A Hierarchical Bayesian Model for Estimating Methane Fields from TROPOMI Observations

William Daniels

1 Introduction

Natural gas is often referred to as a "bridge fuel" between other fossil fuels and renewable energy. At the point of combustion, natural gas (primarily made up of methane) produces less carbon than both coal and oil [1]. If released into the atmosphere, methane decays much faster than carbon dioxide [2]. Both of these points make methane a promising alternative to both coal and oil. However, methane has a global warming potential (GWP) of 84 over 20 years, meaning that while it does not remain in the atmosphere as long as carbon dioxide, its affect on the climate is much more potent [2]. Therefore, if methane is to be considered a cleaner alternative to other fossil fuels, its production, transportation, and storage must be done in such a way as to limit both fugitive emissions (i.e. leaks) and operational emissions (i.e. venting). To ensure that this is being done, effective emissions monitoring is required.

Emissions monitoring can be performed using an array of platforms: ground based "fenceline" sensors, drones and aircraft, and satellite. Here we focus on satellite based monitoring, specifically methane observations from the TROPOMI instrument on board the Sentinel-5 satellite [3]. These data have great potential for emissions monitoring due to their global coverage, but they are limited by their coarse temporal and spatial resolution. TROPOMI has a spatial resolution of 7×5.5 km, meaning that the smallest area it can observe is a 7×5.5 km "pixel." This makes attributing an elevated methane observation to a particular source extremely difficult. **Here we attempt to estimate (with uncertainty) the underlying methane field using these TROPOMI pixels.** This will provide a clearer picture of methane concentrations on a small scale and potentially help pinpoint emission sources.

2 TROPOMI Data Description

TROPOMI observes each geographical point approximately once per day with a spatial resolution of 7×5.5 km (as of August 2019 - earlier observations had a resolution of 7×7 km). The methane observations are packaged along with a number of other fields, including the time of measurement and the geographical footprint of each "pixel." Each methane observation can be considered an average of the continuous methane field within the corresponding pixel. Note that these pixels do not have a fixed shape, rather they depend on their location relative to the nadir of the satellite. Pixels far from nadir are elongated due to their projection onto the Earth's surface, and pixels close to nadir appear more rectangular [3].

In this report, we focus on a single TROPOMI overpass that occurred on September 13, 2019. We further focus on a region centered in northeast Colorado, which spans about a degree in longitude and a degree and a half in latitude. Figure 2(a) shows the TROPOMI methane observations from this overpass. The TROPOMI methane data used in this report are available from the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC).

3 Statistical Methods

The problem of estimating the methane field given the TROPOMI observations is an inverse problem. The quantity we observe (TROPOMI observations, z) is a function of a quantity we do not observe directly (methane field, c). We could write this as $z = \mathcal{F}(c)$. Ultimately, however, we are interested in computing the methane field as a function of the observed data, or $c = \mathcal{F}^{-1}(z)$. To solve this inverse problem, we have created a hierarchical Bayesian model.

3.1 Bayesian Hierarchical Model

The TROPOMI observations, z, make up the highest level of our model. These observations are a function of the underlying methane field, c, which makes up the second level. We model c as a fixed term plus a spatial process, both of which are governed by a set of parameters that, along with a parameter controlling the measurement error in z, make up the lowest level of our model.

The data level of our model is written as

$$z = Wc + \epsilon,$$

where W is a matrix that averages the methane field, c, to produce each entry in z and $\epsilon \sim \text{N}(0, \tau^2)$ is a Gaussian white noise error term. The W matrix is simply a matrix of weights. Each row of W applies a weight to every entry in c, such that Wc produces the values of the TROPOMI observations. This is achieved by giving zero weight to c outside of the pixel for each observation in z. The W matrix is considered known and fixed, although improvements to this assumption are discussed in Section 4.

The second level of our model is written as

$$\begin{aligned} c &= \beta_1 + \beta_2 \text{lat} + \beta_3 \text{lon} + y \\ &= X\beta + y, \end{aligned}$$

where X is a matrix of covariates, β is the vector of coefficients, and $y \sim \text{N}(0,\ \sigma^2 K(\theta))$ is a Gaussian process. In this model, we only consider two covariates, latitude and longitude. This creates a fixed plane that varies with space. We further assume that $\text{cov}(y, \epsilon) = 0$, or that the spatial process, y, is independent of the measurement error governed by τ^2. Finally, we assume an exponential covariance function, namely

$$K(\theta) = \exp\left(-\frac{||s_i - s_j||}{\theta}\right),$$

where $||s_i - s_j||$ is the great circle distance between the two points s_i and s_j. Note that our choice of covariance function is fairly arbitrary at this point. Future work will implement better methods of selecting K, which we discuss further in Section 4.

The third level of our model contains the parameters that govern the data and process levels. We have τ^2, the variance of the measurement error, σ^2, the process variance, θ, the range parameter, and β, the coefficients of the fixed part of the process level model. Instead of a fully Bayesian approach, in which we would create prior distributions for each of these parameters, we instead use an empirical Bayesian approach, in which we will estimate these parameters via maximum likelihood.

Bringing together the three levels and the various assumptions discussed in this section, we get that

$$z = WX\beta + Wy + \epsilon,$$

which is normally distributed according to

$$z \sim \text{N}(WX\beta,\ \sigma^2 WKW^T + \tau^2 I). \tag{1}$$

3.2 Estimating Model Parameters

Here we discuss the maximum likelihood approach used to estimate the covariance parameters $\omega = (\sigma^2, \tau^2, \theta)$ and the fixed coefficients β. Using Equation 1, we can write out the PDF for z given ω and β, and hence the associated log likelihood, $\mathcal{L}(\omega, \beta|z)$. To maximize $\mathcal{L}$ over ω and β, we create a profiled $\mathcal{L}$ that relies solely on θ and $\lambda = \tau^2/\sigma^2$. This is done by first plugging in the GLS estimate of β, given by

$$\hat{\beta}_{GLS} = (A^T \Sigma^{-1} A)^{-1} A^T \Sigma^{-1} z,$$

where in our case $A = WX$. This leaves $\mathcal{L}$ in terms of just ω. We then maximize $\mathcal{L}$ analytically for σ^2 by setting the derivative equal to zero and solving for σ^2. This leaves us with

$$\mathcal{L}(\theta, \lambda|z) = -\frac{m}{2}\ln(2\pi) - \frac{m}{2}\ln(\hat{\sigma}^2) - \frac{1}{2}\ln|WKW^T + \lambda I| - \frac{m}{2}, \tag{2}$$

where m is the number of TROPOMI observations. We can then maximize Equation 2 numerically over θ and λ. The surface resulting from this numerical optimization is shown in Figure 1.

It is important to note, however, that there is additional subtlety to this process, which is required to get a more stable estimate of θ and λ. Since each TROPOMI overpass contains only a few hundred observations, we use multiple days of data to get the MLEs. This requires us to assume that each day is independent, which is a fair but ultimately inaccurate assumption. Future work might take into account the correlation between subsequent TROPOMI overpasses, discussed further in Section 4. When combining multiple days of data, we must also create a single estimate of β, otherwise the details of the spatial process get absorbed into a separate regression for each day. To do this, we break up the likelihood calculation into two steps. The first, described above, gives a separate estimate of β for each day. We then average these estimates to get a single value for $\hat{\beta}$. The second step repeats some calculations using this averaged $\hat{\beta}$ to get the maximum likelihood estimates for θ and λ.

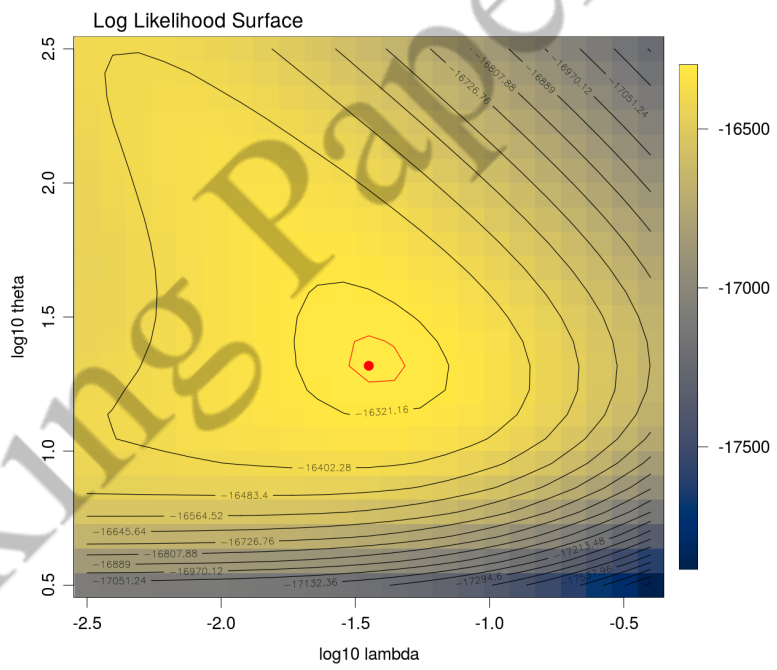

Figure 1: Log likelihood surface used to estimate λ and θ. The MLEs are shown as a red dot and a 95% confidence level is drawn in red.

3.3 Predictions and Uncertainty

With maximum likelihood estimates for θ and λ in hand, we compute estimates for all of the parameters in our hierarchical model. We can then create predictions for the spatial process, y, using properties from the conditional normal distribution. We get that

$$\hat{y} = \text{cov}(y, z)\,\text{cov}(z, z)^{-1}\,(z - WX\hat{\beta}).$$

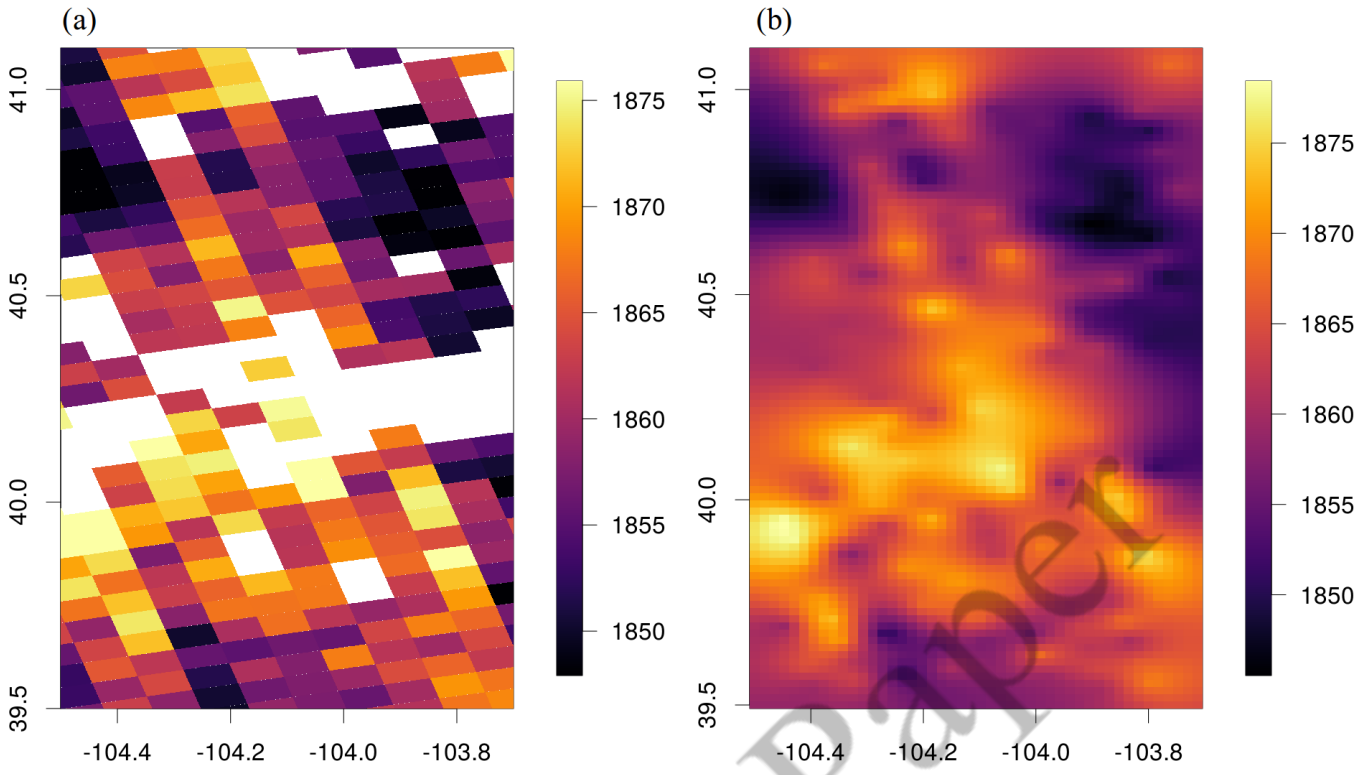

Figure 2: (a) TROPOMI methane observations over northeast Colorado. Note the pixelated structure. (b) Estimated methane field given the TROPOMI observations.

It is straightforward to then estimate the methane field, which is given by

$$\hat{c} = X\hat{\beta} + \hat{y}.$$

Figure 2 shows the TROPOMI pixels and our estimated methane field for a single overpass. However, we are interested not only in predicting the methane field, but also in quantifying the uncertainty in our prediction. To do this, we use conditional simulation to create an ensemble of equally likely methane fields given the TROPOMI observations, z. A general outline of this algorithm is given below:

- Using the MLEs $\hat{\theta}$ and $\hat{\lambda}$, we can generate a synthetic spatial process, y^*, by computing

  $$y^* = L^T u,$$

  where L is the Cholesky decomposition of $\text{cov}(s_{\text{pred}})$, s_{pred} are the locations on which we predict the methane field, and $u \sim \text{N}(0, 1)$. This result follows from the definition of the Cholesky decomposition and the affine transformation properties of the multivariate normal distribution.

- We can then generate synthetic TROPOMI observations, z^*, based on y^* and the W matrix for each day.

- Using z^* and the equations discussed earlier in this section, we can then compute predictions, $\hat{y}^*$, of the synthetic field, y^*.

- Since we created y^*, we know it exactly. This allows us to compute the error of our prediction, $e = \hat{y}^* - y^*$.

- Finally, we can repeat this process for M many synthetic fields. Each e is an equally likely draw from $y - \hat{y}$, making each $\hat{y} + e$ an equally likely spatial process. The collection of M equally likely processes is called an ensemble.

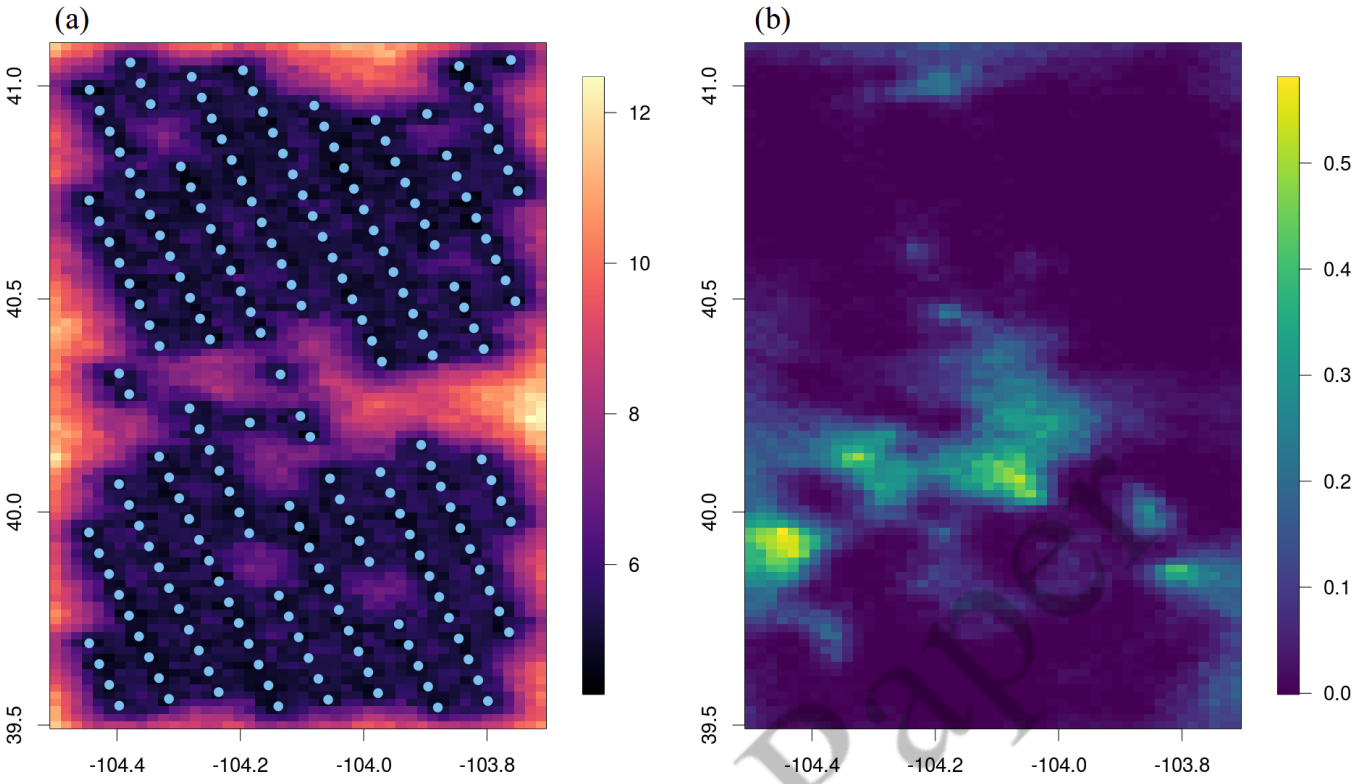

Figure 3: (a) Standard errors from an ensemble with $M = 200$. (b) Estimated probability of each prediction location containing a methane concentration in the top 5% of all predictions within the ensemble.

With this ensemble of equally likely spatial processes, we can compute the standard error of the M fields at each prediction location. This shows how the uncertainty in our estimate varies over space. We plot the standard errors for the TROPOMI overpass from Figure 2 in Figure 3(a). We save the discussion of Figures 3(a) and 3(b) for Section 4.

4 Results, Conclusions, and Future Work

Figure 3(a) shows the standard error of the predictions plotted in Figure 2(b). The center of each TROPOMI pixel is plotted as a blue dot. The standard error clearly decreases as you move closer to one of the observations and increases as you move away. This makes intuitive sense, as the predicted methane field is less constrained as you move away from the actual TROPOMI observations. Figure 3(a) highlights an overpass with relatively dense observations. More frequently, however, TROPOMI observations are quite sparse due to cloud cover or other sources of albedo. These standard error plots will be especially useful in scenarios with less data.

In addition to giving us an estimate of prediction uncertainty, the ensemble of equally likely methane fields allows us to perform any number of interesting inferences. Figure 3(b) highlights one example. Here we plot the estimated probability of a prediction location containing a methane concentration in the top 5% of all predictions within the ensemble. These probabilities clearly highlight three or four locations that are most likely to contain the highest methane concentrations in the region for this given overpass. Further analysis could be performed to see if these locations align with any natural or anthropogenic sources of methane (i.e. wetlands or O&G facilities). This type of inference is of particular interest, as a large percent of anthropogenic methane emissions comes from a small percent of emission events [4]. These so called "super emitter events" therefore provide the greatest opportunity for reducing emissions.

There are many ways in which we could enhance the predictions discussed in this document, some being hinted at throughout the text. Below is a short list of future work that could be performed related to this modeling framework.

- Use cross-validation to select the covariance function, as opposed to arbitrarily selecting an exponential form. A smoother function, such as a Matérn, might be better suited to these data.
- Incorporate the correlation between subsequent TROPOMI overpasses when estimating $\hat{\beta}$. Currently we assume subsequent days are independent, which is likely a fair assumption at best.
- "Sharpen" the W matrix by including fractional weights when a prediction location is partly inside of a given TROPOMI pixel.

To summarize, we have used a hierarchical Bayesian framework to estimate the underlying methane field from a set of TROPOMI observations. This framework takes into account the size and shape of the TROPOMI pixels, rather than simply treating the observations as point concentrations. We have quantified the uncertainty in our estimate using conditional simulation, as well as highlighted an interested inference that can be performed with the resulting ensemble of equally likely methane fields.

References

[1] "How much carbon dioxide is produced when different fuels are burned? | U.S. Energy Information Administration (EIA)," https://www.eia.gov/tools/faqs/faq.php?id=73&t=11, (Accessed on 12/16/2020).

[2] "Understanding Global Warming Potentials | Greenhouse Gas (GHG) Emissions | US EPA," https://www.epa.gov/ghgemissions/understanding-global-warming-potentials, (Accessed on 12/16/2020).

[3] "Methane | Tropomi," http://www.tropomi.eu/data-products/methane, (Accessed on 12/16/2020).

[4] D. Zavala-Araiza, R. A. Alvarez, D. R. Lyon, D. T. Allen, A. J. Marchese, D. J. Zimmerle, and S. P. Hamburg, "Super-emitters in natural gas infrastructure are caused by abnormal process conditions," *Nature Communications*, vol. 8, no. 1, p. 14012, Jan 2017. [Online]. Available: https://doi.org/10.1038/ncomms14012