if either OTM did not produce a valid mass rate estimate or EC did not produce a valid flux estimate. They were also eliminated by a wind filter (WF) if the prevailing wind direction of the period was not within ±45° of the source-to-sensor (STS) direction. Data were not eliminated from contention for quality indicators of either method; however, such ratings were included in the ML datasets. A complete breakdown of the 15-minute continuous averaging periods is presented in Table S2. More detailed information on the data averaging period filtering is found elsewhere [31].

The outputs from OTM and EddyPro® calculations were used to form the primary datasets for combined evaluations. There were 804 valid controlled release periods and 1208 valid background periods. To balance the dataset, 804 randomly selected background periods were used in model evaluation resulting in a full dataset of 1608 periods. The variables used for analysis were from both default processing methods.

The OTM outputs were selected based on those relevant to the method and other available data averages for the period. The number of valid output variables from the OTM calculations was 27. The number of EddyPro® variables in the full output file based on the settings used was 182. From this list all non-number or identifying variables were removed resulting in 146 variables. The distance and StS direction were then added to the set of variables. The total number of possible variables in the dataset was 175 and are described elsewhere in detail [31]. These variables represented inputs to the ML models and were defined as "features." Using all features produced a high ratio (greater than 10%) of features to data instances (1608). It was expected that not all features would have a significant impact on model results. Therefore, methods for reducing the size of the feature set were explored.

Firstly, all features were removed that did not have a Pearson Correlation p-value less than 0.05 when compared to the controlled release rate as described by Eq. S(1). This eliminated all features that had no significant correlation with the controlled release rate. This release-rate-correlated (RRC) dataset contained only 54 features. Secondly, the number of features was reduced by eliminating those that were highly cross-correlated. High correlation reduction was performed by grouping all features which had cross-correlation coefficients greater than 0.75. The two major variable groups were those associated with temperature or atmospheric water vapor, and those associated with methane. There were 32 highly cross-correlated features that were eliminated by this filter. The complete list of cross-correlated features is available in Table S6. This resulted in a dataset that contained only RRC features that were not eliminated by the high-correlation filter (HCF). The total number of features in the RRC-HCF dataset was 22.

Several ML models were considered based on the problem space and recommendations from several sources [34, 35, 36]. The outline and considerations for model selection are detailed in the SI. Python's scikit-learn (sklearn) package was utilized for model building and training. The final evaluated models were the following.

- RF regressor using all scaled features.
- NN using RRC-HCF scaled features.

Since none of the best performing models utilized the same feature sets, other considerations were used to select the desired set. Datasets containing fewer features were desired for both interpretability and simplicity. As such, the RRC-HCF was the most desirable feature set. This was the best performing feature set of the NN. The next best performing dataset on the NN was the RRC which resulted in a 33% increase in the RMSE. Every feature set used in conjunction with the RF produced a lower RMSE then all but one other model. The difference between the best and worst performing RF evaluation was only 11%. The best performing NN resulted in an RMSE that was 87% higher than the worst performing RF. These results suggested that neither scaling nor variable set had a major impact on RF performance, which is further detailed in literature [31]. This was also an indication that the RF required only a few critical features to produce more accurate results than the other models. The NN had the largest difference between scaled and unscaled datasets. The NN was also the most affected by the size of the feature set. A reduction in the number of features by 69% and 87% improved the results of the NN by 74% and 80%, respectively. These results suggested the model was sensitive to the number of features used and performed better when irrelevant features were removed. These factors led to selection of the RRC-HCF as the default dataset moving forward. For the sake of interpretability, scaling was not implemented for the RF since it had minimal impact on results. However, without scaling the NN RMSE increased by 1–5 orders of magnitude depending on the feature set. Scaling the dataset was considered an essential preprocessing step when utilizing the NN. To further evaluate model performance, the following model/feature set combinations were utilized:

1. RF with RRC-HCF unscaled features
2. NN with RRC-HCF scaled features.

During initial ML evaluations, it was clear that the number of features impacted algorithm performance, particularly the NN. In general, fewer features led to lower model prediction RMSEs. To evaluate the impact of the number of features on model performance, both the RF and the NN algorithms were evaluated by using a dataset that grew by one feature at a time. Features were added in order of their Pearson correlation coefficients (PCC) with respect to release rate. The first feature was the one most strongly correlated with release rate, the algorithms were trained and tested, then the next strongest correlated feature was added and so on. This method resulted in 22 evaluations of both the NN and RF and allowed the number of features required for model optimization to be determined. The test set RMSEs as well as the percentage of the model minimum, as a function of the feature added are presented in Table S8.

One advantage of RFs is their ability to quantify the importance of different features. RFs can produce attributes which define the feature importance for each variable in the dataset [37]. The feature importance is a percentage of the impact that each feature has, on average, on the RF's prediction.

The NN attained its best results using just two features with an RMSE of 0.0568 g/s. Though it should be noted that the network used here had only one hidden layer of size 100, a different architecture could have significantly changed results. The RF achieved its lowest RMSE with the use of the first 18 features with an RMSE of 0.0434 g/s. These results further emphasize the importance of input features as optimal selection reduced the difference between the two models from 61% to 27%.

The minimal feature RF and NN models improved upon their default RMSEs by 5% and 33%, respectively, for the controlled release dataset. The NN model was severely limited by only utilizing two features, when tested on data other than the controlled releases. To compensate for this, the 18 features identified by the RF feature analysis were used in both models moving forward as the default feature set. Using the RF identified features on the NN decreased the test RMSE by 13% from the default value 0.085 g/s to the new value of 0.074 g/s. Descriptions of these features and their statistics are presented in Table S6.

A low-cost approach to improve ML algorithms is hyper-parameter optimization. To improve the performance of the RF and NN models, the controlled-release dataset was analyzed while tuning the model hyper-parameters. There are several methods for hyper-parameter tuning which can be utilized depending on the hyper-parameter search space size, computational power, and relevance. A random grid search was used to optimize the parameters of both the RF and the NN. Random searches have been proven to be more efficient at finding optimal hyper-parameters than a strict grid search, in which all iterations are tested [38].

A random search with cross-fold validation was utilized for hyper-parameter search optimization. The search was evaluated for 100 iterations and k cross-fold validation where k was set to three [39]. We note that choice of k is often five or 10, but for small data sets higher k values may bias results and so three was selected during our research as balance of computation effort and to avoid overfitting [40]. We note that Nguyen et al. recommended at least 3-folds [41] and Lan Vu et al. found that 7-fold cross validation reduced RMSE and increased R^2 values for NN models with skewed datasets [42]. The RF parameters evaluated and how each one affected the RF, along with the values used to form the random grid search matrix are presented in Table S7. The total number of possible hyper-parameter combinations was 3960, so utilizing the random search reduced the number of required computations by an order of magnitude. However, results of the randomized search did not improve the RF model performance and the default model was maintained.

For consistency, the same random search with cross-validation was utilized for hyper-parameter tuning of the NN. Scaled features were once again used from the controlled release dataset and the RF identified features made up the inputs to the MLP regressor. NNs have an infinite number of hyper-parameters because the number of layers and the number of neurons per layer can be set to any number. For the sake of minimizing the search space the number of layers tested and the size of those layers were limited. Eq. (1) was utilized to determine a search space for number of total neurons based on recommendations [43].

$$N_{hidden\ neurons} = \frac{N_{training\ samples}}{(\alpha^*(N_{inputs}+N_{outputs}))} \tag{1}$$

where α is a scaling factor between 2 and 10. Based on the controlled release dataset the number of training samples was 1206, the number of inputs was 18, and the number of outputs was one. The α values evaluated were 2, 4, 6, 8, and 10. The total number of layers over which to distribute these neurons was also required. Researchers have found that deeper NNs do not necessarily improve results. To keep the number of hidden layers to a minimum, up to three hidden layers were tested. The total number of hidden neurons for each of the different α levels were distributed evenly amongst the number of hidden layers. As a result, the number and size of hidden layers tested in the random search were those presented in Table S8.

In addition to these hidden layer shapes, the other hyper-parameters tested in the initial evaluation were the activation function, solver, and L2 penalty. The values utilized in the initial random grid search are presented in Table S9. The alpha values tested were five evenly spaced values between 1.0×10^{-5} and 1 on a log-axis.

The search was evaluated for 100 iterations and three cross-fold validations, which reduced the total number of required iterations by a factor of three. The randomized search parameters resulted in a 20% decrease in the test set RMSE. The optimal randomized search parameters are presented in Table S10. The resultant optimal solver was a limited-memory Broyden-Fletcher-Goldfard-Shanno (lbfgs) which is a type of quasi-Newton optimization algorithm for finding local minima or maxima of smaller datasets [44].

A second random search was performed to optimize secondary parameters related to the solver. These features included the maximum number of iterations ('max_iter', default = 200) and the maximum number of function calls ('max_fun', default = 15000). These parameters were tested with the optimal results from the first iteration random search. The 'max_iter' parameter was evaluated on 10 linearly spaced values between 10 and 1000, and the 'max_fun' parameter was evaluated on 10 linearly spaced values between 1000 and 10,000. The random grid for this case tested all 100 possible combinations on three cross-fold validations to determine the optimized hyper-parameters. The optimal values of 'max_iter' and 'max_fun' were 450 and 1000, respectively. However, changing these values only reduced the RMSE by less than 1%. The final NN, which was optimized for the controlled release dataset, was constructed with the settings in Table S10. This was a 10% reduction in RMSE over the minimum feature evaluation RMSE using only two features. The resultant model was also believed to be more robust and less prone to overfitting.

Both the RF and NN models utilized a randomized grid search with cross-fold validation to optimize the hyper-parameters used. The RF improvements were minimal with these searches while the NN model improved by more than 32% when compared to the default model with the RF optimized feature set on the controlled release test dataset. The final models were those that resulted in the minimized RMSE without compromising computational time and expense. The improvements in both models are presented in Table 1.

3. Results and discussion

3.1. Controlled release estimate results

The final RF and NN models were compared to the OTM results. It should be noted that some of the periods were likely used in the training sets of the models, although since the sets were randomized the fraction of the periods used for training is unknown.

Table 1. RF and NN model iterations.

Features/Model Change	Controlled release test RMSE (g/s)	
	RF	NN
RRC-HCF	0.0453	0.085
Minimum Features	0.0434	0.057
RF-Optimized Features	0.0434	0.075
Hyper-parameter Tuning 1	0.0455	0.052
Hyper-parameter Tuning 2	–	0.051
Final	0.0434	0.051
Feature Set	RF Optimized	RF Optimized
Number of Features (#)	18	18
Tuning Iterations (#)	0	2

Comparisons were made between the ML models and the OTM results. These periods were not filtered by any criteria other than the WF that was initially used. The complete release rate dataset of 804 periods and the ML test dataset of 402 periods were evaluated against the default OTM results. The RMSEs of the 402 periods in the test set were compared from the machine learning models in Table 1. The default OTM RMSE for this test set was 0.12 g/s.

The complete controlled release dataset was also analyzed from the perspective of both ML techniques and OTM calculations. The RMSE results were calculated for the entire controlled release dataset (804 periods). The NN and RF improved upon the default OTM by an average of 44% and 78%, respectively. The RMSEs by release rate and method are presented in Table 2. The RF RMSE was 44% less on average than the NN across the three releases. Figure 1 presents the different release rates (Figure 1A – 0.04 g/s, Figure 1B – 0.12 g/s, and Figure 1C – 0.24 g/s) for these shared periods (note difference in scales). Data shows that the ML method means and medians are similar while for the default OTM means tended to be higher than the medians. Further, we see that all methods tend to under predict with an increasing release rate. The ML methods would benefit from increasing the count of periods, which decreased with increasing release rates, see Table 2. When single point sensor approaches are deployed for continuous monitoring, underestimated mean emissions where real release rates are still predicted within the 75th or 95th percentiles are valuable. Industry is seeking to deploy sensor based systems across the supply chain to monitor and detect methane emissions. While under predictions would skew estimates lower, a conservative method such as this would reduce overestimations produced by other methods that would contribute to "false" positives for industry. Such over estimations may alert site operators more frequently to super emitter events that are falsely predicted. The statistics of these box and whisker plots are available in Table S11.

3.2. Results of periods with a DQI < 10

Historically OTM uncertainties have been quantified by using statistics of several measurements. Typically, OTM estimations are eliminated if they produce a DQI value greater than 10, which can significantly limit datasets, especially when data are collected continuously, and no effort is made to collect data during optimal conditions. To compare the uncertainties to previous studies, the entire controlled release dataset was analyzed by the default OTM, RF and NN.

For comparison to other OTM studies, the periods that produced a DQI value less than 10 based on the default OTM analysis (n = 181) were analyzed. The percent error distributions are presented in Figure 2 for the release rates of 0.04 g/s (Figure 2A, n = 43), 0.12 g/s (Figure 2B, n = 99), and 0.24 g/s (Figure 2C, n = 39). Comparisons to previous controlled release studies are presented in Table 3. The RF and NN both improved upon the default OTM estimates. It is noteworthy that they contained more periods within ±50% than previous studies, suggesting that these methods may help to reduce the spread of predictions when the models are properly tuned and trained on similar scenarios. A comparison to previous studies is presented in Table 3.

We note that our default OTM data had a larger range of errors than any previous studies, but sample size was 1.6–9.5 times larger than other

Table 2. Comparison of RMSE results of full control release dataset with various methods.

Release Rate (g/s)	Count	RMSE (g/s)		
		Default OTM	Random Forest	Neural Network
0.036	395	0.059	0.012	0.024
0.119	325	0.19	0.030	0.054
0.239	84	0.25	0.073	0.12

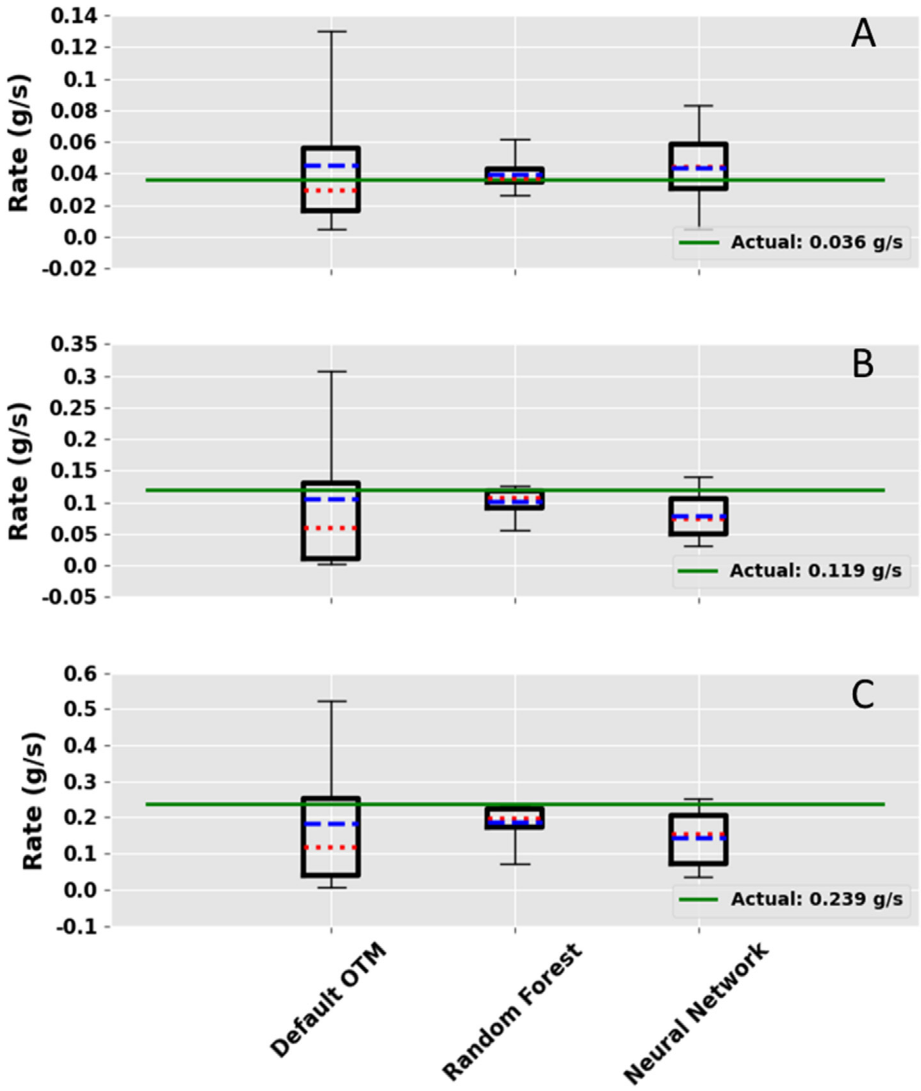


Figure 1. Comparison of OTM, RF, and NN against controlled releases for (A): 0.04 g/s (B): 0.12 g/s (C): 0.24 g/s. Box data are the predictions from the given method. The boxes encapsulate the lower and upper quartiles, the whiskers extend to the 5th and 95th percentiles, the blue lines represent the means, the magenta lines represent the medians, green lines represent the actual release rate. Unlike RF and NN models, the distributions of default OTM rates are heavily skewed regardless of the release rates (0.04–0.24 g/s).

studies. In addition, our work tended to focus on smaller emissions rates that may be more indicative of normal operating conditions at well sites [4, 5]. However, with the use of RF the combined OTM and EC data reduced the full range of error to that of Edie et al. While the RF and NN methods tended to underpredict for larger emissions rates, the range of estimates within the generally referenced range of ±50% improved over others giving more confidence within this range. In addition, we also recently showed that the minimum attainable uncertainty due to measurement uncertainty and stochastic nature of micrometeorological conditions was ±17.4% for OTM [45]. Here, 85% of the RF results were within the ±30% range.

Both ML methods improved over the default OTM for our analysis. Our analysis was conducted on relatively flat terrain with typical emissions rates for well pads which further challenges IQTs. However, the reduction in overpredicted values is important as industry moves towards deploying unmanned monitoring systems at well sites. Erroneously high estimates could trigger unwarranted site visits and reduce confidence in such systems. Our ML methods improved results, specifically the RF improves the range of estimates for ±30, ±50, and the 68th percentile. Additional training data would further improve the ML models as is discussed below.

3.3. ML model discussion

The RF and NN models were trained on a subset of the controlled release dataset and used to predict the mass emissions. Model inputs were selected based on correlations with known release rates, cross-correlations with other features, and model improvement and robustness. Model improvements were quantified by a reduction in the RMSE of a test dataset. The final feature set of the models consisted of 18 variables produced during calculations of OTM and EC methods. Model hyper-parameters were optimized using random searches with cross-validation. Optimized feature selection and hyper-parameters reduced the RMSE produced by the RF and NN by 4% and 40%, respectively. The ML models were then compared to traditional methods across several datasets.

Across the three different controlled release rates the NN and RF reduced the standard deviation of estimates by an average of 70% and 82%, respectively, compared to the default OTM estimates.

The RF, NN, and default OTM models were compared across a dataset of periods with DQI values less than 10 (n = 181). Previous studies involving OTM have typically discarded measurements with a DQI above 10. The 1σ errors of the default OTM from this dataset were ±64%, which were similar to the 1σ results of previous studies The NN and RF produced 1σ errors of ±34% and ±13%, respectively.

The RF and NN models were therefore able to produce estimates across the entire release dataset (n = 804). Default OTM values were compared to these estimates, although based on historical OTM research, many of these periods would have been discarded. Across the three release rates of the dataset, the NN and RF reduced the RMSE by an average of 61% and 78%, respectively. Both the NN and RF models also

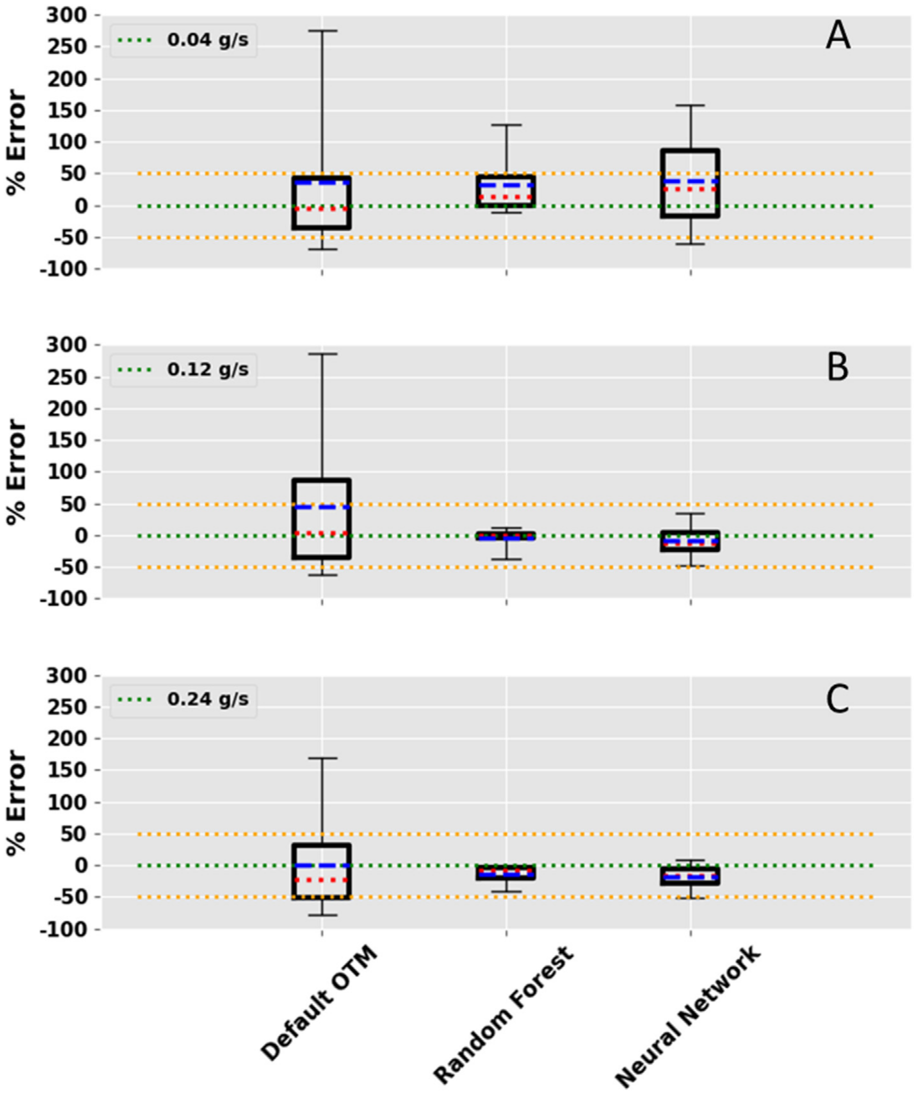


Figure 2. Comparison of OTM, RF, and NN estimate errors for (A): 0.04 g/s, (B): 0.12 g/s (C): 0.24 g/s for periods with a DQI < 10. Note the dotted lines highlight the ±50% error range about the 0% error reference.

Table 3. Comparison of results to previous OTM studies (DQI < 10).

	Previous Studies			This Work		
	Robertson et al.	Edie et al.	Brantley et al.	Default OTM	Random Forest	Neural Network
Count (#)	19	24	107	181		
Release Rates (g/s)	0.03–0.56	0.04–0.6	0.19–1.2	0.04–0.24		
Full Range of % Error	−75% to 60%	−60% to 175%	−60% to 52%	−95% to 1070%	−75% to 186%	−78% to 226%
Tests within ±30%	–	–	71%	30%	85%	65%
Tests within ±50%	85%		–	56%	93%	85%
68th Percentile Error	±28%	±38%	–	±66%	±13%	±34%

produced estimates closer to zero during periods where no controlled release was present, compared to OTM results.

3.4. Overall method discussion

Traditional methods for estimating mass emissions from stationary single sensors are poorly constrained and have high uncertainties. We demonstrated that methods for improving estimates was possible through a novel combined (EC and OTM) approach with ML algorithms (RF or NN). However, neither of the ML algorithms were fully optimized at the conclusion of this research. The limited dataset utilized here allowed for comparison to a small number of controlled releases. The efficacy of either of these techniques as a realistic solution for improving the accuracy of measurements depends on several factors.

The main drawback of the RF was its inability to extrapolate beyond its initial training scope. This is a critical flaw in the method when moving from a controlled experiment with a maximum rate of 0.24 g/s to the potentially higher emissions rates of actual production sites. However, this could be overcome with a wider range of controlled release rates, distances, and release geometries. The problem with such an approach is that it is a slippery slope. Where does one stop expanding a controlled release matrix that is meant to encapsulate any real-world scenario? Even with this downfall, the RF could still provide value and insight into emissions predictions. The RF could be combined with other methods such as a NN. In such a scenario, the RF would be relied upon when the predicted emissions were below a pre-defined threshold based on its training and a factor of safety. Say that the RF was trained on release rates up to 0.24 g/s, as was performed here. The algorithm could say that if the prediction from the RF was above 80% of this threshold, then an alternative algorithm which allows for extrapolation should be used. If a RF were not used for any predictions of mass emissions estimates, its ability to identify key features could still be utilized. This would be valuable when combining with any type of NN as they are notoriously difficult to interpret and have the stigma of being a “black box” methodology. The use of a RF to identify features which could be used as inputs to a robust NN has been used in medical research with the use of Forrest Deep Neural Networks (fDNN) [46].

A key drawback of using NNs is the requirement that the features be scaled or normalized. This makes recursive training difficult because the model cannot simply be “updated” with new data. Instead, if a MLP was used, as it was here, the entire potential dataset would have to be rescaled and the NN retrained. This could be overcome by initially testing what are believed to be the distributions of each variable in a controlled release dataset. The difficulty of such a task would depend on the features believed to be required for inputs. Suggested limits could be easily defined for some variables such as temperature, air heat capacity, signal strength, and StS direction. Such variables have natural bounds. However, it would be much more difficult to set the bounds of variables such as methane flux, OTM estimates, or methane variance. The limits of these variables would not only depend on the ranges of mass emissions rates, but also on-site geometries, tower placement, and surrounding ecological conditions. These factors combined with the stochasticity of micrometeorological measurements, would make recursive training without rescaling a near impossibility. In addition, if the standard scaler were used, as in this work, the distributions of such variables would also be required. This data would be unknown no matter the scenario. These factors point towards the data requiring rescaling when the set is expanded.

With such drawbacks or alternatives in mind, the ML models utilized in this work could overcome significant obstacles with more concurrent direct measurements. Most models can be improved with higher quality input data. As measurements were performed in the field, the size and scope of the available data would continue to grow. Models trained on datasets of the highest quality would become more robust in their ability to interpret complex scenarios. The training and testing sets used in this research were limited to a small series of controlled releases. Ideally training sets could be expanded in the following ways:

1. More release rates – expanding the controlled release rate limits would expand the bounds of both models.
2. Different site geometries – better representation of NG infrastructure sites with the potential for multiple leaks would be of use.
3. Varied atmospheric conditions – changing the time of year of the training data and ensuring that the same releases were performed under various conditions would eliminate reliance on unrealistic correlations, which could be a source of error in the models. However, we note that the controlled release experiments covered a broad range of stability classes as did controlled releases from previous studies.

One way to rapidly increase the training set would be to deploy the MECT during LDAR programs that were coupled with quantification efforts. Even if the MECT was only on a site for a few days during a LDAR and quantification audit, it would add valuable training data to models. As new regulations move towards multiple annual audits, more training data would become available. In such a scenario, during a direct quantification campaign the MECT would be positioned on site at the start of the audit. Observation of wind direction would help with positioning of the MECT. Once deployed, a range finder or GPS coordinates could be used to estimate distances and StS directions. While the data would be limited if only deployed for a single day, over the course of a series of campaigns the dataset would expand and ideally encapsulate a wide range of scenarios on which the model could be trained. The MECT could then be deployed at similar sites, long-term, to allow for better understanding of the temporal variability of emissions.

Better understanding of the temporal distribution of emissions could help target reductions. Improvements in long-term IQTs that are low cost will help researchers and industry understand the problem of “super-emitters” as well. It may turn out that sites are only “super-emitters” intermittently, which would drastically change their emissions profiles compared to simply assuming that emissions are constant. Elucidating a clearer picture of emissions profiles both temporally and geo-spatially will drastically improve our understanding of the NG methane problem. While this work has not solved the problem explicitly, the research has identified methods that could enhance OTM alone, by reducing variability and increasing average accuracy. By using data from multiple approaches and with the inclusion of ML methods, the work performed here could provide a pathway to a better understanding of emissions.

4. Conclusions

Recent research has led to the development of new, cost effective methane sensors and proposed regulations may enable their deployment at natural gas sites to monitor emissions as part of leak detection and repair programs. To improve accuracy and confidence in indirect quantification techniques, we examined a novel approach of combining OTM and EC methods along with ML including RF and NN. We collected extensive controlled leak experiments over a broad range of atmospheric conditions to test our approach, its impacts on accuracy (RMSE), and the spread of estimates (standard deviations and various accuracy windows). In summary, some key quantifiable findings were:

- RF and NN models were developed and optimized, reducing baseline RMSEs by 4% and 40%, respectively.
- The optimized models reduced the RMSE of all datasets compared to the default OTM measurements.
- The NN and RF reduced the standard deviation of estimates with a DQI less than 10 by an average of 70% and 82%, respectively, compared to default OTM estimates.
- Across the full release dataset, the NN and RF RMSEs were 61% and 78% lower than default OTM; however, OTM errors were initially higher than other studies due in part to larger more variable data set conditions.
- Future research should examine 5-fold and 7-fold cross validation to examine benefits on error reduction as compared to model overfitting.

With these benefits, our approach could be deployed with new single sensor monitoring systems to improve accuracy of continuous estimates, while reducing anomalously high estimates that could trigger unwarranted site visits. Further, as more systems are deployed, periodic quantification efforts could expand the training datasets to further improve the methods.

Declarations

Author contribution statement

Robert S. Heltzel, Derek R. Johnson: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Mohammed T. Zaki, Aron K. Gebreslase: Analyzed and interpreted the data; Wrote the paper.

Omar I. Abdul-Aziz: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Funding statement

Derek Ross Johnson was supported by Division of Chemical, Bioengineering, Environmental, and Transport Systems [1804024], National Energy Technology Laboratory [DE-FE0024297].

Data availability statement

Data will be made available on request.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e11962.

References

[1] D. Zavala-Araiza, D.R. Lyon, R.A. Alvarez, K.J. Davis, R. Harriss, S.C. Herndon, A. Karion, E.A. Kort, B.K. Lamb, X. Lan, A.J. Marchese, S.W. Pacala, A.L. Robinson, P.B. Shepson, C. Sweeney, R. Talbot, A. Townsend-Small, T.I. Yacovitch, D.J. Zimmerle, S.P. Hamburg, Reconciling divergent estimates of oil and gas methane emissions, Proc. Natl. Acad. Sci. USA (2015) 201522126.

[2] S.M. Miller, S.C. Wofsy, A.M. Michalak, E.A. Kort, A.E. Andrews, S.C. Biraud, E.J. Dlugokencky, J. Eluszkiewicz, M.L. Fischer, G. Janssens-Maenhout, B.R. Miller, J.B. Miller, S.A. Montzka, T. Nehrkorn, C. Sweeney, Anthropogenic emissions of methane in the United States, Proc. Natl. Acad. Sci. USA 110 (50) (2013) 20018–20022.

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

[4] D. Johnson, R. Heltzel, D. Oliver, Temporal variations in methane emissions from an unconventional well site, ACS Omega 4 (2) (2019) 3708–3715.

[5] D. Johnson, R. Heltzel, On the long-term temporal variations in methane emissions from an unconventional natural gas well site, ACS Omega (2021) acsomega.1c00874.

[6] T.L. Vaughn, C.S. Bell, C.K. Pickering, S. Schwietzke, G.A. Heath, G. Pétron, D.J. Zimmerle, R.C. Schnell, D. Nummedal, Temporal variability largely explains top-down/bottom-up difference in methane emission estimates from a natural gas production region, Proc. Natl. Acad. Sci. U.S.A. 115 (46) (2018) 11712–11717.

[7] Thoma, E. D.; Brantley, H.; Squier, B.; DeWees, J.; Segall, R.; Merrill, R. Development of Mobile Measurement Method Series. 15.

[8] H.L. Brantley, E.D. Thoma, W.C. Squier, B.B. Guven, D. Lyon, Assessment of methane emissions from oil and gas production pads using mobile measurements, Environ. Sci. Technol. 48 (24) (2014) 14508–14515.

[9] R. Edie, A.M. Robertson, R.A. Field, J. Soltis, D.A. Snare, D. Zimmerle, C.S. Bell, T.L. Vaughn, S.M. Murphy, Constraining the accuracy of flux estimates using OTM 33A, Atmos. Meas. Tech. 13 (1) (2020) 341–353.

[10] A.M. Robertson, R. Edie, D. Snare, J. Soltis, R.A. Field, M.D. Burkhart, C.S. Bell, D. Zimmerle, S.M. Murphy, Variation in methane emission rates from well pads in four oil and gas basins with contrasting production volumes and compositions, Environ. Sci. Technol. 51 (15) (2017) 8832–8840.

[11] G. Burba, Eddy Covariance Method for Scientific, Industrial, Agricultural, and Regulatory Applications: A Field Book on Measuring Ecosystem Gas Exchange and Areal Emission Rates, LI-COR Biosciences, Lincoln, Nebraska, 2013.

[12] P.C. Stoy, A.A. Cook, J.E. Dore, W. Kleindl, E.N.J. Brookshire, Methane Efflux from an American Bison Herd 30, 2020.

[13] P. Prajapati, E.A. Santos, Measurements of methane emissions from a beef cattle feedlot using the Eddy covariance technique, Agric. For. Meteorol. 232 (2017) 349–358.

[14] R. Kormann, F.X. Meixner, An analytical footprint model for non-neutral stratification, Boundary-Layer Meteorol. 99 (2) (2001) 207–224.

[15] N. Kljun, P. Calanca, M.W. Rotach, H.P. Schmid, A simple two-dimensional parameterisation for flux footprint prediction (FFP), Geosci. Model Dev. (GMD) 8 (11) (2015) 3695–3713.

[16] C.-I. Hsieh, G. Katul, T. Chi, An approximate analytical model for footprint estimation of scalar fluxes in thermally stratifed atmospheric flows, Adv. Water Resour. 8 (2000).

[17] K. Fawagreh, M. Medhat Gaber, E. Elyan, Random forests: from early development to recent advancements, Syst. Sci. Control Eng. (2014).

[18] Z. Wang, Z. Zhao, C. Wang, Random forest analysis of factors affecting urban carbon emissions in cities within the Yangtze River Economic Belt, PLoS One (2021).

[19] J. Mascaro, G.P. Asner, D.E. Knapp, T. Kennedy-Bowdoin, R.E. Martin, C. Anderson, A tale of two “forests”: random forest machine learning aids tropical forest carbon mapping, PLoS One 9 (1) (2014), e85993.

[20] R.L.M. Tavares, S.R.D.M. Oliveira, F.M.M.D. Barros, C.V.V. Farhate, Z.M.D. Souza, N.L. Scala Junior, Prediction of soil CO2 flux in sugarcane management systems using the Random Forest approach, Sci. Agric. 75 (4) (2018) 281–287.

[21] A. Philibert, C. Loyce, D. Makowski, Prediction of N2O emission from local information with Random Forest, Environ. Pollut. 177 (2013) 156–163.

[22] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, M. Dehmer, An introductory review of deep learning for prediction models with big data, Front. Artif. Intell. (2020).

[23] L. Alfaseeh, R. Tu, B. Farooq, M. Hatzopoulou, Greenhouse gas emission prediction on road network using deep sequence learning, Transport. Res. Transport Environ. 88 (2020).

[24] C. Arif, B. Indra Setiawan, N.A. Iswati Hasanah, Predicting methane emission from paddy fields with limited soil data by artificial neural networks, in: 2020 International Conference on Computer Science and its Application in Agriculture (ICOSICA), 2020, pp. 1–4.

[25] S. Russel, C. Vines, G. Borher, D. Johnson, J. Villa, R. Heltzel, C. Rey-Sanchez, J.H. Matthes, Quantifying CH4 concentratino spikes above baseline and attributing CH4 sources to hydraulic fracturing activities by continous monitoring at an off-site tower, Atmos. Environ. 228 (2020).

[26] B. Travis, M. Dubey, J. Sauer, Neural networks to locate and quantify fugitive natural gas leaks for MIR detection system, Atmos. Environ.: X (2020).

[27] J. Wang, L.P. Tchampi, A.P. Ravikumar, M. McGuire, C.S. Bell, D. Zimmerle, S. Savarese, A.R. Brandt, Machine Vision for natural gas methane emissions detection using an infrared camera, Appl. Energy (2020).

[28] LI-COR LI-7700. LI-COR Biosciences.

[29] Gill Instruments Limited, WindMaster 3-Axis Ultrasonic Anemometer. http://www.gillinstruments.com/products/anemometer/windmaster.htm (accessed 2020-01-02).

[30] R.S. Heltzel, M.T. Zaki, A.K. Gebreslase, O.I. Abdul-Aziz, D.R. Johnson, Continuous OTM 33A analysis of controlled releases of methane with various time periods, data rates and wind filters, Environments 7 (9) (2020) 65.

[31] R.S. Heltzel, On the Improvement of the Indirect Quantification of Methane Emissions: A Stationary Single Sensor Approach, Dissertation, West Virginia University, Morgantown, WV, 2021.

[32] US EPA, O. EMC Other Test Methods https://www.epa.gov/emc/emc-other-test-methods (accessed 2020-01-02).

[33] EddyPro 7 | What is EddyPro? https://www.licor.com/env/support/EddyPro/topics/introduction.html (accessed 2021-03-08).

[34] SAS: Analytics, Artificial Intelligence and Data Management. https://www.sas.com/en_us/home.html (accessed 2021-03-08).

[35] Scikit-Learn: Machine Learning in Python—Scikit-Learn 0.24.1 Documentation https://scikit-learn.org/stable/(accessed 2021-03-08).

[36] Cloud Computing Services | Microsoft Azure https://azure.microsoft.com/en-us/(accessed 2021-03-08).

[37] sklearn.ensemble.RandomForestRegressor — scikit-learn 0.24.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed 2021-03-08).

[38] Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. 25.

[39] sklearn.model_selection.RandomizedSearchCV — scikit-learn 0.24.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html (accessed 2021-03-08).

[40] M. Kuhn, K. Johnson, Applied Predictive Modeling, second ed., Springer, 2018.

[41] X.C. Nguyen, T.T.H. Nguyen, D.D. La, G. Kumar, E.R. Rene, D.D. Nguyen, S.W. Chang, W.J. Chung, X.H. Nguyen, V.K. Nguyen, Development of machine learning - based models to forecast solid waste generation in residential areas: a case study from vietnam, J. Res. Con. Rec. 167 (2021), 105381.

[42] H. Lan Vu, K. Tsun Wai Ng, A. Richter, C. An, Analysis of input set characteristics and variances on k-fold cross validation for recurrent neural network model on waste disposal rate estimation, J. Environ. Manag. 311 (2022), 114869.

[43] Model Selection - How to Choose the Number of Hidden Layers and Nodes in a Feedforward Neural Network?. https://stats.stackexchange.com/questions/181/how-to-choose-the-number-of-hidden-layers-and-nodes-in-a-feedforward-neural-netw (accessed 2021-03-08).

[44] sklearn.neural_network.MLPRegressor—Scikit-Learn 0.24.1 Documentation. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html#sklearn.neural_network.MLPRegressor (accessed 2021-03-08).

[45] R. Heltzel, D. Johnson, M. Zaki, A. Gebrelase, O.I. Abdul-Aziz, Understanding the Accuracy Limitations of Quantifying Methane Emissions Using Other Test Method 33A. Environments, 2022.

[46] Y. Kong, T. Yu, A deep neural network model using random forest to extract feature representation for gene expression data classification, Sci. Rep. 8 (2018), 16477.