

The Effects of Enhanced Hydrocarbon Extraction Technologies on Greenhouse Gas Emissions from Upstream and Midstream Oil and Gas Industry

Problem Statement:

The purpose of this project is to examine established Greenhouse Gas (GHG) Emissions data and determine if there is a correlation between accelerated hydrocarbon (Crude oil and Natural gas) production through horizontal drilling and hydraulic fracturing in the most recent decade and GHG emissions rates. If a correlation is found, additional analysis may be able to reveal the causal source. The project will focus on three GHGs, Carbon-Dioxide (CO₂), Methane (CH₄), and Nitrous-Oxide (N₂O).

Questions to Answer:

Is there a correlation between increased hydrocarbon production and industry sector GHG emissions? Does the correlation apply to the industry as a whole or to one or more individual components?

Team Composition:

The Greenhouse Gas Emissions investigation team is composed of Rafee Shaik and Greg Myers, UMSI MADS students. Both members have professional backgrounds in domestic Crude Oil and Natural Gas production.

Source Datasets:

1. Name: Greenhouse Gas Emissions Data

Source: US Oil and Gas Upstream (Exploration & Production) and Midstream (Pipelines and Storage) facilities can be accessed from the U.S. Environmental Protection Agency (EPA) FLIGHT database.

Location: <https://ghgdata.epa.gov/ghgp/main.do>

Access Method: Facility Level Information on GreenHouse gases Tool (FLIGHT) database can be accessed through the website: <https://ghgdata.epa.gov/ghgp/main.do>
Download the data in excel format after applying appropriate filters.

Format: Excel spreadsheets

Dataset Size: Six excel spreadsheets with a total of 41K records.

2. Name: Emission Data from other Industrial Sectors

Source: Emissions data from other Industrial sectors can be downloaded from the EAP data store. We will use this data to compare GHG emissions from Oil & Gas systems and other industrial sectors.

Location: <https://cfpub.epa.gov/ghgdata/inventoryexplorer/#industry/allgas/source/all>

Access Method: Web Scraping if there are no popups preventing web scraping, otherwise use the manual download option.

Format: CSV or Web Scraping, CSV name will be IndustryWiseGHGEmissions.csv, if we're not able to scrape it from the web.

Dataset Size: 243 records

3. Name: Crude Oil and Natural Gas Production

U.S. Field Production of Crude Oil, U.S. Natural Gas Gross Withdrawals; Yearly

Source: Energy Information Administration (EIA) datastore.

Location: <https://www.eia.gov/opendata/qb.php?category=371>

Access Method: API query

Format: JSON

Dataset Size: Crude records: 161; Gas records: 84; ~3 kilobytes each for crude and oil

Data Manipulation Methods:

1. Processing Emissions datasets:

- CH₄ and N₂O emissions are represented in CO₂ equivalent quantities.
- Data is present in multiple sheets, a sheet for each year. Load each of these files into a separate data frame and flag what type of gas and which sector it is.
- Aggregate the emission quantities by the parent company
- Combine all the six datasets into a single data frame. Fill 'NaN' values with zero

2. Processing Emissions from Other industrial sectors:

- Data is present in wide-format, unpivot the data to put it in long-format, use pandas melt method.

3. Processing Crude and Natural Gas Production volumes datasets:

- Save API query results to a JSON file as an immutable source data reference.
- Import JSON data (crude & gas) into Pandas data frames.
- Perform Explode operations to separate date and production data.
- Transform Date column into a DateTime data type.
- Create and populate a product type column and drop unused columns.
- Append crude and gas data frames (long-format).
- Save the data frame in CSV format as a source for continued analysis.

4. Data Integration:

- Combine the above two datasets (Emissions and Production volumes) by the year, with this combined dataset we can compare greenhouse gas emission volumes with O&G production volumes.

Analysis and Visualization:

A regression analysis will be conducted to estimate missing emissions data from 1990 to 2010 and for 2019. If an acceptable R-Squared value can be obtained from the regression analysis, the estimated data points will be used in the following visualizations.

A scatterplot will be used to observe the potential correlation between hydrocarbon production and Greenhouse Gas emissions. Another visualization of interest will plot GHG trend curves against U.S. Natural Gas and Petroleum production curves in the time frame beginning year 2010 to present. If time permits, an interactive heatmap will be prepared to show the GHG emissions and Hydrocarbon production at the US state level.

Team contribution:

Rafee Shaik will download and prepare all emissions related datasets. Greg Myers will download and prepare all production-related datasets. Both members will collaborate on data integration methods and combine effort into a single code base. Both members will collaborate on visualizations and select the best combinations and integrations.