Monitoring Insights

Substitute data in EPA CAMD's Power Sector Emissions Data

February 16th, 2021

Substitute data represents a fraction of a percent of total emissions data.

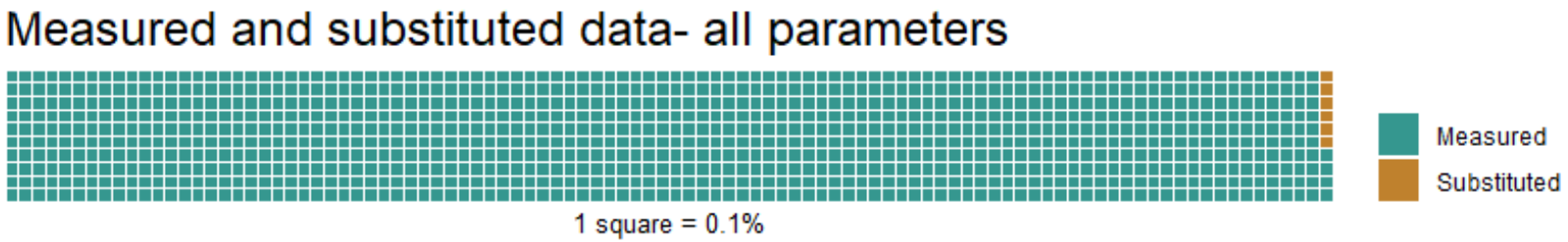
Under the Clean Air Act, most fossil fuel-fired power plants must continuously monitor and report their emissions of carbon dioxide (CO_2), nitrogen oxides (NO_X), and sulfur dioxide (SO_2) to EPA.[1] Emissions are monitored by continuous emission monitoring systems (CEMS) or equivalent that power plants install and maintain. The power plants collect and report the hourly emissions data to EPA every calendar quarter and EPA publishes those data on its website.

While the CEMS measure continuously, there may be operating hours when a CEMS does not provide valid data due critical system malfunctions, missed or failed quality assurance tests, routine maintenance, or other problems. When data is missing or invalid, EPA's regulation[2] specifies how to estimate and substitute the data. The longer or more frequent the missing or invalid data, the more conservative (i.e., likely to overestimate emissions) the substitute data algorithm becomes. Because the emission data are used to assess compliance with several cap-and-trade programs, affected power plants do not want to overreport emissions. Therefore, they have an incentive to minimize the amount of missing or invalid emission data.

[1] Refer to 40 CFR part 75—Continuous Emission Monitoring for more information about the monitoring and reporting requirements

[2] Refer to 40 CFR part 75 Subpart D, §§75.30-37 for more information about the missing and substitute data provisions

Substitute data are a minor portion of total emission data



99.4%

of data for all parameters was measured (2015-2019)

As shown in the figure above, in years 2015-2019, 0.6% of operating hours reported to EPA had missing or invalid data for the following parameters.

- CO_2 concentration
- NO_X concentration
- O_2 concentration
- SO_2 concentration
- stack gas flow

Substitute data are a minor portion of total emission data

Percent of substitute data by parameter

Parameter	Substituted
CO_2 concentration	0.8%
NO_X concentration	0.4%
O_2 concentration	0.2%
SO_2 concentration	0.9%
Stack gas flow	1.1%

Refer to the table above to review the percent of substitute data by parameter from years 2015-2019.

Substitute data range from 0.2 to 1.1 percent of total operating hours

Substitute data have different effects on emission estimates

The data substitution algorithms vary in their likelihood to overestimate emissions. A power plant must apply the appropriate algorithm based on the duration and frequency of missing or invalid data. For example:

- When the duration and frequency of missing or invalid data is low, the power plant averages the valid hours before and after the missing data period and applies that value to the missing data period. This approach is unlikely to significantly overestimate emissions.
- When the duration and/or frequency of missing or invalid data is long, the power plant may have to report maximum potential concentration or flow rate, regardless of operating level. This approach is likely to overestimate emissions.

Substitute data can be categorized into three tiers of estimation to indicate the likelihood that emissions are overestimated.

Tier 1: Low likelihood of overestimation

Tier 2: Moderate likelihood of overestimation

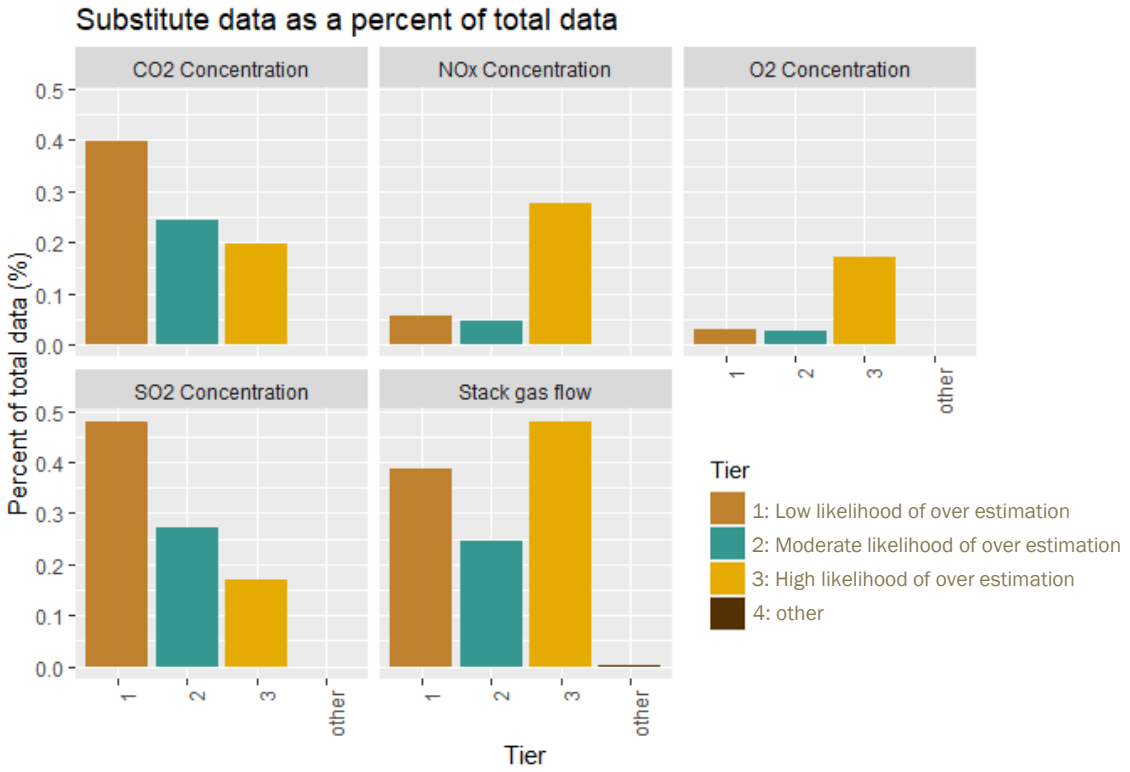
Tier 3: High likelihood of overestimation

A small portion of substitute data hours cannot easily be categorized because they are reviewed on a case-by-case basis and do not fall into the traditional algorithm. These are listed as “other”.

Substitute data varies by parameter

This figure illustrates the percent of total operating hours using substitute data and separates the data into the three tiers of estimation by parameter.

- For CO_2 concentration and SO_2 concentration tier 3 (highest likelihood for overestimation) substitution accounts for 0.2% or less of total data and represents the smallest percent of substitute data for those parameters.
- For NO_X concentration, O_2 concentration, and stack gas flow tier 3 substitution accounts for 0.5% or less of total data and represents the largest percent of substitute data. Although tier 3 represents the majority of the data, 0.5% is an insignificant percentage of the total data.



For more information about the data or this analysis…

EPA's part 75 monitoring and reporting program

- 40 CFR part 75—Continuous Emission Monitoring
- Plain English Guide to Part 75 (PDF)

Power Sector Emissions Data

- CAMD's Power Sector Emission Data
- CAMD's Power Sector Emissions Data Guide (PDF)

Contact information

Stacey Zintgraff
EPA's Clean Air Markets Division
zintgraff.stacey@epa.gov

Categorizing method of determination codes

Every hourly measurement includes a method of determination code (MODC) to inform how the value was measured or calculated.

The MODCs can be categorized into measured and calculated. The calculated MODCs can be further categorized by the likelihood that the associated algorithm will overestimate emissions.

For more information about MODCs, refer to the part 75 reporting instructions.

Categories of MODC values

Category	MODC
Measured	1-5, 14, 16-17, 21-22, 26, and 40
Tier 1—Low likelihood of overestimation	6-7 and 11
Tier 2—Moderate likelihood of overestimation	8-9
Tier 3—High likelihood of overestimation	10, 12-13, 15, 18-20, 23-25, and 46-48
Other	53-55

Analytical methodology

This analysis was completed in RStudio. If you would like to review the code or the source data, contact Stacey Zintgraff to make the request. To complete this analysis, we …

Summarized steps

1. Created a data frame consisting of all operating hours, including measured and substituted data by parameter.
2. Calculated percent of operating hours measured versus substituted by parameter.
3. Categorized substituted data into the tiers of estimation and calculated percent of operating hours in each tier by parameter.

By the numbers (All Parameters)

- Power plant combustion units:
  - 3,963 units
- Hours of operation:
  - 181,189,654 hours
- Measured hours of operation:
  - 180,084,949 hours
- Tier 1 substituted data hours of operation:
  - 388,837 hours
- Tier 2 substituted data hours of operation:
  - 246,793 hours
- Tier 3 substituted data hours of operation:
  - 468,781 hours
- Other substituted data hours of operation:
  - 294 hours