Probabilistic Gas Leak Rate Estimation Using Submodular Function Maximization With Routing Constraints

Kalvik Jakkala and Srinivas Akella

***Abstract*—Harmful greenhouse gases such as methane are prone to leak during extraction, transportation, and storage in oil fields. Therefore we must monitor gas leak rates to keep such fugitive emissions in check. However, most currently available approaches incur significant computational costs to generate informative paths for mobile sensors and estimate leak rates from the collected data. As such, they do not scale to large oil fields and are infeasible for real-time applications. We address these problems by deriving an efficient analytical approach to compute the leak rate distribution and Expected Entropy Reduction (EER) metric used for path generation. Moreover, a faster variant of a submodular function maximization algorithm is introduced, along with a generalization of the algorithm to find informative data collection walks with arc routing constraints. Our simulation experiments demonstrate the approach's validity and substantial computational gains.**

***Index Terms*—Environment monitoring and management, fugitive emissions, integrated planning and learning, informative path planning, probabilistic inference.**

I. Introduction

METHANE accounted for 10% of the global greenhouse gas emissions in 2018 [1]. However, in the form of natural gas, methane is a viable energy source that can slow global emissions since it has a smaller carbon footprint than most other fossil fuels [2]. Unfortunately, it is not possible to extract without leaking, and its carbon footprint is small only if under 4% of its total production volume leaks [3]. Since methane leaks are unavoidable, we need to estimate leak rates in oil fields and take appropriate actions depending on the estimated rate.

However, estimating the leak rates of gas sources is non-trivial. Recent work [4] has shown that current methods have heavily underestimated methane leaks. The reason leak rate estimation is difficult is because it is inherently an ill-posed problem. Even if a single source is considered, multiple leak rates could result in sensing the same gas concentration at a given location as environmental factors such as wind speed and temperature could affect the dispersion of the gas. Furthermore, when multiple leak sources are considered (Fig. 1), it is possible to have overlapping gas plumes making it difficult to attribute the data to each source.

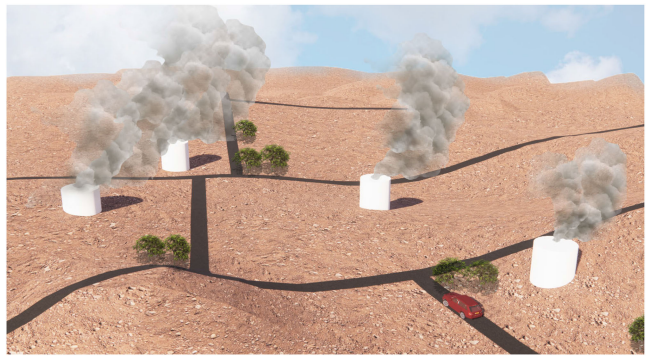


Fig. 1. Illustration of an oil field depicting storage tanks. Note that methane gas is not in the visible spectrum; it is shown gray for visualization.

Robots can collect gas concentration data from oil fields to estimate leak rates. The data collection locations have to be planned so that each location is highly informative and reachable within the robot's distance budget. So we also have to consider the informative path planning problem (IPP) for mobile robots. Even if we restrict data collection to a road network, depending on the size of the road network and distance budget, there could be a prohibitively large number of possible data collection walks.[1]

We address two main problems in this letter. First, we derive a computationally efficient probabilistic approach for estimating gas leak rates. We improve on the approach of Albertson *et al.* [5] by introducing a simplifying Gaussian assumption that results in substantial computational gains while retaining leak rate estimate convergence. Second, we address the IPP problem; we use the Generalized Cost-benefit (GCB) algorithm [6] to find data collection walks. However, the GCB algorithm does not consider arc routing constraints needed to find informative data collection walks in road networks; we introduce a variant that considers such constraints. We also present a modification to the GCB algorithm that substantially improves its runtime efficiency.

Manuscript received September 9, 2021; accepted January 16, 2022. Date of publication February 7, 2022; date of current version March 11, 2022. This letter was recommended for publication by Associate Editor H. Il Son and Editor Y. Choi upon evaluation of the reviewers' comments. This work was supported in part by NSF under Award Number IIP-1919233. *(Corresponding author: Kalvik Jakkala.)*

The authors are with the Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC 28223 USA (e-mail: kjakkala@uncc.edu; sakella@uncc.edu).

Digital Object Identifier 10.1109/LRA.2022.3149043

[1]A *walk* is any sequence of alternating vertices and edges $v_1, e_1, v_2, e_2, \ldots, v_k, e_k, v_{k+1}$ in a graph such that each edge e_i has endpoints v_i and v_{i+1}. A walk could contain repeated edges or vertices. It is considered closed if the first and last vertices are the same, and open otherwise.

This letter makes the following contributions:

1) Presents a fast and effective Bayesian approach for leak rate estimation from gas concentration data.
2) Derives an efficient analytical solution to an information metric–Expected Entropy Reduction (EER)–that is used in the IPP problem.
3) Improves the runtime efficiency of the GCB algorithm used to solve the IPP problem.
4) Introduces an arc routing variant of the GCB algorithm for IPP in graph networks.

II. Problem Statement

We are given a road network graph $G = (V, E)$ with intersections modeled as vertices V and roads as edges E. We are also given a set of candidate leak locations and environmental factors such as wind speed and ambient temperature. We need to accurately estimate the leak rate of each leak source. The problem entails identifying informative data collection walks within a distance budget b and using the collected data to estimate the leak rates. Furthermore, depending on the leak rate of each source, we can detect the gas leaks at varying distances from the source. Therefore, our goal is to find minimal length data collection walks by selectively approaching the sources and getting only as close as needed to the sources. Additionally, the estimates could have high variance or become obsolete as environmental factors and leak rates continually change over time. Thus the solution approach would have to be fast, accurate, and iterative to update the leak rate estimates whenever needed.

III. Related Work

Leak rate estimation: There are several approaches to leak rate estimation based on whether the sensors are fixed or mobile, the type of gas concentration sensors, and resolution of the sensors.

Pandey *et al.* [7] showed that it is possible to estimate methane leak rates from satellite measurements. However, satellites are expensive to deploy and maintain. The approach also has a limited ground pixel resolution and suffers occlusion from clouds, making it challenging to estimate small scale leaks omnipresent in oil fields.

An alternative to satellite data based sensing is to deploy a methane sensor at each oil well. To ascertain the feasibility of such a method, Project Astra [8] aims to build a sensor network for an entire oil field and develop a network monitoring method. However, it might be challenging to deploy and maintain a sensor network in an oil field like the Permian Basin in the US, spanning about 220,000 square kilometers with over 3500 drilled but uncompleted wells (DUC) [9], given the large number of sensors required.

Travis *et al.* [10] addressed methane leak rate estimation using fixed sensors by training a Neural Network (NN) on data from gas leak simulations. The NN could predict leak rates with reasonable accuracy but assumed stationary methane gas concentration sensors. Furthermore, the NN was an entirely data-driven, black-box approach that does not generalize to oil fields that do not follow the same leak rate distribution as the simulated data.

Albertson *et al.* [5] developed a Bayesian model for leak rate estimation in an oil field. The Generalized Extreme Value (GEV) Type II distribution [11] was used as the prior distribution in their approach. They also used Expected Entropy Reduction (EER) as an optimization metric to find data collection paths for mobile sensors. However, using the GEV distribution necessitated approximation methods like numerical quadrature to evaluate the nested integrals involved in the computation of EER and the posterior distribution of leak rates.

Furthermore, the framework of [5] is an iterative approach wherein one generates a data collection path, collects data, updates the leak rate estimates, and repeats the process until convergence to the true leak rates. However, every new iteration introduces an additional nested integral, each incurring substantial computation costs. Although the approach is theoretically elegant, it is computationally prohibitive and infeasible for large oil fields.

Informative Path Planning: Finding the most informative walk for data collection is known as the Informative Path Planning (IPP) problem. Despite its name, IPP is not limited to just paths but includes tours and walks as well.[2]

Usually, information metrics such as mutual information are used to quantify the informativeness of data collection locations in IPP. But the IPP problem is known to be NP-hard [13], and as a consequence, only suboptimal solutions can be obtained for most real-world problems.

Hollinger and Sukhatme [14] presented branch and bound techniques for IPP and established asymptotically optimal guarantees; their algorithms converge to the optimal solution as the run time approaches infinity.

A recent approach presented by Bottarelli *et al.* [15] developed active learning algorithms for the IPP problem with a complexity of $\mathcal{O}(|D|^5)$ where D is the discretized data collection space. They also suggested optimizations that trade search space complexity for reduced computation time.

Some IPP methods exploit structure in their optimization functions, mainly submodularity [16]–[19], a property often found in information metrics used in IPP. Submodular functions have a diminishing returns property that makes them amenable to greedy optimization with known approximation factors. Iyer and Bilmes [20] established tight approximation factors for the maximization of submodular functions with submodular constraints.

Zhang and Vorobeychik [6] developed the Generalised Cost-benefit (GCB) algorithm with approximation guarantees to find a subset of vertices in a graph that maximize submodular functions with node routing constraints. They imposed node routing constraints by solving the Travelling Salesperson Problem (TSP) while ensuring that the walk is within the distance budget and includes the selected vertices of the graph. In practice, the method therefore requires numerous computations of the TSP, making it relatively expensive, and the method was limited to node routing constraints.

A closely related problem is gas distribution estimation [21], [22]. In this problem, gas leak sources are at unknown locations.

[2]A walk with no repeated edges is called a *tour*, and a walk with no repeated vertices is called a *path* [12].

The task is to estimate the gas density at each region and locate the leak. The problem also entails determining the data collection locations. Arain *et al.* [23] presented an approach that discretizes the search space and solves a convex relaxation of an integer linear program for near-optimal environment coverage.

Nonetheless, our problem is intrinsically different from the conventional IPP problem. We are interested in collecting data only to predict the leak rates of potential sources in an oil field instead of building a model of the entire data collection space. Therefore, an optimal walk for our problem might be substantially different from an optimal walk for the conventional IPP problem.

Albertson *et al.* [5] addressed our variant of the IPP problem by iterating over every possible path within the distance budget b between a given start and end location. The authors then computed each path's EER and selected the maximal EER path as the solution. However, a road network will have exponentially many possible walks, some of which might not even go near any leak source in the field. As such, the method incurs exponential computational costs to find the solution route and is feasible only for small road networks.

Arc Routing: Arc Routing Problems (ARP) [12] are closely related to TSP. But unlike node routing problems such as TSP that look for a walk that visits all nodes, ARPs look for a walk that traverses the arcs or edges of a graph at least once. The Rural Postman Problem (RPP) [24], a variant of ARP, is to find the shortest walk that traverses a specified subset of edges, the required edges of a graph. Since a walk needs to be continuous, RPP solvers may additionally use non-required edges.

IV. Preliminaries

Foster-Wittig *et al.* [25] developed a gas dispersion model to calculate fugitive gas concentration at any location given the leak rate of the source. The gas concentration $C(s,x,y,z)$ at the location (x,y,z) when the leak rate is s is given by:

$$C(s,x,y,z) = \frac{s}{U}\left(\frac{\bar{A}}{\overline{z}(x)} exp\left[-\left(\frac{Bz}{\overline{z}(x)}\right)^2\right]\right) \times \left(\frac{1}{\sqrt{2\pi}\sigma_y} exp\left[-\frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2\right]\right) \quad (1)$$

Here, U is the observed speed of the gas plume advection, $\bar{A}$, B, and $\overline{z}$ are functions of atmospheric stability, and σ_y is the length scale of the plume along the horizontal axis. The (x,y,z) coordinates are relative to the origin centered at the leak's source, with the x-axis along the wind direction.

Albertson *et al.* [5] used the above gas dispersion model in a Bayesian model to compute the posterior distribution of the leak rates given the methane gas concentration data from field measurements. A new instance of the Bayesian model was associated with each source.

The Bayesian model was also used to evaluate the Expected Entropy Reduction (EER) information metric φ to set up an optimization problem and find maximally informative data collection walks. EER measures mutual information [26], the amount of information one random variable contains about another. Mutual information can also be interpreted as the reduction in uncertainty of one variable due to the knowledge of another variable. Mutual information is treated as a dimensionless metric that can only be interpreted in its relative sense [26].

In [5], EER φ quantifies the information relevant to a source's leak rate in gas concentration data collected from a path. It also allows us to quantify this information without knowing the true leak rate. Here, M is the set of gas concentration data from the data collection path. Each $m \in M$ represents the measured gas concentration in parts per million (ppm). $S \subseteq \mathcal{R}_{\geq 0}$ is the domain of the leak rate s.

$$\varphi[S;M] = -\log_2 \int_{s\epsilon S} p^2(s)ds + \int_{m\epsilon M} \log_2 \int_{s\epsilon S} \left(\frac{p(m|s)p(s)}{\int_{s_1\epsilon S} p(m|s_1)p(s_1)ds_1}\right)^2 ds\, p(m)dm \quad (2)$$

EER is submodular [27]. Submodular functions are set functions with returns that diminish as the input set size increases. Any submodular function f satisfies the following property for sets X, Y, and T, with u being an element of the set T that is not already in Y.

$$f(X \cup \{u\}) - f(X) \geq f(Y \cup \{u\}) - f(Y)$$
$$\forall X \subseteq Y \subset T \text{ and } u \in T \backslash Y$$

Consider the EER function–adding more locations to a data collection path will increase the EER. However, the size of the incremental increases in the EER will diminish as the number of data points increases, as the amount of additional information in a path decreases with each newly collected data sample.

Like prior approaches, we assume that the oil field is on flat terrain without any large obstacles obstructing the gas plumes. Source detection is done by thresholding the leak rate of every well. Table I lists all the variables used in this letter along with their definitions.

V. Approach

Our approach assumes a Gaussian prior for the leak rates, which we use to derive an analytical EER and posterior for the leak rates; we use the analytical EER to perform informative path planning.

A. Gaussian Assumption

To avoid the drawbacks of using the GEV Type II distribution as the prior over the leak rates, we instead use the Gaussian distribution as the prior. The Gaussian distribution is conjugate–if the likelihood is Gaussian, using a Gaussian prior over its mean will result in a Gaussian posterior. Moreover, its mean and variance compactly parameterize the distribution and are amenable to analytical computations.

The Gaussian prior assumption facilitates our derivation of an analytical equation to compute the EER and the posterior in time linear in the number of gas concentration samples collected from the field. Since Gaussian distributions have the maximal likelihood at the mean, instead of sampling the entire domain of s as was done in [5], we only have to compute the mean

TABLE I
DEFINITIONS OF VARIABLES

Variable	Definition
$G = (V, E)$	Road network graph G, with vertices V and edges E
b	Distance budget
s	Leak rate of a source
S	Domain of leak rates
C	Gas dispersion model or gas concentration function
x, y, z	Coordinates along X, Y, and Z axes respectively
U	Observed speed of gas plume advection
$\hat{A}$, B, $\hat{z}$	Functions of atmospheric stability
σ_y	Length scale of gas plume along the Y axis
φ	Expected Entropy Reduction (EER)
M	Set of gas concentration data
m	Individual gas concentration data sample
$\mathcal{N}$	Normal distribution
μ_s	Mean of leak rate s of Gaussian prior
σ_s	Standard deviation of leak rate s of Gaussian prior
σ_e	Combined error of gas dispersion model and data measurement
$\mathcal{A}$	Function of all the terms except s in the gas dispersion model C
A_{xyz}	Scalar output of the function $\mathcal{A}$ at location x, y, z
$\mathbf{A}$	Vector containing A_{xyz} values computed at different locations
γ, μ, β	GEV Type II distribution parameters
$\hat{c}$	Routing problem solver (returns solution route cost)

to determine the most likely leak rate s (i.e., the maximum a posteriori probability estimate).

However, assuming a Gaussian distribution for the prior has its shortcomings. It is not consistent with the results of Brantley *et al.* [28] whose findings showed that the leak rates follow a log-normal distribution. Nevertheless, we found that the computational gains from computing both the EER and posterior analytically outweigh aligning the prior to a log-normal distribution. Moreover, our experiments show that our approach converges to the simulated leak rate despite the Gaussian prior. We next describe the critical steps in our derivations.

B. Analytical EER and Posterior

We formulate the prior $p(s)$ and likelihood $p(m|s)$ functions of the leak rate s as follows.

$$p(s) \sim \mathcal{N}(\mu_s, \sigma_s^2)$$

$$p(m|s) = \frac{1}{\sigma_e\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{m - C(s,x,y,z)}{\sigma_e}\right)^2\right] \quad (3)$$

Here, μ_s and σ_s are the mean and standard deviation of the leak rate s. The combined gas dispersion model and concentration measurement error is σ_e. The gas concentration m is measured at coordinates x, y, z.

The evaluation of EER and the posterior involves integrating the likelihood function $p(m|s)$ with respect to s. However, the likelihood $p(m|s)$ contains the gas dispersion model C (1), which is a function of numerous parameters and seemingly intractable to analytical integration.

We found that the dispersion model C can be factorized into a product of s and the remaining terms independent of s. Therefore, we can combine all the terms other than s into a single function $\mathcal{A}$, dependent on the (x, y, z) coordinates where the dispersion is calculated. This factorization allows us to treat the output of the function $\mathcal{A}(x, y, z)$ as a constant with respect to the likelihood $p(m|s)$, allowing us to treat the likelihood as a Gaussian. And since both the prior $p(s)$ and likelihood $p(m|s)$ are Gaussian, our posterior $p(s|m)$ will also follow a Gaussian distribution. We omit the arguments of $\mathcal{A}$ for brevity.

$$p(m|s) = \frac{1}{\sigma_e\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{m - s\mathcal{A}(x,y,z)}{\sigma_e}\right)^2\right]$$
$$= \mathcal{N}\left(s\mathcal{A}, \sigma_e^2\right) \quad (4)$$

We derived[3] the EER φ and Gaussian posterior $p(s|M)$ using the factorized likelihood $p(m|s)$ and Gaussian prior $p(s)$. Here, M is a vector containing gas concentration data m at every sampling location. A_{xyz} is the scalar output of the function $\mathcal{A}$ computed for the sampling location with coordinates (x, y, z), $\mathbf{A}$ is the vector representation of all the A_{xyz} values, and c is a constant.

$$\varphi[S; M] = -\log_2\left(\frac{1}{2\sigma_s\sqrt{\pi}}\right) + \sum_{m\in M} \frac{(m - \mu_s A_{xyz})^2}{2(A_{xyz}^2\sigma_s^2 + \sigma_e^2)} + c$$

$$p(s|M) \propto \mathcal{N}\left(\frac{M^T\mathbf{A}\sigma_s^2 + \mu_s\sigma_e^2}{\mathbf{A}^T\mathbf{A}\sigma_s^2 + \sigma_e^2}, \frac{\sigma_e^2\sigma_s^2}{\mathbf{A}^T\mathbf{A}\sigma_s^2 + \sigma_e^2}\right) \quad (5)$$

Note that while computing the EER, M is calculated using a simulated leak source (1) with an arbitrary leak rate since we are only interested in the amount of information in a data collection walk, independent of the actual leak rate. However, to compute the posterior leak rate, we use gas concentration data collected from the generated data collection route.

Moreover, unlike the GEV prior model, the Gaussian prior model's posterior can be represented by its mean and variance. So when the posterior needs to be used as the new prior distribution to determine subsequent data collection walks, it will not introduce any nested integrals as we can update the prior by changing its mean and variance.

C. Informative Path Planning (IPP)

Our approach to evaluating EER and the posterior substantially decreases the computation time. However, we still need to solve the IPP problem. Using the EER function as an optimization metric, we wish to find a data collection walk that maximizes the EER. Such a walk will result in the most informative sensor data for leak rate estimation.

We reformulate the problem as one where we have to find the edges in the graph G that maximize the aggregate EER and find a walk within the distance budget that includes all the selected edges. This problem, belonging to the class of arc routing problems with profits, is NP-hard [12].

Since EER is a submodular function, we could use the GCB algorithm [6] to maximize EER while imposing routing constraints. However, the GCB algorithm operates on nodes and imposes only node routing constraints. We need arc routing constraints, as the EER function operates on edges to quantify

[3]The derivation and code can be found at https://github.com/UNCCharlotte-CS-Robotics/Gas-Leak-Estimation

Algorithm 1: The modified Generalized Cost-benefit Algorithm (MGCB). $\hat{c}$ is the routing function: TSP when W is the node set V of the graph G, and ARP when W is the edge set E. φ is the submodular cost function (EER), and $W'\backslash x^*$ is the set W' without the element x^*.

Data: $b > 0, W$
Result: Walk $S \subset W$
1 $A := \arg\max\{\varphi(x) \mid x \in W, \hat{c}(x) \leq b\}$
2 $Z := \emptyset$
3 $W' := W$
4 **while** $W' \neq \emptyset$ **do**
5 **for** $x \in W'$ **do**
6 $\Delta^x_\varphi := \varphi(Z \cup x) - \varphi(Z)$
7 $\Delta^x_c := \hat{c}(Z \cup x) - \hat{c}(Z)$
8 **end**
9 $Y := \{x \mid x \in W', \hat{c}\,(Z \cup x) \leq b\}$
10 **if** $|Y| == 0$ **then**
11 **break**
12 **end**
13 $x^* = \arg\max\{\Delta^x_\varphi / \Delta^x_c \mid x \in Y\}$
14 $Z := Z \cup x^*$
15 $W' := W'\backslash x^*$
16 **end**
17 **return** $\arg\max_{S \in \{A,Z\}} f(S)$

information. Solving the routing problem as a node routing problem does not always give us the best solutions. Furthermore, it nullifies the approximation guarantee established for the GCB algorithm [6].

We address this problem by proposing an ARP variant of the GCB algorithm. Our ARP variant selects edges of the graph instead of vertices and imposes arc routing constraints. We set up the routing problem as the Rural Postman Problem (RPP) [24] and solve it using an RPP solver. The RPP solver finds the shortest walk within the distance budget (if one exists) while including all the edges in the subset selected by the GCB algorithm. We found that the RPP variant of the GCB algorithm often results in walks with higher or equivalent EER than those generated using the original TSP variant. Furthermore, the RPP variant retains the approximation guarantee of the GCB algorithm since both the subset selection and routing constraints operate on the edges of the graph.

Moreover, we also improve the runtime efficiency of the GCB algorithm by adding a conditional break statement (Algorithm1). Let $S \subseteq W$ be the solution set generated by the GCB algorithm. The original algorithm takes $|W|$ iterations of the while loop. In contrast, our modified GCB algorithm (MGCB) takes only $|S| + 1$ iterations of the while loop and returns the same result as the original algorithm.

The MGCB algorithm (Algorithm 1) starts by ensuring that at least one element (nodes if using TSP or edges if using ARP) is reachable within the distance budget (Line 1). Then it computes the increments in the route cost Δ^x_c and submodular function cost Δ^x_φ upon adding each available element $x \in W'$ to the solution set Z (Lines 5–8). Any infeasible routes are filtered, and it checks if any feasible elements remain that it could add to the solution route (Lines 9–12). If none remain, the algorithm returns the best-known solution up to that point. Else it adds the element x^* with the highest increment ratio (of submodular cost to route cost) to the solution set Z and removes it from the available elements set W' (Lines 13–15). The algorithm iterates until either the available elements set is empty or the condition for the break statement is satisfied (i.e., there are no more feasible elements).

VI. Simulation Experiments

This section compares the EER and posterior computations using the GEV Type II and Gaussian priors. Additionally, it illustrates the advantages of the modified GCB algorithm for IPP. Real-world experiments on gas leak rates are stochastic, influenced by environmental conditions such as wind or humidity. To control for such variations, we conducted our experiments in simulation with the fixed environmental conditions documented in [29] and used the dispersion model (1) from Foster-Wittig *et al.* [25].

In all our experiments, we used the GEV Type II prior with model parameters γ, μ, β, and σ_e set to 1, 0.19, 0.23, and 0.01 respectively, obtained from [5], and the Gaussian prior with μ_s, σ_s, and σ_e set to $0.15, 0.65$, and 0.03 respectively. The Gaussian prior's mean was estimated by fitting to data sampled from the GEV Type II distribution with the parameter values mentioned above. We can generalize this fitting process to any oil field of interest by replacing the GEV distribution data with aggregate historical leak data from that oil field. The σ_s and σ_e parameters were tuned so that the largest leak rate in our sampled data used to fit the Gaussian mean was within two standard deviations, which gave us Gaussian posterior predictions close to that of the original GEV prior model.

A. Leak Rate Posterior

First, to establish our approach's validity, we simulated a source leaking at three different leak rates and calculated gas concentrations M at ten random locations around the source. Using the GEV Type II and Gaussian prior-based approaches, we then used the simulated gas concentration data M to estimate the true leak rate s. The results are shown in Fig. 2.

We found that using either prior we can accurately estimate the true leak rate. Even when the leak rate has a low likelihood in the prior distribution, such as 5 g/s, our approach's estimates are close to the true leak rate. But the Gaussian prior based model underestimates the leak rate as the true leak rate moves far away from the prior distribution's mean.

However, this is a degenerate scenario, as we limited the data to only 10 points to show our approach's limits. Moreover, gas concentration sensors have much higher sampling rates, and according to Brantley *et al.* [28], most leaks occur at much lower rates, around 0.16 g/s, which is close to the mean μ_s of our prior distribution $p(s)$. As such, we anticipate that our approach's estimated leak rate would be closer to the true leak rate in a real-world scenario.

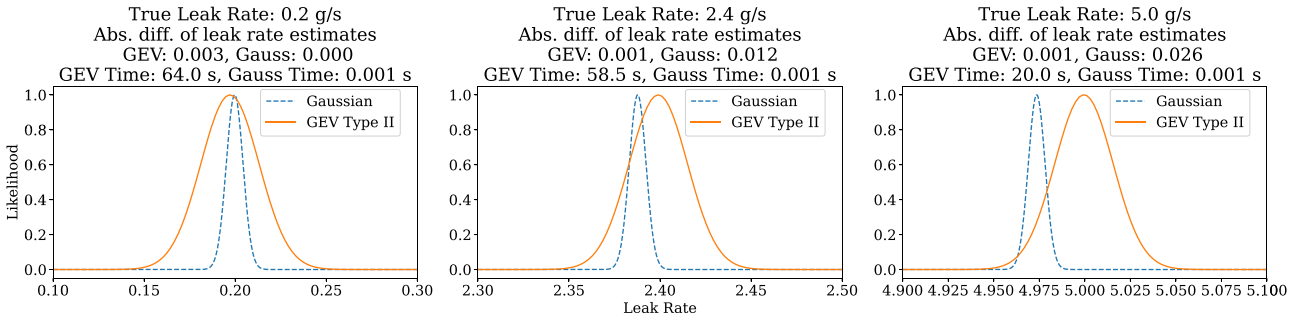


Fig. 2. Posterior leak rate distribution with GEV Type II and Gaussian priors for different true leak rates. The posterior with GEV prior is shown in orange (solid line), and the posterior with Gaussian prior is shown in blue (dashed line). Subplot titles show the true leak rate, the absolute difference between each distribution's mode and the true leak rate, and the computation time. The GEV Type II and Gaussian priors were parametrized with a mode (most likely leak rate) of 0.09 g/s and 0.15 g/s respectively.

TABLE II
EER COMPUTATION TIME WITH GEV TYPE II AND GAUSSIAN PRIORS FOR PATHS AT VARYING DISTANCES FROM THE LEAK SOURCE. THE RESULTS WERE AVERAGED OVER 10 ITERATIONS

Distance to oil well along X-axis (meters)	Compute Time (secs)	
	GEV Type II prior	Gaussian prior
0.2	22.42391	0.00017
0.8	16.90775	0.00016
1.4	30.65850	0.00017
2.4	41.51963	0.00021
3.0	42.78649	0.00017
5.0	41.21629	0.00020

Furthermore, the Gaussian prior based approach is five orders of magnitude faster than the GEV prior based approach. This is because, unlike the GEV based approach, we can analytically compute our model's posterior.

B. Expected Entropy Reduction (EER)

The following experiment details the computational gains of our approach when computing the EER. We are interested in paths with the most information about leak rates quantified by EER. The EER computation cost is substantial when using the GEV Type II prior. We demonstrate the advantages of the Gaussian prior based approach to EER computation by evaluating the EER of six paths and sorting them in decreasing order. The EER values are not interpretable by themselves. Therefore, we are only interested in the order of the paths sorted by EER. We considered straight-line paths (each with 100 evenly sampled data points) parallel to the Y-axis at different distances along the X-axis from a simulated leak source.

We found that using either prior distribution to compute the EER gives the same ordering of paths. However, the Gaussian prior based model took five orders of magnitude less computation time compared to the GEV model, as shown in Table II. Furthermore, our approach's benefits multiply in an actual oil field where hundreds of leak sources are considered, and the computation is repeated for numerous paths. We also observed fluctuations in the GEV model's computation time due to the stochastic nature of adaptive quadrature used to evaluate the integrals in its EER. In contrast, the Gaussian prior model takes an almost constant amount of time to compute the EER, given the analytical solution to its integrals.

Our results above establish that our method converges to the true simulated leak rate with significantly reduced computation time despite the Gaussian assumption.

TABLE III
STATISTICS OF THE GRAPHS USED TO BENCHMARK THE MGCB ALGORITHM

Statistic	Min	Max	Mean
Number of oil wells	11	145	35.50
Number of vertices	15	420	76.28
Number of edges	16	484	81.07
Avg. connectivity [31]	1.0	1.10	1.04

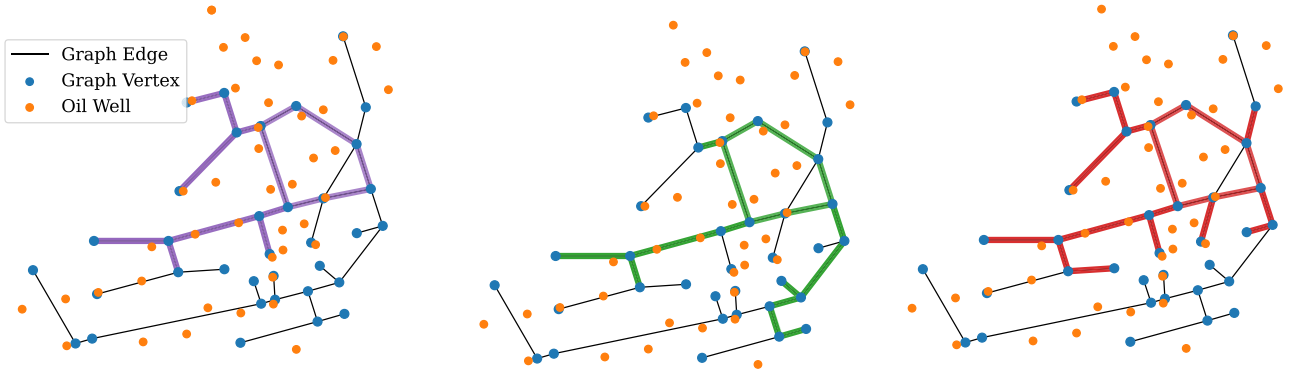
C. Informative Path Planning (IPP)

We also improved the IPP approach as mentioned in Section V-C. The following experiment establishes the improvement in computation time of the IPP approach using our modified GCB algorithm.

We considered a corpus of 80,000 oil wells in the Permian basin in Texas, USA [30]. The wells were clustered into 1000 clusters based on their relative positions. We then extracted the road network associated with each cluster. To ensure that the path iteration algorithm's runtime is feasible, we empirically chose the connectivity range and total graph distance. We filtered out graphs with average node connectivity [31] higher than 1.1 and graphs with total road network length less than two times the distance budget b, which gave us 134 graphs. The statistics of the considered graphs are shown in Table III.

We then generated data collection walks with a distance budget of 15 km, which we found to be the range feasible for selected unmanned aerial vehicles (UAVs) and unmanned ground vehicles (UGVs). We generated the walks using the path iteration approach of Albertson *et al.* [5], the GCB algorithm with TSP, our modified GCB algorithm with TSP, the GCB algorithm with RPP, and the modified GCB algorithm with RPP. The TSP [32] and RPP [24][4] solvers we used had a 3/2-approximation guarantee. To compute the EER, we sampled

[4]We used the Line Coverage Library available at https://github.com/UNCCharlotte-CS-Robotics/LineCoverage-library



(a) Path iteration (1.80E-5 MSE) (b) Modified GCB with TSP (2.20E-5 MSE) (c) Modified GCB with RPP (1.08E-5 MSE)

Fig. 3. Example graph extracted from the oil well corpus and the walk generated with each algorithm. The graph has 35 vertices, 36 edges, and 43 oil wells.

TABLE IV
LEAK RATE PREDICTION MEAN SQUARED ERROR (MSE) AND COMPUTATION TIME FOR EACH METHOD (LOWER IS BETTER). PATH ITER IS THE PATH ITERATION, GCB IS THE ORIGINAL GCB ALGORITHM, MGCB IS OUR MODIFIED GCB ALGORITHM, AND THE TSP/RPP POSTFIX REFERS TO THE ROUTING CONSTRAINT SOLVER USED IN GCB

Method	Mean MSE	Std. Dev. of MSE	Mean Time (secs)	Std. Dev. of Time
Path Iter	1.6132E-01	1.2989E+00	1203.81	4.18
GCB TSP	2.0583E-04	1.4948E-03	1209.58	22.64
GCB RPP	2.5680E-05	1.3238E-04	1211.50	17.45
MGCB TSP	2.0583E-04	1.4948E-03	129.41	305.48
MGCB RPP	**2.5680E-05**	1.3238E-04	450.02	512.24

the gas concentration at ten evenly spaced points along each edge in the route generated by the routing algorithm.

In this experiment, we used the Gaussian prior based EER computation method as it would be far too expensive to compute EER with the GEV prior. We randomly sampled the oil well leak rates from a uniform distribution over the range 0 to 6 g/s and assumed that we were given no prior leak rate estimates. Therefore we used the default μ_s and σ_s obtained from historical leak rate data [5]. Furthermore, we allocated a maximum of 20 mins to each algorithm for each graph. The mean squared error (MSE) of the leak estimates for each method and the computation times are shown in Table IV. Additionally, one of the generated graphs, along with its walks, is shown in Fig. 3.

Path iteration is infeasible for large graphs as the number of possible walks increases exponentially with the number of graph edges; this is reflected in its computation time. In almost all cases, it terminates with a timeout and returns the best-known solution up to that point, which also explains the low standard deviation of computation time.

In contrast, the GCB and MGCB algorithms' computational complexity does not grow exponentially. But GCB is still costly to compute and usually terminates with a timeout and returns the best-known solution up to that point, thereby underutilizing the distance budget. This also explains the low standard deviation of GCB's computation time.

Even though GCB finds the solution early on in its computation, it continues to iterate through the algorithm as there is no test to detect convergence and terminate the algorithm. However, MGCB converges to the same solution as the GCB algorithm in a fraction of the time on all considered graphs.

We notice a higher standard deviation in the computation time of MGCB because the graphs are of varying sizes, therefore taking a varying amount of time to solve. Finally, we also note that the RPP variant of GCB performs better than the TSP variant as the arc routing constraints align with the EER function that measures information along the edges of the graph.

VII. DISCUSSION

Suppose the routing constraint solver used in MGCB is stochastic, which is sometimes the case when using heuristics to solve the TSP/ARP. In this case, one should expect to see the original GCB algorithm improve its solution even after $|S|+1$ iterations of the algorithm. This is because even though the GCB algorithm converges to the solution subset of nodes/edges in $|S|+1$ iterations, the algorithm keeps solving the routing problem with the same required solution subset for $|W|-|S|-1$ iterations. Heuristic-based solvers usually find better solutions after a few such iterations. However, one could always use an exact solver once MGCB converges to improve the solution and get similar results.

Also, note that even though the MGCB algorithm takes $|S|+1$ while loop iterations, $|S|$ could still be close to $|W|$ in the worst case. This would be the case when the distance budget b is large enough, and the total distance of the graph edges is small enough that the solution could traverse the whole graph. Nonetheless, our approach does not incur any significant additional computation costs. As such, it results in the same computation time as the original GCB algorithm.

A limitation of our work is that our experiments were conducted only in simulation. However, we did not change the gas dispersion model, which is the only component influenced by real-world conditions. Since the validity of the dispersion model and the Bayesian approach was already established by Albertson

et al. [5], we believe field experiments would be consistent with our simulations.

Additionally, the dispersion model we considered assumed flat terrain and steady state environmental conditions. One could potentially develop a more sophisticated dispersion model that can be factorized into $\mathcal{A}$ and s to handle dynamic environmental conditions and use it in our approach.

VIII. Conclusion

We presented a method for efficient and accurate gas leak rate estimation of greenhouse gases such as methane. We derived a closed-form equation for EER, a mutual information metric, and substantially improved the runtime efficiency of the GCB algorithm used to maximize the EER to find informative data collection walks. Since the GCB algorithm did not consider arc routing constraints, we presented a GCB variant that addressed such constraints. We also derived an efficient analytical approach for computing the posterior distribution of the gas leak rate for each leak source.

Our simulated experiments, using oil well data, established the convergence of our approach to the true leak rate. We also showed that our approach computes the EER and posterior leak rates five orders of magnitude faster than the prior approach. Furthermore, our modified GCB algorithm (MGCB) was shown to be at least an order of magnitude faster than the original GCB algorithm. Finally, we demonstrated that our ARP variant of the MGCB algorithm obtains data collection walks for oil fields that on average result in more accurate leak rate estimates when compared to the original GCB algorithm.

The TSP/ARP routing algorithm is invoked numerous times, with only one new element added to the required set in each iteration. One could reduce this computation cost by incorporating incremental solutions in each iteration. We plan to address this in our future work.

Acknowledgment

We thank Saurav Agarwal for thoughtful discussions and making his RPP code available in the Line Coverage Library. We also thank Eben Thoma and Sayantan Datta for helpful comments and Geethika Jakkala for creating Fig. 1.

References

[1] "Inventory of US greenhouse gas emissions and sinks," U. S. Environmental Protection Agency, Tech. Rep. EPA 430-R-21-005, 2021.

[2] K. Hayhoe, H. S. Kheshgi, A. K. Jain, and D. J. Wuebbles, "Substitution of natural gas for coal: Climatic effects of utility sector emissions," *Climatic Change*, vol. 54, no. 1, pp. 107–139, 2002.

[3] D. Farquharson, P. Jaramillo, G. Schivley, K. Klima, D. Carlson, and C. Samaras, "Beyond global warming potential: A comparative application of climate impact metrics for the life cycle assessment of coal and natural gas based electricity," *J. Ind. Ecol.*, vol. 21, no. 4, pp. 857–873, 2017.

[4] R. A. Alvarez *et al.*, "Assessment of methane emissions from the U.S. oil and gas supply chain," *Science*, vol. 361, no. 6398, pp. 186–188, 2018.

[5] J. D. Albertson *et al.*, "A mobile sensing approach for regional surveillance of fugitive methane emissions in oil and gas production," *Environ. Sci. Technol.*, vol. 50, no. 5, pp. 2487–2497, 2016.

[6] H. Zhang and Y. Vorobeychik, "Submodular optimization with routing constraints," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 819–825.

[7] S. Pandey *et al.*, "Satellite observations reveal extreme methane leakage from a natural gas well blowout," *Proc. Nat. Acad. Sci.*, vol. 116, no. 52, pp. 26376–26381, 2019.

[8] D. Allen, "Project Astra," The University of Texas at Austin, 2020. [Online]. Available: https://dept.ceer.utexas.edu/ceer/astra/index.cfm

[9] "Drilling Productivity Report," U.S. Energy Information Administration (EIA), 2021. [Online]. Available: https://www.eia.gov/petroleum/drilling/pdf/dpr-full.pdf

[10] B. Travis, M. Dubey, and J. Sauer, "Neural networks to locate and quantify fugitive natural gas leaks for a MIR detection system," *Atmospheric Environ.: X*, vol. 8, pp. 100092–100104, 2020.

[11] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An Introduction to Statistical Modeling of Extreme Values*, vol. 208. Berlin, Germany: Springer, 2001.

[12] A. Corberan and G. Laporte, Eds., *Arc Routing: Problems, Methods, and Applications*. Philadelphia, PA, USA: SIAM, 2014.

[13] G. Hollinger and S. Singh, "Proofs and experiments in scalable, near-optimal search by multiple robots," in *Proc. Robot.: Sci. Syst. IV*, 2009, pp. 206–213.

[14] G. A. Hollinger and G. S. Sukhatme, "Sampling-based robotic information gathering algorithms," *Int. J. Robot. Res.*, vol. 33, no. 9, pp. 1271–1287, 2014.

[15] L. Bottarelli, M. Bicego, J. Blum, and A. Farinelli, "Orienteering-based informative path planning for environmental monitoring," *Eng. Appl. Artif. Intell.*, vol. 77, pp. 46–58, 2019.

[16] A. Krause and C. Guestrin, "Submodularity and its applications in optimized information gathering," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 4, Jul. 2011, Art. no. 32.

[17] C. Chekuri and M. Pal, "A recursive greedy algorithm for walks in directed graphs," in *Proc. 46th Annu. IEEE Symp. Foundations Comput. Sci.*, 2005, pp. 245–253.

[18] A. Singh, A. Krause, C. Guestrin, and W. J. Kaiser, "Efficient informative sensing using multiple robots," *J. Artif. Int. Res.*, vol. 34, no. 1, pp. 707–755, Apr. 2009.

[19] J. Binney, A. Krause, and G. S. Sukhatme, "Optimizing waypoints for monitoring spatiotemporal phenomena," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 873–888, 2013.

[20] R. Iyer and J. Bilmes, "Submodular optimization with submodular cover and submodular knapsack constraints," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2436–2444.

[21] P. P. Neumann, V. H. Bennetts, A. J. Lilienthal, M. Bartholmai, and J. H. Schiller, "Gas source localization with a micro-drone using bio-inspired and particle filter-based algorithms," *Adv. Robot.*, vol. 27, no. 9, pp. 725–738, 2013.

[22] C. Stachniss, C. Plagemann, and A. J. Lilienthal, "Learning gas distribution models using sparse gaussian process mixtures," *Auton. Robots*, vol. 26, no. 2, pp. 187–202, Apr. 2009.

[23] M. A. Arain, V. H. Bennetts, E. Schaffernicht, and A. J. Lilienthal, "Sniffing out fugitive methane emissions: Autonomous remote gas inspection with a mobile robot," *Int. J. Robot. Res.*, vol. 40, no. 4-5, pp. 782–814, 2021.

[24] G. N. Frederickson, "Approximation algorithms for some postman problems," *J. Assoc. Comput. Mach.*, vol. 26, no. 3, pp. 538–554, 1979.

[25] T. A. Foster-Wittig, E. D. Thoma, and J. D. Albertson, "Estimation of point source fugitive emission rates from a single sensor time series: A conditionally-sampled gaussian plume reconstruction," *Atmospheric Environ.*, vol. 115, pp. 101–109, 2015.

[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1991.

[27] C. Cai and S. Ferrari, "Information-driven sensor path planning by approximate cell decomposition," *IEEE Trans. Syst., Man, Cybern., Part B. (Cybern.)*, vol. 39, no. 3, pp. 672–689, Jun. 2009.

[28] H. L. Brantley, E. D. Thoma, W. C. Squier, B. B. Guven, and D. Lyon, "Assessment of methane emissions from oil and gas production pads using mobile measurements," *Environ. Sci. Technol.*, vol. 48, no. 24, pp. 14508–14515, 2014.

[29] E. Thoma and B. Squier, "OTM 33 and OTM 33A Geospatial Measurement of Air Pollution-Remote Emissions Quantification-Direct Assessment (GMAP-REQ-DA)," *U.S. Environmental Protection Agency*, 2014.

[30] "2016 U.S. Oil & gas activity," FracTracker Alliance, Sep. 2017. [Online]. Available: https://www.fractracker.org/map/national/us-oil-gas/

[31] L. W. Beineke, O. R. Oellermann, and R. E. Pippert, "The average connectivity of a graph," *Discrete Math.*, vol. 252, no. 1, pp. 31–45, 2002.

[32] L. Perron and V. Furnon, "OR-Tools," Google, [Online]. Available: https://developers.google.com/optimization/