Guide on Data

Revision 1, February 6, 2022

General Notes:

- These were generated large image datasets from methane plume enhancements captured from airborne measurement surveys taken between 2019 and 2022.
- There are roughly **6800** individual plumes available
- They cover a large area in the US and are concentrated in California and the Permian Basin. A "dummy" dataset without plumes was also created., i.e. the "no-cat" dataset (9750 images). These just showcase images without the methane plumes.
- The image sizes provided have been kept small (217x217 dim, 72x72 resolution) in case some ambitious participants choose to train a deep learning model
- The original geotiff files are available. The originals are rasters (.tif files). Each plume has one rgb tiff for the background image and one raster file for the plume. The png's were generated by combining both.
- Each plume is associated with a "source" and "plume" data entry in associated spreadsheets
    - The source data entries contain information about the identified source of the emissions (e.g., location, type of facility, rate, ipcc...)
    - The plume data entries contain information about the plume
- There are a few ways to aggregate the info together/visualize. Some examples have been provided to show this visualization.

Permian Specific

Note that the Permian specific datasets are also contained within the general CarbonMapper list below, so you may find some duplicates, but that should be easily evident.

Satellite Data Set: https://zenodo.org/record/5610307

Plume List Data: https://pubs.acs.org/doi/10.1021/acs.estlett.1c00173

Data Sets

- permian_plume_list_2019.csv
    - Source ID: Unique source (e.g. facility) identifier for a specific plume.
    - plume_lat: Latitude of emission source
    - plume_lon: Longitude of emission source
    - candidate_id: Unique identifier for each plume. The first three characters (i.e., "GAO" or "ang") represent the instrument platform: GAO=Global Airborne Observatory; ang=Next Generation Airborne Visible Infrared Imaging Spectrometer (AVIRIS-NG). The next characters represent the timestamp of acquisition. The suffix (e.g., "-A") uniquely identifies plumes that may have been quantified in the same flight line
    - date: Date of acquisition
    - time: time of the observation
    - qplume: Raw, unadjusted methane plume rate in kg/hr
    - sigma_qplume: Raw, unadjusted uncertainty of methane plume rate in kg/hr
    - active_flair_detected: Note this is spelled wrong. It should say “flare”. This is true/false if an active flare is detected on site.

    - inactive_flair_detected: Note this is spelled wrong. It should say “flare”. This is true/false if an inactive flare is detected on site.
- permian_source_list_2019.csv
    - source_id: Unique source (e.g. facility) identifier for a specific plume.
    - source_lat: Latitude of emission source
    - source_lon: Longitude of emission source
    - type: See table below for categories and their definitions:

category	**what pertains to this category**
tank	a source located near* a tank at an O&G upstream production site
well	a source not located at a tank* at an O&G upstream production site (e.g., pumpjack, flare, etc)
compressor	any emission at a compressor station (tank, flare, compressors, etc). If just one small compressor, don't put in this category
pipeline	any emission at a pipeline or gathering line
other oil and gas	use very sparingly - but represents other midstream elements, like gas storage
processing	any emission at a processing plant (tank, flare, compressor, etc)
compost	windrows not located on a larger landfill face
landfill	landfill
coal mine	emission from the face of a coal mine
coal vent	emission from a coal vent not necessarily at the site of extraction
dairy digestor	rubber over a lagoon
livestock	everything not a digestor (manure lagoon, etc)
powerplant	plant when gas is made into electricity
wastewater	waste water treatment plant. Usually digestors
refinery	refinery or petrochemical plants

    - ipcc_sector: the IPCC sector the source/plume falls into
    - number_overflights: number of types the source was observed
    - source_persistence: frequency of the observation (number of detected plumes divided by number of observations)
    - confidence_in_persistence: Persistence weighting for each CH_4 (methane) source detected in the survey. This is estimated by accounting for the frequency ($f = M/N$) in which a plume was detected, where M is the number of detected plumes at that source location during the campaign and N is the number of overflights over that location during the campaign.
    - qsource: Persistence adjusted average source emissions (units in kg/hr)
    - sigma_qsource: Uncertainty for persistence adjusted average source emissions (units in kg/hr)

All Airbourne Plume Data

Satellite and Plume List Data Set (July 2020 to August 2021): https://zenodo.org/record/5606120

Satellite and Plume List Data Set (July 2020 to May 2022): https://zenodo.org/record/7072824

Data Sets

- plume_attribution_2019-2022.csv
    - source_id: Unique source (e.g. facility) identifier for a specific plume.
    - candidate_id: Unique identifier for each plume. The first three characters (i.e., "GAO" or "ang") represent the instrument platform: GAO=Global Airborne Observatory; ang=Next Generation Airborne Visible Infrared Imaging Spectrometer (AVIRIS-NG). The next characters represent the timestamp of acquisition. The suffix (e.g., "-A") uniquely identifies plumes that may have been quantified in the same flight line
    - plume_lat: Latitude of emission source
    - plume_lon: Longitude of emission source
    - date: Date of acquisition
    - source_type: See table from Permian basin specific for source types
    - ipcc: the IPCC sector the source/plume falls into
    - qplume: Quantified emission rate of plume, estimated using the Integrated Methane Enhancement method (see Duren et al., 2019 - "California's Methane Super-Emitters", Nature) (units in kg/hr)
    - sigma_qplume: Uncertainty in emission rate, derived from uncertainty in IME and wind speed. Wind speeds used in this analysis are from the HRRR meteorological reanalysis (NOAA) (units in kg/hr)
    - time_of_detection: time of the observation
    - plume_tif: Can be ignored
    - rgb_tif: Can be ignored
- source_list_2019-2022.csv
    - source_id: Unique source (e.g. facility) identifier for a specific plume.
    - source_lat: Latitude of emission source
    - source_lon: Longitude of emission source
    - number_overflights: number of types the source was observed
    - source_persistence: frequency of the observation (number of detected plumes divided by number of observations)
    - qsource: Quantified emission rate of the source, the mean of all emissions attributed to that source weighted by the persistence. (units in kg/hr)
    - sigma_qsource: Uncertainty in emission rate, derived from uncertainty in IME and wind speed. The uncertainty combines all plume uncertainty attributed to the source weighted by persistence (units in kg/hr)
    - source_type: See table below for categories and their definitions:

category	**what pertains to this category**
tank	a source located near* a tank at an O&G upstream production site
well	a source not located at a tank* at an O&G upstream production site (e.g., pumpjack, flare, etc)
compressor	any emission at a compressor station (tank, flare, compressors, etc). If just one small compressor, don't put in this category
pipeline	any emission at a pipeline or gathering line

other oil and gas	use very sparingly - but represents other midstream elements, like gas storage
processing	any emission at a processing plant (tank, flare, compressor, etc)
compost	windrows not located on a larger landfill face
landfill	landfill
coal mine	emission from the face of a coal mine
coal vent	emission from a coal vent not necessarily at the site of extraction
dairy digestor	rubber over a lagoon
livestock	everything not a digestor (manure lagoon, etc)
powerplant	plant when gas is made into electricity
wastewater	waste water treatment plant. Usually digestors
refinery	refinery or petrochemical plants

- confidence_in_persistence: indicates how confident the analyst was in attribution

Dummy Data

This dataset contains imagery without any plumes for users to explore and see if they can come up with a way to detect plumes and draw a box around the source or the plume itself:

- dummy_data.zip: Specific to the general all airbourne data
- dummy_data_permian.zip: Specific to the Permian basin