# Bayesian Methods for Enhanced Greenhouse Gas Emissions Inventories

Michael Pekala[1,3], Zoheyr Doctor[2,3], Elizabeth Reilly[1,3], Gavin McCormick[2,3], Gary Collins[1,3], Daniel Moore[2,3], Amy Piscopo[2,3], Krsna Raniga[2,3], Michael Robinette[1,3], Ting So[2,3]

[1] Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland, USA, [2] WattTime, Oakland, USA, [3] Climate TRACE

## Overview

**Key question: What emissions properties can we estimate for facilities we don't directly observe?**

Modern emissions inventories can provide fine-grained sources of GHG emissions, some even down to the facility level [CT25]. However, in settings where facilities are not known a priori but must be detected (e.g., via remote sensing), facility location data may be incomplete. How might priors and auxiliary data of varying resolution help enrich these inventories?

### Selection Bias

For known facilities, inventories provide estimates for one or more relevant characteristics (e.g., activity level):

$$\{d_i\},\ i = 1, \ldots, N_{\text{obs}}$$
$$d_i = \theta_i + \sigma_i$$

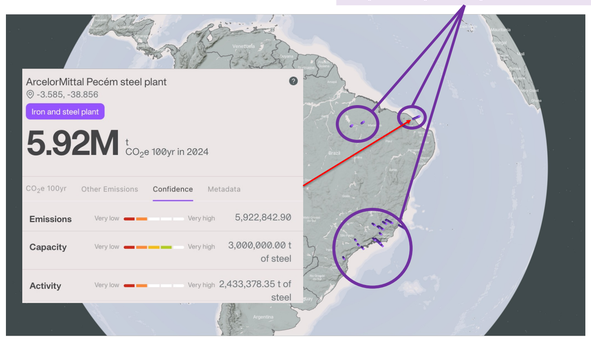


**Fig. 1: Example facility-level data from Climate TRACE for the iron and steel sector.**

In settings where facilities are not known but can be sensed, the probability of observing a facility may be related to the magnitude of its characteristic. For example, a very active steel blast furnace may produce a more persistent heat signature than a less active one.

However, we may have (a) good priors about population-level distributions of these characteristics, and (b) reliable auxiliary information about these characteristics, such as an aggregate value:

$$d_{\text{net}} = \sum_{i=1}^{N_{\text{obs}}} \theta_i + \sum_{j=1}^{N_{\text{unobs}}} \theta_j + \sigma_{\text{net}} \qquad (1)$$

How can we leverage these priors and auxiliary data to obtain useful information about the overall population?

## Approach

We propose that Bayesian methods offer a principled framework for integrating heterogeneous emissions and metadata across multiple scales and can help infer properties of unobserved data.

We adopt the framework of [Man19], which accounts for selection effects as well as measurement uncertainty. In this work we augment that framework with additional terms to accommodate the aggregate observations.

## Bayesian Model

Let:

| | | |
|---|---|---|
| $\lambda$ | Parameters of the latent facility population | Posterior estimates of interest |
| $N$ | Total number of facilities | |
| $d_i$ | Noisy observations of the $i^{\text{th}}$ facility's characteristics | Observed data |
| $N_{\text{obs}}$ | The number of detected facilities | |
| $d_{\text{net}}$ | Aggregate characteristic observation | |
| $N_{\text{det}}$ | The expected number of detected facilities | Nuisance or assumed |
| $\theta$ | Facility characteristics | |
| $\pi(.)$ | Prior distribution for the associated parameter | |

[Man19] derive the following posterior estimate for the population parameters and total population count:

$$p(\lambda, N|\{d_i\}) = \pi(\lambda)\pi(N) \prod_{i=1}^{N_{\text{obs}}} \frac{\int p(d_i|\theta) p_{\text{pop}}(\theta|\lambda) d\theta}{\int p_{\text{det}}(\theta) p_{\text{pop}}(\theta|\lambda) d\theta} \cdot e^{-N_{\text{det}}} (N_{\text{det}})^{N_{\text{obs}}} \qquad (2)$$

We expand this model by introducing a term that relates the above model to our aggregate activity data:

$$d_{\text{net}} \sim \mathcal{N}(N \cdot \mathbb{E}[p_{\text{pop}}(\theta|\lambda)], \sigma_{\text{net}}) \qquad (3)$$

We implement this model using the NumPyro [Pha19] probabilistic programming framework:

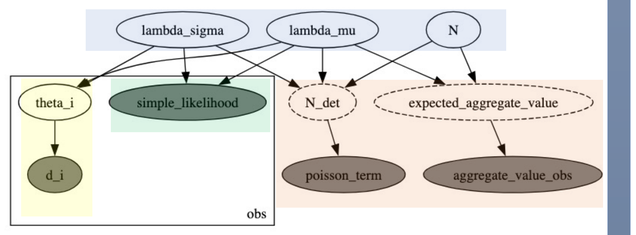


**Fig. 2: Graphical model corresponding to Eq. (2) and (3).**

## Motivating Example

Consider the scenario depicted in Fig. 1. Based on the observed activity data from [CT25], together with country-level (i.e., aggregate) economic estimates of steel activity [WAS24], how many unobserved facilities might exist and approximately how active might we anticipate them to be?

Answers to this question, applied across multiple sectors and geographic locations, can help inform future data collection strategies or triage intervention opportunities.

## Preliminary Exploration

**Assumptions:**

- We are modeling BF/BOF facilities only
- Facility activity levels can be modeled as independent draws from a (latent) normal population distribution
- Our economic data is for steel facilities only; we currently assume this is a good approximation for the entire iron and steel sector
- Model components (e.g., priors, noise terms) have not yet been tuned

### Real-world and Synthetic Data

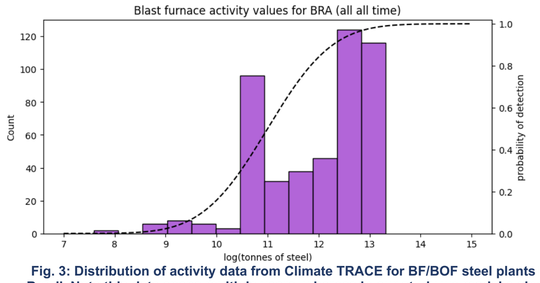


**Fig. 3: Distribution of activity data from Climate TRACE for BF/BOF steel plants in Brazil. Note this data spans multiple years, whereas in our study we model a single month of activity (2022-01-01). A presumed probability of detection is shown on the second y axis (black dashed line).**

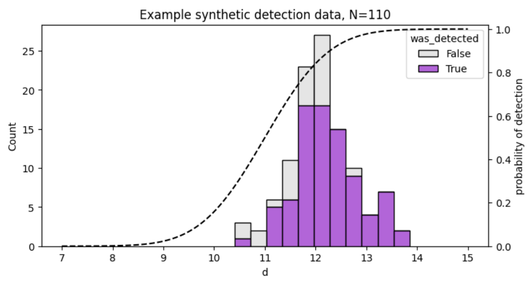


**Fig. 4: Synthetic activity data generated for the case where N=110 total facilities. Observed facilities are shown in purple; missing (unobserved) facilities are in gray.**

### Posterior Estimates

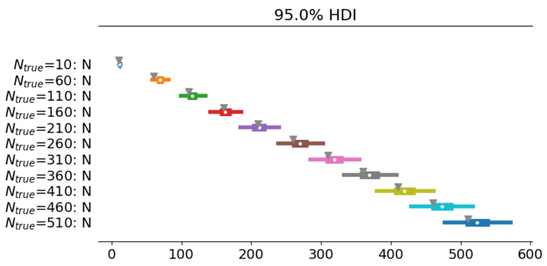


**Fig. 5: 95% highest density interval (HDI) posterior estimates on synthetic data as we sweep over various values of N (for the same population parameters). Note that the posterior estimates cover the true value (gray triangles).**

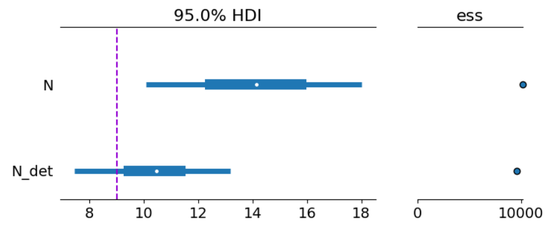


**Fig. 6: Posterior estimates for N (total population size) and $N_{\text{det}}$ (expected number of detections) for real-world data. Purple dashed line shows true number of observations. These posterior estimates are combined across 4 Markov Chain Monte Carlo (MCMC) chains; ess denotes the effective sample size estimate.**

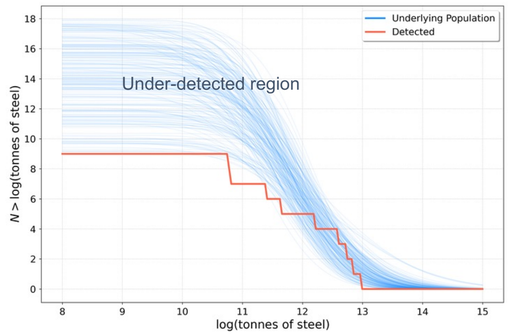


**Fig. 7: Cumulative number of facilities with activity > *x* tonnes of steel for real-world data. The orange line depicts observed samples while blue denotes realizations drawn from the posterior estimate of the underlying population.**

## Conclusions & Next Steps

Early indications suggest the model does a good job capturing parameters of interest in synthetic data (Fig. 5). Furthermore, the distributional assessments (Fig. 7) provides insight into how active these facilities might be.

- Enriched modeling, including:
  - explore other distributions for modeling facility populations (e.g., power law, uniform); run sensitivity studies and perform model selection,
  - temporal extensions (i.e., move beyond a single month), and
  - tuning detection models and noise priors.
- Generalizing to other sectors and countries.
- Cross-validation & additional studies using real-world data.

## References


- [CT25] Climate TRACE website, https://www.climatetrace.org
- [Man19] Mandel, Ilya, Will M. Farr, and Jonathan R. Gair. "Extracting distribution parameters from multiple uncertain observations with selection biases." Monthly Notices of the Royal Astronomical Society 486.1 (2019): 1086-1093.
- [Pha19] Phan, Du, Neeraj Pradhan, and Martin Jankowiak. "Composable effects for flexible and accelerated probabilistic programming in NumPyro." arXiv preprint arXiv:1912.11554(2019).
- [WSA24] World Steel Association. World steel in figures. https://worldsteel.org/data/world-steel-in-figures/world-steel-in-figures-2024/, 2024.


## Acknowledgements and Contact

The authors would like to thank the Climate TRACE coalition for its organizational support, and Climate TRACE's funders for their financial support. For a full list of funders, please visit the Climate TRACE website: https://climatetrace.org/team.

**Contact:** michael.pekala@jhuapl.edu, elizabeth.reilly@jhuapl.edu, or zoheyr.doctor@watttime.org



2025.12.07