JPL Document D-101028

IS-GEO Dataset JPL-CH4-detection-2017-V1.0: A Benchmark for Methane Source Detection from Imaging Spectrometer Data

David R. Thompson[1]
Anuj Karpatne[2]
Imme Ebert-Uphoff[3]
Christian Frankenberg[4]
Andrew K. Thorpe[1]
Brian D. Bue[1]
Robert O. Green[1]

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109; [2]Computer Science & Engineering, University of Minnesota, Minneapolis, MN, karpa009@umn.edu; [3]Electrical & Computer Engineering, Colorado State University, Fort Collins, CO, iebert@colostate.edu; [4]Division of Geology and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

October 2017

Abstract

Methane (CH_4) is a powerful greenhouse gas. CH_4 enters the atmosphere from various sources: naturally, from phenomena like geologic seeps; and artificially, through a range of processes like animal husbandry, decomposition in landfills, and oil and gas extraction and production. Institutions like the Jet Propulsion Laboratory are developing automated methods for the reliable detection, and potentially classification, of CH_4 sources from imaging spectrometer data. The key challenge is to distinguish CH_4 sources from background noise. Domain experts currently perform this task manually by visual inspection of the imaging spectrometer data. In consort with the NSF-sponsored Research Coordination Network on Intelligent Systems Research to Support Geosciences (IS-GEO, https://is-geo.org/), we have prepared a benchmark dataset to promote future research on automated detection. The dataset, designated JPL-CH4-detection-2017-V1.0, contains sample imaging spectrometer data acquired in the "Four Corners" area - a hotspot of CH_4. This document describes the dataset to researchers developing and testing algorithms for CH_4 detection from imaging spectrometer data. The benchmark includes, (1) this description and (2) the dataset. The data was provided by the Jet Propulsion Laboratory using the Next Generation Airborne Visible Infrared Imaging Spectrometer (AVIRIS-NG, https://avirisng.jpl.nasa.gov/). The benchmark dataset is available online for use by researchers.

1. Introduction

Atmospheric Methane (CH_4) is an important Greenhouse Gas whose potential impact on global climate is many times higher than an equivalent quantity of CO_2. It enters the atmosphere from various sources: naturally, from phenomena like geologic seeps; and artificially, through a range of processes like animal husbandry, decomposition in landfills, and oil and gas extraction and production. The lifespan of CH_4 in the atmosphere is far shorter than CO_2, so reducing anthropogenic CH_4 could be an efficient path to mitigate humanity's climate impact. However, many anthropogenic sources of CH_4 remain poorly understood. Spurious releases and leaks of natural gas from oil and gas infrastructure, including extraction and storage facilities, play some role but often go undetected. These *fugitive releases* result in excess greenhouse gases and loss of revenue for the extraction industry. Recently, airborne imaging spectrometers have been increasingly used to detect, map, and characterize these emissions over wide areas. These instruments are a potent new tool to understand fugitive CH_4 emissions.

This document describes a dataset, named JPL-CH4-detection-2017-V1.0, which comes from NASA's Next Generation Visible Infrared Imaging Spectrometer (AVIRIS-NG). AVIRIS-NG looks down from aircraft at altitudes from a few hundreds to thousands of meters. AVIRIS-NG was not originally designed to detect CH_4, but its spectral range encompasses diagnostic features of this trace gas and it has proven useful for this application (Thorpe et al., 2016). This dataset contains examples of intermediate products from this sensor, acquired during a campaign to seek fugitive emissions in the Four Corners area of the United States. As a result of the Four Corners campaign, over 250 distinct sources of CH_4 have been identified, leading to a significant improvement in our understanding of fugitive CH_4 sources (Frankenberg et al., 2016). This dataset offers an opportunity to develop and benchmark new algorithms for detection and mapping of CH_4. Advances in computing and algorithm technology can significantly improve the value of this and future surveys.

This benchmark is an effort of the NSF-sponsored RCN on Intelligent Systems Research to Support Geosciences *(IS-GEO)*, which is engaged in creating several benchmarks representative of different geoscience tasks to attract more data scientists to collaborate on these tasks. See Ebert-Uphoff et al. (2017) for information on this effort.

1.1 Background

Observations from space indicated a hotspot of CH_4 emissions in the Four Corners region of the US, which is located near the New Mexico-Colorado border (Kort et al. 2014). The area of interest covers about 80 x 40 km^2. The resolution of current space observations is currently too low to identify the actual sources of CH_4 emissions, so an airborne campaign was conducted in April 2015 with multiple instruments including NASA's Next Generation Visible Infrared Imaging Spectrometer (AVIRIS-NG) (Frankenberg et al. 2016). AVIRIS-NG measured the area of interest in parallel lines, called flightlines, at an altitude of about 3 km above ground. Each flightline yielded infrared measurements for one narrow section of the area. This benchmark dataset corresponds to a portion of the Four Corners survey, split into flightlines representing individual contiguous image strips acquired by the spectrometer as it flew over the terrain.

The campaign aimed to characterize the distribution of different CH_4 sources in the Four Corners area, identifying leaks in oil and gas infrastructure as well as natural sources. Analysis of the data revealed more than 250 CH_4 plumes (Frankenberg et al. 2016), some of which which were detected in real time and immediately investigated by teams on the ground. Sources included a leak in a pipeline, natural seeps, a leaking underground storage tank, and a coal mine venting shaft. Post-analysis determined that just 10% of the sources were responsible for up to 60% of CH_4 emissions – indicating that that remedial action at a small number of sites could gain large overall reduction in atmospheric greenhouse forcing (Frankenberg et al. 2016).

1.2 Future Perspective

Further insights into the nature of CH_4 sources and their emission rates are important for policy and resource allocation decisions. At the time of this writing, additional measurements are underway in the state of California (see for example https://www.arb.ca.gov/research/methane/ab1496_research.htm). Future orbital instruments could augment the airborne picture with global observations of large sources (Thompson et al. 2016). The increasing volume of data and coverage areas makes *automatic* detection from spectrometer data even more important.

2. Data Acquisition and Format

2.1 Instrumentation

AVIRIS-NG acquires data in the Visible-Shortwave Infrared (VSWIR) range, measuring spectra comprised of over 400 color channels reaching far beyond the range of light visible to human eyes, and ranging from ultraviolet (380 nm) through Shortwave Infrared (2500 nm). Images are taken over large areas, creating a three-dimensional data cube of radiance (roughly, a measure of light intensity), where two dimensions are spatial and the third is wavelength, i.e. for each wavelength the data cube contains one 2D image (Figure 1).

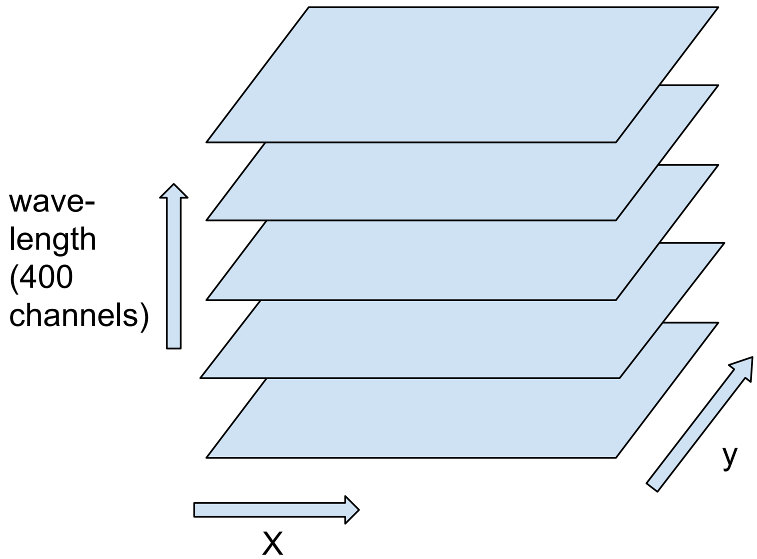

Fig. 1: Radiance data cube with two spatial dimensions and one wavelength dimension. Each plane contains radiance values for one wavelength (one channel).

Patterns and signatures in these spectra are sensitive to physical and chemical properties of the atmosphere and surface, enabling quantitative measurement of surface and atmospheric properties throughout the scene. In particular, atmospheric CH_4 absorbs light, leaving a characteristic "fingerprint" in several infrared regions.

2.2 Analysis Chain

Typical analyses use a two-stage process for CH_4 source detection (Figure 2). This benchmark dataset focuses on the second step.

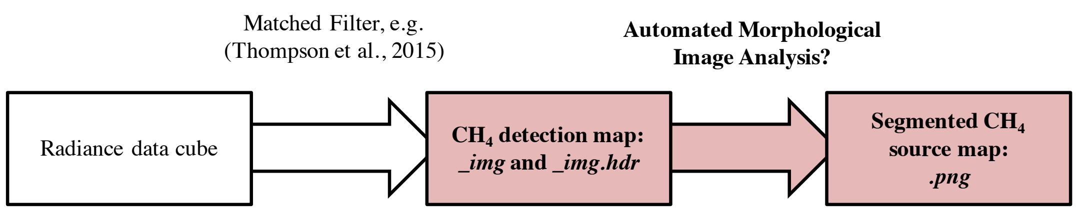

Figure 2: Complete CH_4 analysis chain. Step 1 (left arrow) maps the radiance data cube to a CH_4 estimate. This is done automatically and yields **img** and **img.hdr** files (RGB image and methane estimate). Step 2 (right arrow) is the task proposed here. It takes the img and img.hdr files and detects actual methane sources. Step 2 is currently performed manually and the best approach to automatically segment the image into real plumes (sources) is not yet known.

Step 1 (Preprocessing of data): The AVIRIS-NG science data system uses the known spectral signature of methane light absorption to convert the 400-channel data at each grid point to a single estimate of CH4 concentration at that location. Specifically, it compares the shape of CH_4 absorption to each radiance measurement in the cube to generate an initial abundance map of gas enhancements. This operation reduces the data cube from over 400 wavelength channels to a scalar value representing CH_4 abundance. Currently, AVIRIS-NG uses a matched filter approach to detect and quantify CH_4 enhancements (Thompson et al., 2015; 2016; Frankenberg et al., 2016). The matched filter characterizes the background using multivariate Gaussian distributions, whitens the data by this covariance structure, and matches the result to a spectral signature of methane absorption. The image is then orthorectified, resulting in a 2D image that contains an estimate of CH_4 enhancement for each pixel. This image, informally called a CH_4 map, is the starting point for Step 2, the focus of this benchmark. The CH_4 map is not perfect. Occasionally, erroneous high concentrations are estimated due to confusing features on the surface ("clutter") such as very bright surfaces or hydrocarbon paints, the signatures of which are very similar to CH_4. One of the key challenges of this application is thus to find algorithms that can distinguish between actual CH_4 sources and such erroneous signals.

Step 2: The second step aims to address these errors to reduce the image to a list of sources, or ideally, a binary mask of enhanced areas. Candidate approaches include morphological filtering or segmentation. This is currently done manually, by analysts inspecting the CH_4 map to build a list of discrete source features. These sources are typically plumes from oil and gas infrastructure but occasionally include geologic sources. As of the posting of this benchmark in 2017, this step currently requires manual intervention; human analysts must inspect the CH_4 map for each flightline (a strip of images obtained from one flight), manually spot and delineate plumes, and separate them from any non-plume artifacts. This allows the further reduction of the plume map to a simple source list. The manual approach is effective but not scalable to very large datasets, and the time is a significant challenge to applying this technique over global scales.

An automatic process would improve this situation by providing consistent, objective, and fast CH_4 source lists. Such an automatic plume labeling process would necessarily incorporate rejection of spurious detections, and permit automatic cataloging of the actual sources. In practice, the analysis often distinguishes between *point sources* that arise from specific locations and exhibit a plume-like morphology, and *diffuse sources* that arise from wider areas. It would be desirable, though not essential, to recognize these features automatically.

2.3 Data Format

The dataset consists of flightlines from the Four Corners campaign in April 2015. The dataset is a portion of this survey, split into flightlines representing individual contiguous image strips acquired by the spectrometer as it flew over the terrain.

Each flightline is associated with three files, two files representing the input to the desired task and one file representing the desired output of the task, obtained by hand.

1) **A header file that provides META data, with a filename suffex *img.hdr:**

   A detached ASCII human-readable header in ENVI format. This file describes the specific dimensioning and layout of the flightline, and embeds geographic projection information. It allows ENVI-compatible software toolboxes to directly read the binary data contained in the *img file.

2) **The imaging spectrometer data, with a filename suffix *img:**

   A binary file representing the CH_4 image data, with four channels:

   a. Radiance in visible red wavelengths
   b. Radiance in visible green wavelengths
   c. Radiance in visible blue wavelengths
   d. Estimated CH_4 enhancement above background, given in units of parts per million meters (ppm x m). The unused periphery of the image is marked by a "bad data" fill flag.

   The data is provided in Band Interleaved by Line (BIL) ordering. BIL ordering signifies the 3D matrix is indexed first by image row, then by channel, and then by the image column. The data is in 32 bit IEEE little endian floating point format.

3) **The results from the manual source analysis, with a filename suffixed *.png:**

   An image file in PNG format, representing a human labeling of the methane sources in the flightline. We will refer to this file informally as the “label file.” It has three channels with values in the [0-255] range, coded according to:

   a. Red (255,0,0): plume, believed to be associated with a point source
   b. Blue (0,0,255): plume, believed to be associated with a diffuse source
   c. Black (0,0,0): no plume (or unlabeled).

2.4 Reading and Visualizing the Data

To read the CH_4 data files, we highly recommend using existing open source toolboxes such as spectral python or the Matlab ENVI toolbox. Below we demonstrate how to read the data into Matlab.

Preparation - download and update the ENVI package:

1. Download Matlab toolbox ‘ENVI’ from https://www.mathworks.com/matlabcentral/fileexchange/27172-envi-file-reader-writer
2. You may need to update the ENVI package to newer Matlab versions by changing Line 66 of envihdrread.m to the following:
   ```
   line=textscan(line,'%s','Delimiter',',');
   ```
3. Add the directory with the ENVI files to the Matlab search path.

Download one flightline and load data into Matlab:

1. Download one data set, i.e. choose a flightline such as name = ' ang20150419t192226_cmf_v1f' and download the files 'name_img.hdr', 'name_img', and 'name_img_mask.png'.
2. Run the following command to load the spectrometer data:
   ```
   [D,info] = enviread(filename_image);
   ```
   where filename_image is the complete path to file 'name_img'.

3. Run the following command to load the manually labeled source data (png):
   ```
   [A_png,map_png,transparency_png] = imread( filename_png );
   ```
   where filename_png is the complete path to file 'name_img_mask.png'.

Appendix A provides sample Matlab code to read and visualize all files for one flightline. That code was used to generate Figures 3-5, by first running the code, then zooming in at the top of the resulting figures (the figures are linked so that they always show the same region). Fig. 3 shows the RGB image of the region including roads, buildings, vegetation and creeks. Fig. 4 shows the enhanced methane image, which indicates one primary methane source. Fig. 5 shows the hand-labeled methane identification result for this region, namely a single point source that results in a large methane plume upwind.

Figure 3: RGB image (Channels 1-3 of *img) of selected region.

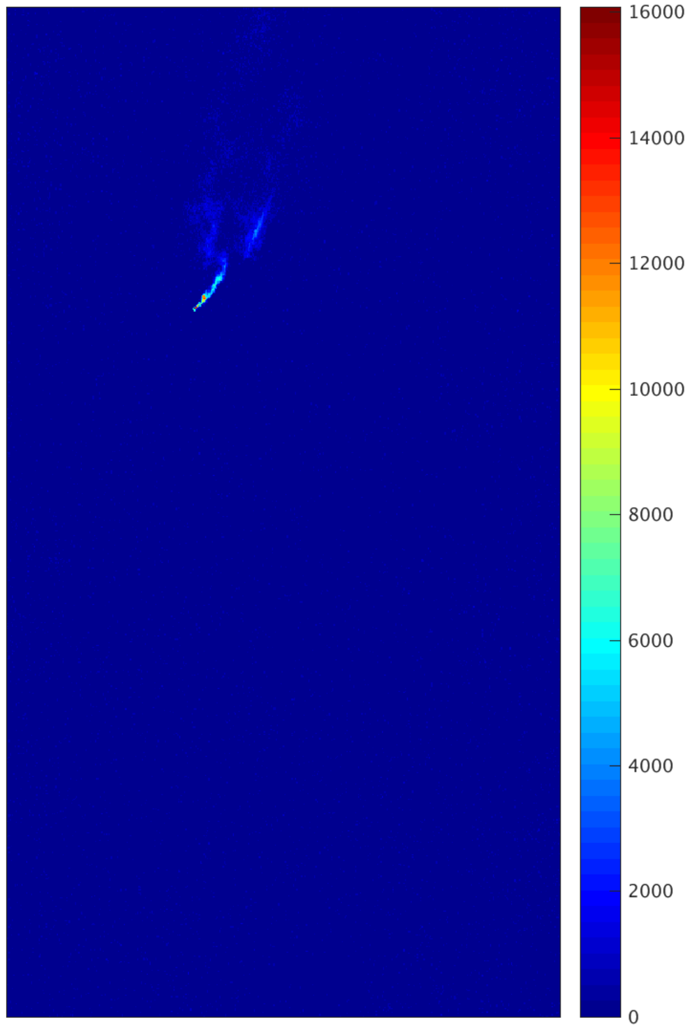

Figure 4: Plot of methane estimate (Channel 4 of *img) of same region as shown in Figure 3.

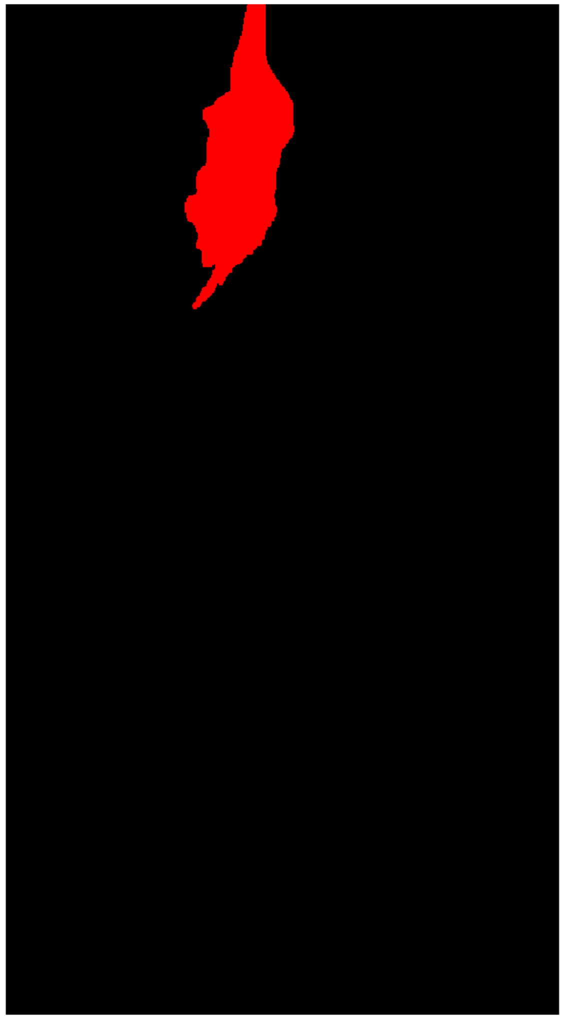

Figure 5: Hand-labeled PNG image indicating a CH_4 plume from a single point source in red (a diffuse source would be shown in blue).

3. Desired Analysis

This section discusses the specifics of the analyses that would most help geoscientists, in order to best understand and catalog greenhouse gas point source emissions. Key science questions include:

- Where are the most significant sources of CH_4 emissions?
- How do anthropogenic and natural sources of different classes compare?
- How do sources change over time?

At large scales, these questions rely accurate, consistent, and fully-automated interpretation of the CH_4 data.

3. 1 Analysis "Wishlist"

Objective 1: The primary goal of this analysis is to reliably identify sources from the CH_4 and RGB data. The desired output format for this main task consists of two elements:

1. A list of lat, lon coordinates of CH_4 sources, and
2. Visualization of segmentation in a format similar to the provided label images, for easy interpretation and to allow comparison with the existing label images.

The key challenge for Objective 1 is to find the right way to distinguish true CH_4 sources from background noise. Approaches that might be helpful are:

- Looking only at the CH_4 image, study/estimate the background noise (for each image separately or for all images together), then choose the right threshold to separate background noise from CH_4 source, to get segmentation.
- Morphology, looking at the shape of the plume and its intensity distribution to help decide which signatures are CH_4.

Learning from the RGB image, modeling what types of things create what kind of background noise ("clutter"), e.g. streets and other bright surfaces.

Objective 2 (optional): Estimate severity of CH_4 source, e.g. estimate amount of CH_4 in a plume; or even rate of CH_4 emission, as in Frankenberg et al. (2016).

Objective 3 (optional): Classify point sources vs. diffuse sources, or even the specific type of source. Classification of point and diffuse sources may be based on shape of intensity profile in CH_4 estimate. Classification of the specific type source may be based on the CH_4 estimate, but also objects recognized in close proximity of plume in RGB image. In a point source, CH_4 is emitted from a single location, i.e. one can identify a single pixel with peak values in the CH_4 estimate. Its signature in a CH_4 estimate consisting of a peak concentration at the source, and lower concentration downwind. These generally produce compact plumes with the source near the emission point at the upwind periphery. Typical point sources are anthropomorphic features such as pipeline leaks and venting. In a diffuse source, the CH_4 emits from an area significantly larger than a single pixel. Typical diffuse sources are geological seeps and dairy farms.

3.2 Potential Performance Measures

Suitable performance measures are still an open question and formulating them is part of ongoing research. Potential error measures for Task 1 include:

1. Comparing segmentation alignment between the label image and the automatic segmentation. For this purpose one would not distinguish between point and diffuse sources in the labels.
2. Mapping both label and input images to lat-lon coordinates and comparing the results. The lat-lon coorindates can define centroids of plumes, or, for point sources, the actual source pixel. If the latter, one should take intensity into account when calculating source locations, e.g.by using the label image as a mask over the CH_4 image, and then finding the highest intensity pixel.

Potential error measures for optional Tasks 2 and include estimating CH_4 flux (Task 2) and correct classification (Task 3) based on algorithm output and the manual labels.

4. How to Cite This Benchmark

Several sources contributed to this dataset. The Frankenberg et al. manuscript (2016) is the authoritative citation for the Four Corners campaign, and the Thompson et al. manuscript (2015) describes the step 1 analysis method which produced the CH_4 enhancement maps. For referencing the manual labels and problem description, we suggest citing this document:

D. R. Thompson, A. Karpatne, I. Ebert-Uphoff, C. Frankenberg, A.K. Thorpe, B.D. Bue, R.O. Green, *IS-GEO Dataset JPL-CH4-detection-2017-V1.0: A Benchmark for Methane Source Detection from Imaging Spectrometer Data,* JPL Records Document D-101028, available from http://avirisng.jpl.nasa.gov

The IS-GEO benchmark team has adopted the following convention for referencing this dataset:
JPL-CH4-detection-2017-V1.0
Using this unique identifier will make it easy to identify all publications using this benchmark.

5. Acknowledgements

This benchmark description was developed as part of the IS-GEO Research Collaboration Network funded by NSF Award #1632211, *EarthCube RCN IS-GEO: Intelligent Systems Research to Support Geosciences*, with some initial auxiliary support by NSF #1526667, *NRI*: *Collaborative Research: Co-Exploration using Science Hypothesis Maps*. AVIRIS-NG is sponsored by The NASA Earth Science Division. The data set was provided by and a part of the research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

A. Sample Matlab Code to Read and Visualize a Flightline

Below we include a very basic sample code to read and display the data for one flightline. This demonstrates how to *load and visualize* the data.

```
my_dir = '../NEW_SAMPLE_DATA';  % path to whereever your data files are
flightline_name = 'ang20150419t192226';  % choose the flightline

my_filename = sprintf('%s_cmf_v1f_img', flightline_name); % basic filename
my_filename_image = sprintf('%s/%s', my_dir, my_filename); % absolute path for *img
my_filename_png = sprintf('%s_mask.png', my_filename_image ); % absolute path for *.png

%------------------------------------------ READ DATA
% Read spectrometer data:
[D,info] = enviread(my_filename_image);

% Read hand-labeled methane source file:
[A_png,map_png,transparency_png] = imread( my_filename_png );


%------------------------------------------- ANALYZE DATA

NaN_code = info.data_ignore_value;
cut_off = 500;  % sample cut-off value for methane (min value)

D(D == NaN_code) = nan; % replacing fill values with nan
[nr, nc] = size(D(:,:,1));

ch4 = D(:,:,4);
ch4(ch4 < cut_off) = 0; % applying cutoff to fourth channel

%------------------------------------------- Create figure for methane estimate
figure(1);
set( gcf, 'Position', [100 0, 1100, 1100] );
imagesc( ch4 ); colorbar; axis equal;
g_handle_methane_estimate = gca;
my_title = sprintf( 'Methane Estimate for flightline %s', flightline_name );
title(my_title);

%------------------------------------------ Create figure for RGB image
% Creates a figure for a selected region with values in all three channels scaled
% to vary from 0 to 1 in the selected region. Run this code snippet every
% time Figure 1 is updated for best stretching of RGB image.

% get the dimensions of the region currently selected in Figure 1.
figure(1); axis_curr = axis;
rmin = max(floor(axis_curr(1)),1); rmax = min(ceil(axis_curr(2)),nr);
cmin = max(floor(axis_curr(3)),1); cmax = min(ceil(axis_curr(4)),nc);
RGB_image = D(rmin:rmax,cmin:cmax,1:3);
for i=1:3   % for red, green, blue
   im_curr = RGB_image(:,:,i);
  % find minimum and maximum values
  val_min = nanmin(im_curr(:)); val_max = nanmax(im_curr(:));
  % stretching the band to range from 0 to 1 in the selected region
```

```
 im_curr = (im_curr – val_min)./(val_max - val_min);
 im_curr(isnan(im_curr)) = 1; % nan values are shown as white.
  RGB_image(:,:,i) = im_curr;  % increase contrast
end

figure(2);
set( gcf, 'Position', [100 0, 1100, 1100] );
imagesc(RGB_image); axis equal;
g_handle_RGB = gca;
my_title = sprintf( 'RGB image for flightline %s', flightline_name );
title(my_title);

%---------------------------------------- Create figure for hand-labeled source results (label file).
% Run this code snippet every time Figure 2 is updated.
figure(3);
set( gcf, 'Position', [100 0, 1100, 1100] );
imagesc(A_png(rmin:rmax,cmin:cmax,:)); axis equal;
g_handle_PNG = gca;
my_title = sprintf( 'Manually deteted methane sources for flightline %s', flightline_name );
title(my_title);

%---------------------------------------  Link zooming and shifting actions of all three figures.
linkaxes( [g_handle_methane_estimate, g_handle_RGB, g_handle_PNG], 'xy');
```

References

Ebert-Uphoff, I., D. R. Thompson, I. Demir, Y.R. Gel, M.C. Hill, A. Karpatne, M. Guereque, V. Kumar, E. Cabral-Cano, P. Smyth, *A Vision for the Development of Benchmarks to Bridge Geoscience and Data Science*, Proceedings of the Seventh International Workshop on Climate Informatics (CI 2017), NCAR Technical Note NCAR/TN-536+PROC, 2017.

Frankenberg, C., A. Thorpe, D. R. Thompson, G. Hulley, E. Kort, N. Vance, J. Borchard, T. Krings, K. Gerilowski, C. Sweeney, S. Conley, B. Bue, A. Aubrey, S. Hook, and R. O. Green. Airborne methane remote measurements reveal heavy-tail flux distribution in Four Corners region. *Proc. National Academy of Sciences*, 113 (35), 2016.

Kort, E. A., C. Frankenberg, K. R. Costigan, R. Lindenmaier, M. K. Dubey, and D. Wunch. "Four corners: The largest US methane anomaly viewed from space." *Geophysical Research Letters* 41:19, 6898-6903, 2014.

Thompson, D. R., A. K. Thorpe, C. Frankenberg, R. O. Green, R. Duren, A. Hollstein, L. Guanter, E. Middleton, L. Ong, S. Ungar, Orbital Measurement of the Aliso Canyon CH_4 Super-emitter. *Geophysical Research Letters*. 43, 6571–6578, 2016.

Thompson, D. R., I. Leifer, H. Bovensmann, M. Eastwood, M. Fladeland, C. Frankenberg, K. Gerilowski, R. O. Green, S. Kratwurst, T. Krings, B. Luna, and A. Thorpe. Real-Time Remote Detection and Measurement for Airborne Imaging Spectroscopy: A Case Study with Methane. *Atmospheric Measurement Techniques*, 8, 2015, p. 4383-4397.

Thorpe, A. K., C. Frankenberg, A. Aubrey, D. A. Roberts, A. A. Nottrott, T. Rahn, J. A. Sauer, M. K. Dubey, K. R. Costigan, C. Arata, A. Steffke, S. Hills, C. Haselwimmer, D. Charlesworth, C. C. Funk, R. O. Green, S. R. Lundeen, J. W. Boardman, M. L. Eastwood, C. M. Sarture, S. H. Nolte, I. B. Mccubbin, D. R. Thompson, J. P. McFadden. Mapping methane concentrations from a controlled release experiment using the next generation airborne visible/infrared imaging spectrometer (AVIRIS-NG). *Remote Sensing of Environment,* 179, 2016.