Preamble

This page is supposed to contain abbreviations, acronyms, and symbols with links to those in text. Not currently working as of 08 Feb 2019.
The following is a list of notation used in the model.

1. $F = [0, b_1] \times [0, b_2]$–The observation field.
2. $f : \mathbb{R} \to \mathbb{R}$–The measurement certainty function based on distance to a measured site.
3. $\mathcal{M}$–An index set of the sensors.
4. P–An index set of the discretized field.
5. $G \subseteq F$–The gravel terrain (or other terrain) in which sensors cannot be placed.
6. r–The minimum distance between sensors
7. $\widehat{\mathcal{D}} : F \to \mathbb{R}$–The estimated isotope level at a site based on the measurements taken.
8. $\mathcal{D} : F \to \mathbb{R}$–The true isotope level at a site.

1 Overview

1.1 Introduction

The Earth remains a habitable planet largely in part to the greenhouse gases (GHG) such as methane (CH_4) and carbon dioxide (CO_2) that insulate the Earth by absorbing and emitting infrared radiation. Properly accounting for the pathways of GHGs into the atmosphere is essential to mitigating climate warming.

Studies and long-term monitoring for changes in CO_2 have been heavily documented for decades (**Keeling1960**; **Friedlingstein2014**). However, CH_4, is also an important GHG; one molecule of CH_4 traps 86 times more thermal energy than a molecule of CO_2 over a 20-year period and there are even claims that estimate is too low (**Howarth2015**; **Bridgham2013**). This ratio, also known as global warming potential has stimulated research during the last decade or so on the routes CH_4 takes to enter the atmosphere. Though human-derived, or anthropogenic, sources are important to quantify, natural reservoirs of CH_4 contribute about 51% of total emissions in the previous decade (**Kirschke2013**). But estimates for natural sources currently contain a greater degree of uncertainty than their anthropogenic counterparts (**Kirschke2013**).

Wetlands are the dominant source of natural production of CH_4, but they alone only account for about 60% of natural emission, leaving the remaining 40% to microbial activity in the soil microbes, termites, and even the deep ocean (**Kirschke2013**; **Etiope2002a**). Another important natural source of CH4 are derived from ongoing geologic processes, otherwise known as geogenic. Currently, geogenic emissions are estimated to produce 30-70 Mt CH_4 yr^{-1}, which is on par with other non-wetland natural sources (**Etiope2002a**). This estimate was calculated by measuring CH_4 from active points of emission (e.g. vents and fumaroles) and diffuse areas (thermally altered earth).

However, previous estimates are only from European locations, thus stimulating a need for diffuse CH_4 measurements in N. America. This research seeks to further constrain estimates of CH_4 emissions by providing the first geologic measurements of diffuse CH_4 in N. American volcanic caldera and hydrothermal environments, specifically, Valles caldera (VC) and Yellowstone caldera (YC). Furthermore, collection of CH_4 and CO_2 emissions from numerous sites within VC and YC will permit the characterization of spatial and temporal variability; upon integrating that variability we will be able to: **estimate total emissions, assess correlations between emissions and/or geyser activity with time and space to constrain subsurface transport mechanisms, and lastly, make overall comparisons between the two calderas to highlight controls between emission rates**.

1.2 Site Description

A collapsed caldera is a large, basin-shaped depression formed by the collapse of the roof of a volcano following a large eruption. Valles Caldera is located in northcentral New Mexico and is one of several features within the Jemez Mountain Volcanic Field. VC initially erupted 1.14 million years ago, creating a caldera 20 km in diameter (**Goff2009**). Yellowstone Caldera is

located in northwestern Wyoming and is a product of the North American plate moving west across a stationary mantle hotspot, causing linear chain of volcanic eruptions (**Smith1994**). An eruption 640 thousand years (kya) created YC and left behind a caldera that is roughly 64 km across. The most recent eruption at VC was 40 kya and for YC, 70 kya; these recent eruptions coupled with active hydrothermal systems reveal that both calderas are supplied with anonymously high subsurface heat sources.

However, one important difference between the sites is that YC is more active than VC, which is exemplified with the greater abundance of thermal features, higher surface heat flow, and larger and more frequent earthquakes (**Lowenstern2008**). Together, VC and YC are two of the three major Quaternary calderas in the US and both are considered active volcanoes, making them relevant candidates for quantifying the contribution of geologic CH4 into the atmosphere.

1.3 Research Design & Data Collection Methods

Emissions will be measured by capturing CH4 and CO2 as it migrates from the subsurface toward the atmosphere using a chamber sealed tightly to the surface. This Eosense autochamber (eosAC) will be connected to a Picarro G2201-i Cavity Ring-Down Spectrometer (CRDS) that can measure both gases rapidly (4 Hz) and in situ for concentrations ($[CH_4]$ and $[CO_2]$) and carbon isotopic composition (δ^{13}C-CH_4 and δ^{13}C-CO_2). Over several years, I have created a successful methodology that I will continue to use for these measurements. The concentration data will be necessary for making emission estimates and the carbon isotopic composition will be essential for characterizing the source of the gases and the temperature at which these gases formed.

Research shows the effectiveness of the CRDS-eosAC method (**Christiansen2015a**), but mobility is limited due to weight of the equipment (> 54 kg). For sites that I cannot access with the CRDS-eosAC method, I will use static PVC chambers (**Livingston2006**). Gas samples are withdrawn from the chamber headspace with a syringe and injected into a vial. CH_4 concentration and δ^{13}13C-CH4, and δ^{2}H-CH_4 will be shipped to UC Davis for GC-IRMS analysis. CO2 concentration and 13C-CO2 will be measured using a GC-MS at the Center for Stable Isotopes (Univ. of New Mexico) with our collaborator, Prof. Tobias Fischer.

2 Sensor Optimization Formulation

We formulate a nonlinear optimization algorithm to place sensors in the field in order to characterize the gaseous output in a given area. The model captures phenomena intrinsic to the "area" of the specific application and the "sensor" hardware. For instance, it is known that once a sensor's location has been chosen, the sensor can continuously record the concentration of carbon dioxide ($[CO_2]$), methane ($[CH_4]$), and carbon isotopes of methane ($\delta^{13}C-CH_4$), and carbon dioxide ($\delta^{13}C-CO_2$) measurements. Each of these measurements can be used to estimate, with relatively high precision (**Christiansen2015**), the isotopic composition at a specific location. That is, if $\widehat{\mathcal{D}}_{meas}$ is a function that represents the estimated **isotope** level after a complete set of measurements, and $\mathcal{D}$ is a function that represents the

true isotope level, then the difference between the two function values at a measured location is small. Hence, we make the following assumption:

A1. *Let x be a chosen measurement site. Then $\mathcal{D}(x) = \widehat{\mathcal{D}}(x)$.*

We should be able to relax this assumption to say $||\mathcal{D}(x) - \widehat{\mathcal{D}}(x)|| < \epsilon$.

$$\min \sum_{p\in P}\sum_{m\in\mathcal{M}} \frac{f(x)}{|\mathcal{M}|} \tag{1a}$$

$$\text{s.t.} x_m \in F\backslash G, \forall\ m \in \mathcal{M}, \tag{1b}$$

$$x \in \mathcal{P}, \forall\ m, m' \in \mathcal{M}, m \neq m'. \tag{1c}$$

General constraints are represented by $x \in \mathcal{P}$.

There are multiple choices of objectives functions one can use to model the decay in a measurement value as a function of the distance to the specified site. For ease of exposition, we introduce variables $y_{p,m} = x_m - X_p$, where x_m is the location of a sensor and X_P is a desired measurement site. Note that this produces an affine constraint in the formulation.

Consider $f_1 : \mathbb{R}^2 \to \mathbb{R}$ defined by $f_1(z) = ||z||_2^2$. Then f_1 is a convex function. Define $F_1 : \mathbb{R}^{2\times|\mathcal{M}|} \times \mathbb{R}^{2\times|\mathcal{M}|\times|P|}$ by $F_1(x, y) = \sum\limits_{p\in P}\sum\limits_{m\in\mathcal{M}} \frac{1}{\mathcal{M}} f_1(y_{p,m})$. Then F_1 is a convex objective function. A similar result occurs if $f_2 : \mathbb{R}^2 \to \mathbb{R}$ is defined by $f_2(z) = e^{||z||_2^2}$ and $F_2 : \mathbb{R}^{2\times|\mathcal{M}|} \times \mathbb{R}^{2\times|M|\times|P|}$ is defined by $F_2(x, y) = \sum\limits_{p\in P}\sum\limits_{m\in M} \frac{1}{M} f_2(y_{p,m})$.

A common objective function is the *Gaussian*. Let $f_3 : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f_3(z) = e^{-\frac{||z||_2^2}{\sigma}}$, where $\sigma > 0$ is a parameter. The Gaussian function f_3 is a log-concave function; hence, it is a quasiconcave function. Let $F_3 : \mathbb{R}^{2\times|M|} \times \mathbb{R}^{2\times|M|\times|P|}$ be defined by $F_3(x, y) = \prod\limits_{p\in P}\prod\limits_{m\in M} \frac{1}{M} f_3(y_{p,m})$. Then, F_3 is the product of log-concave functions and is log-concave. Moreover, $-F_3$ is quasiconvex.

2.1 A Connection to K-means

Formulation (1) can be adapted slightly into a special case of K-means. To do so, we introduce binary variables and "big-M" constraints that model which sensor is closest to each location. In addition, we first assume that $G = \emptyset$ and that F is a convex set.

$$\min \sum_{p\in P}\sum_{m\in\mathcal{M}} \frac{y_{p,m}}{|\mathcal{M}|} \tag{2a}$$

$$\text{s.t. } x_m \in F \forall\ m \in \mathcal{M}, \tag{2b}$$

$$y_{p,m} \geq ||x_m - X_p||_2^2 - M(1 - z_{p,m}), \tag{2c}$$

$$\sum_{m\in\mathcal{M}} z_{p,m} = 1, \tag{2d}$$

$$y_{p,m} \geq 0, \forall\ m \in \mathcal{M}, p \in P, \tag{2e}$$

$$z_{p,m} \in \mathbb{B}, \forall\ m \in \mathcal{M}, p \in P. \tag{2f}$$

Conjecture. *There exists an optimal solution* (x^*, y^*, z^*) *to* (2) *such that* x^*_m *is the mean of* $\{X_p\}_{P(m)}$*, where* $P(m) = \{p \in P : z_{p,m} = 1\}$.[1]

In general, $G \neq \emptyset$. Hence, it is possible that the the K-means solution is infeasible. This leads to a *constrained K-means* model. There are examples of constrained K-means models in the literature. For instance, (**Wagstaff2001**) considers constraints that either ensure two instances are assigned to the same cluster or prevent two instances from sharing a cluster. (**Luo2003**) considers a spatially constrained K-means algorithm with applications to image segmentation. Spatial constraints are useful in image segmentation because they require that the clusters are contiguous; in an image, the pixels in each cluster form a connected subset.

To the best of our knowledge,[2] this is the first study of what we call *terrain constrained* K-means: the centroids are constrained away from subsets of the feature space.

In addition, we also propose using alternative distance metrics from the Euclidean norm, as proposed by (**Aggarwal2001**).

2.2 Reverse Convex Programming

Due to the constraint (1b), this optimization will probably be at best a Reverse Convex Program (**Hillestad1980**). Further reading to be done at `https://link.springer.com/content/pdf/10.1007/BF01442883.pdf`. **Hillestad1980** describe how to identify basic solutions and they also provide a cutting plane algorithm.

If one assumes that the untestable areas $\mathcal{G}$ in the interior of the cite can be modeled by circles, then this can be modeled by $g_k(x) = ||x - c^k||_2^2 - d^k \geq 0$, where c_k is the center of the restricted area. Because g_k is strictly convex, g_k is also strictly quasiconvex. Moreover, g_k is continuous so the constraint is a reverse convex constraint. Other restrictions are also possible. For example, an ellipsoidal constraint can be modeled as $g_k(x) = ||x - c^k||_{E_k}^2 - d^k \geq 0$, where E_k is a diagonal matrix in $\mathbb{R}^2_{++}$ in which its nonzero entries dictate the lengths of the semi-axes. Polyhedral restrictions are not given by reverse convex constraints; however, one can approximate them using strictly convex functions. For instance, given the reverse L_1-norm constraint $||x - c^k||_1 - d^k \geq 0$, one can approximate this constraint by $g_k(x) = ||x - c^k||_p^p - (d^k)^p \geq 0$, for some $p > 1$.

Suppose that the restricted area $\mathcal{G}$ is given by multiple ellipsoids centered around point $c_k, k = 1, ..., K$. Given that x_m represents the placement of sensor $m \in \{1, ..., M\}$ in the 2-dimensional space, the restricted area can be modeled by multiple reverse convex constraints with convex functions $g_{m,k} : \mathbb{R}^{2M} \to \mathbb{R}$ defined by $g_{m,k}(x) = ||x_m - c_k||_{E_k}^2 - d_k$, where $E_k \in \mathbb{R}^2_{++}$ is the diagonal matrix defining the ellipsoid semi-axes and $d_k \geq 0$. Then, $\mathcal{G} = \{x \in \mathbb{R}^{2M} \mid g_{m,k}(x) < 0, \text{for some } m \in \{1, ..., M\}, k \in \{1, ..., K\}\}$.

[1] Maybe we can also show that it is the unique optimal solution. Or analyze that case.

[2] Need to increase our knowledge here...

The optimization problem now has multiple reverse convex constraints:

$$\min \sum_{p\in P}\sum_{m\in\mathcal{M}} \frac{f(x)}{|\mathcal{M}|} \tag{3a}$$
$$\text{s.t.} x_m \in F, \forall\ m \in \mathcal{M}, \tag{3b}$$
$$g_{m,k}(x) \geq 0, \forall\ m \in \{1,...,M\}, k \in \{1,...,K\}, \tag{3c}$$
$$x \in \mathcal{P}, \forall\ m, m' \in \mathcal{M}, m \neq m'. \tag{3d}$$

Many reverse convex algorithms are designed for only a single reverse convex constraint (see Tuy other works to cite). It is shown in (Jacobsen Encyclopedia entry) that one can transform a problem with multiple reverse convex constraints into one with a single reverse convex constraint, and we provide details specific to this setting.

Let $s, p_{m,k}, q : \mathbb{R}^{2M} \to \mathbb{R}$, for all $k \in \{1,...,K\}, m \in \{1,...,M\}$, where $s(x) = \sum_{k=1}^{K}\sum_{m=1}^{M} g_{m,k}(x)$, $p_{m,k}(x) = s(x) - g_{m,k}(x)$, and $q(x) = \max_{k\in\{1,...,K\},m\in\{1,...,M\}} p_{m,k}(x)$. Then constraint (3b) is equivalent to the following two constraints with the auxiliary variable $t \in \mathbb{R}$:

$$q(x) - t \leq 0,$$
$$s(x) - t \geq 0,$$

where the first constraint is a convex constraint and the second constraint is a reverse convex constraint. Thus, if (3) is a convex program with multiple additional reverse convex constraints, then (4) is a convex program with a single reverse convex constraint.

$$\min \sum_{p\in P}\sum_{m\in\mathcal{M}} \frac{f(x)}{|\mathcal{M}|} \tag{4a}$$
$$\text{s.t.} x_m \in F, \forall\ m \in \mathcal{M}, \tag{4b}$$
$$s(x) - t \leq 0, \tag{4c}$$
$$q(x) - t \geq 0, \tag{4d}$$
$$x \in \mathcal{P}, \forall\ m, m' \in \mathcal{M}, m \neq m'. \tag{4e}$$

Returning to the K-means structure, the z variables indicate whether a sensor is assigned to a location. These variables are binary, but these binary constraints can be expressed with reverse convexity:

$$\{z \in \mathbb{B}^{MK}\} = \{z \in \mathbb{R}_+^{MK} \mid h(z) \geq 0\},$$
$$h(z) = \sum_{m=1}^{M}\sum_{k=1}^{K} z_{mk}(z_{mk} - 1).$$

Thus, one can adapt s and q to include h as one of the functions along with g_{mk} to create a reverse convex program without integrality constraints.

3 Some Implementation Notes

The workflow for solving the problem looks like:

1. Read in image of the field site
2. Using colors determine which areas are off limits
3. Use ellipses to cover the off-limit areas for the reverse convex model
4. Solve the sensor placement problem
5. Using some simulation, determine the collected "data" by the algorithmically placed sensors
6. Create model for entire field
7. Compare to simulated truth