|  |
| --- |
| Engineering and Economics of Energy Research Group |
| Guide to the Data Matchup Procedure |
| Matching the EIA, EPA, and EV datasets |
|  |
| **Charles Marquet, Jubo Yan, Daniel Shawhan** |
| **3/31/2013** |

|  |
| --- |
|   |

Table of Contents

[Section 1: Introduction 1](#_Toc352859360)

[Section 2: Preparing the Datasets 1](#_Toc352859361)

[Preparing the EIA Dataset 1](#_Toc352859362)

[Preparing the EPA Dataset 2](#_Toc352859363)

[Section 3: Matching Energy Information Administration (EIA) data files 2](#_Toc352859364)

[Description of software use: 2](#_Toc352859365)

[Description of algorithm: 2](#_Toc352859366)

[Section 4: Matching the EIA dataset to the EPA dataset 3](#_Toc352859367)

[Description of our matching algorithm: 3](#_Toc352859368)

[Summary of algorithm: 5](#_Toc352859369)

[Inputs needed for the program: 5](#_Toc352859370)

[Output from the program: 5](#_Toc352859371)

[Section 6: Estimating Missing Data 5](#_Toc352859372)

[Description of our algorithm: 5](#_Toc352859373)

[Inputs needed for the programs: 6](#_Toc352859374)

[Output from the programs: 6](#_Toc352859375)

# Section 1: Introduction

Simulation modeling of the power grid that includes costs and emissions requires several types of data about the generators on the system. The US federal government collects and publishes a considerable amount of data about generators, but those data are in multiple datasets published by the Energy Information Administration and the Environmental Protection Agency. We have developed methods of identifying which data in each dataset corresponds to which generator in the Energy Information Administration’s annual basic dataset of currently operating generation units, of predicting missing values, and adding this additional information to that basic dataset. The result is a dataset with the electrical capabilities, heat rates, and emission rates of almost all generators in the contiguous United States. This document summarizes how we produced this dataset.

# Section 2: Preparing the Datasets

## Preparing the EIA Dataset

1. All files mentioned in this section are downloaded from [www.eia.gov](http://www.eia.gov)
2. Energy Information Administration (EIA) updates their dataset annually. Before using the most updated datasets, you need to check their have the same file structure as what we used here. If not, you need to manually adjust the column names and orders to match the files we use here. If not properly adjusted, errors may occur without warning!
3. In specific, you will need files named “GeneratorY20XX”, “EnviroAssocY20XX” and “EnviroEquipY20XX”. “20XX” represents the specific year EIA releases their datasets. The posted version here uses 2011 datasets.
4. We have included three sample files in the package for your use. You should compare data files from future years with the sample files if you decide to use the newer version.

## Preparing the EPA Dataset

Two columns were added to the original EPA dataset. One was a fuel type column, for which a description can be found in Section 3. The second column contains our estimated value for the capacity of each EPA unit, the simplest way to calculate this column is by dividing GLOAD by the number of operating hours. We calculated it using the hourly CEMS data, using GLOAD only when above 70% of max GLOAD.

# Section 3: Matching Energy Information Administration (EIA) data files

In the first part of this section, we describe how to utilize the matching software. In the second part, we describe the algorithm we used in the matching process.

### Description of software use:

1. After downloading the necessary files described in Section 2, put all files into a SINGLE folder. On sheet “Control”, correctly specify the folder address and file names. For example, if you downloaded all files to your “C:\data” folder, specify “C:\data” as your folder address (no quotes). The best way to avoid mistake would be to copy the address from your explore. Files names require correct extensions (.xls in our case).
2. Click buttons on sheet “Control” ONE BY ONE. Depending on the computer you use, each step could take a significant time to run. Do NOT proceed to the next step until you see the successful message. Usually, you will be notified several times in each step, just click OK to continue.
3. If you make ANY mistake in ANY step, you need to delete all sheets and start over the process. Continuing from the middle of the process may result in an incorrect matched dataset. After successfully executing each step, you may choose to keep only the result sheet and delete all other sheets by clicking the corresponding button.
4. In case you need to use a newer EIA dataset, go to sheet “Arrange” to correspond the new column names to the column names we use here. The standard column names are in italic font and newer column names are in bold font. Do NOT change the standard names. Manually match newer names to the standard names. For example, in a newer dataset, PLANT\_CODE is called FACILITY\_CODE, and then write FACILITY\_CODE in the cell under PLANT\_CODE. You need to repeat this for each column in each sheet. In addition, you need to change the names correspondingly in sheet “index”. This sheet is used to keep columns you want. If you want less/more columns in the match dataset, you can remove/add column names on line 6.

### Description of algorithm:

1. Program copies all necessary sheets into the Excel file.
2. Match environment equipment information to each boiler and eliminate duplicate rows. This information is matched to each boiler using boiler IDs. In the case a boiler is connected to multiple equipment, the matching results in multiple records for a single boiler, we consolidate these records into one by taking average.
3. Match generators with ESH and environmental information using boiler\_id. Generators are matched to each boiler by boiler\_id.
4. Since a generator may be connect to one or multiple boilers, the matching in step 4 results in multiple records for a single generator. We consolidate these records into one by taking average of corresponding information.
5. Match existing generator list with environment information. This step match all information we have collected so far to the EIA generator EIA list using generator ID.
6. Remove unnecessary columns. This step removes all columns except the ones specified in line 6 in sheet “index”. As mentioned above, you may choose to customize the columns to keep.

# Section 4: Matching the EIA dataset to the EPA dataset

Run the match.java program. This program reads the EIA and EPA data files and tries to match the generator data based on generator ID, capacity and type. The result is a file with all EIA units and the matched EPA units.

### Description of our matching algorithm:

* 1. Match EIA plants with EPA plants using the “plant code” AKA “ORIS code” or “ORISPL code.”
	2. Narrow each pair of matchable sets of units as much as possible, to make the matching easier.
		1. Narrow them to just the units at that plant that use a particular energy source category. The EIA “layout”/”reference tables 1” sheet in the “f860y…” folder (which you have) has a key to the meaning of EIA energy source codes, if you’d like to see what they mean.

|  |  |  |
| --- | --- | --- |
| Category name | EIA “ENERGY\_SOURCE\_1” codes that fit in this category | EPA energy sources that fit in this category |
| Coal | “ANT”, “BIT”, “LIG”, “RC”, “SUB”, “WC” | “Coal”“Coal Refuse”“Coal, Pipeline Natural Gas”“Coal, Wood” |
| Liquids | DFORFOJF, KER, BLQ, OBL, WDL, WO | “Diesel Oil”“Diesel Oil, Other Gas”“Residual Oil”“Other Oil” |
| Petroleum Coke | PC | “Petroleum Coke” |
| Gases | NG(ENERGY\_SOURCE\_1 equals any of these: BFG, LFG, OBG, OG, PG, SG, SGC) | “Natural Gas”“Pipeline Natural Gas”“Other Gas”“Process Gas” |
| Other solid | ABMSWOBSSLWWDS | “Wood”“Other solid fuel” |
| Tires | TDF | “Tire derived fuel” |

* + 1. Further narrow each pair of matchable sets of gas-fired and liquid-fired units according to the “unit type” (“UNIT\_TYPE\_INFO” column in EPA data, “PRIME\_MOVER” in EIA’s Generator/Exist sheet). Subcategorize the gas & liquid units into the following categories:

|  |  |  |
| --- | --- | --- |
| Category Name | EIA | EPA |
| Combustion turbine part of combined cycle set | “CT” or “CS” in PRIME\_MOVER column of Generator/Exist sheet.  | “Combined Cycle” in UNIT\_TYPE\_INFO column of EPA sheet. |
| Steam part (or “steam turbine part”) of combined cycle set. | “CA” in PRIME\_MOVER column of Generator/Exist sheet. | Based on observation, we decided not to try to match these EIA units because we don’t think they are in the EPA dataset.  |
| Combustion turbine that is not part of combined cycle set | “GT” in PRIME\_MOVER column of Generator/Exist sheet. | “Combustion Turbine” or “Other Turbine” in UNIT\_TYPE\_INFO column of EPA sheet. |
| All other gas-fired units | All other PRIME\_MOVER codes | All other UNIT\_TYPE\_INFO codes |

* 1. Exclude from the matchable sets any units that we know are unlikely to appear in the other dataset.
		1. Do not try to match an EIA unit with an EPA unit if any of the following applies in the EIA data for that unit:
			1. STATUS is “OS”. This means the unit was out of service for some or all of the year and is not expected to return to service during the coming year.
			2. PRIME\_MOVER is “CA”.
		2. If we are using an EIA dataset that is from the year preceding the EPA dataset, do not try to match an EPA unit with an EIA unit if any of the following applies in the EPA data for that unit:
			1. “OP\_STATUS\_INFO” = “Operating (Started…” This means it started operation after the beginning of 2010. In that case, it would not be in the EIA dataset which is for Dec. 31, 2009, so we should not try to match it with the EIA dataset.
	2. Once these three steps are done, you have the units divided by plant, further divided by primary fuel type and unit type, and you have removed the units in one list that are unlikely to appear in the other list. The hard part is the remaining EIA-EPA matchup, because in many cases there is more than one unit with a given primary fuel at a given plant.
		1. To select the matching unit, we used a score based on multiple criteria:
			1. Unit code
			2. Generation capacity
			3. Secondary energy source (“Energy\_source\_2” in EIA list, “Secondary fuel info” in the EPA list)
			4. The unit’s estimated capacity from the EPA data. Units with nameplate capacities under 25 MW are less likely to have to report, so EIA units with nameplate capacities less than 25 MW, especially gas-fired ones, are less likely to have a match in the EPA dataset.

### Summary of algorithm:

The program goes through each EIA unit one at a time. For each EIA unit:

1. It finds the set of units in the EPA data corresponding to the same plant code as the current EIA unit.
2. It assigns a score to each of the EPA units in the set, based on the similarities the unit matches the current EIA unit
3. The EPA unit with the highest score is matched with the current EIA unit, as long as it passes some validity tests (e.g. close enough capacities)

### Inputs needed for the program:

1. EIA-Stack dataset, csv format
2. EPA data set, csv format (modified as shown in section 2)

### Output from the program:

1. The match.csv file, with all EIA units and the matched EPA units.

# Section 6: Estimating Missing Data

Run the regression\_prepare.java program. This program prepares a valid dataset for Stata regression. Run emission\_regression\_new-full.do in Stata. This code runs multiple linear regression to predict emission and heat rates for all units (regressions are ran separately for each type of generator). Run the regression\_finnish.java program. This program uses the output of the regression analysis to add the emission and heat rates to all units. It also checks the results of the regression are reasonable, and replaces extreme values if necessary.

### Description of our algorithm:

1. First we create a valid dataset for regression
	1. Explanatory variables include: prime mover, energy source, capacity, age, latitude, longitude, use of certain pollution reducing equipment, and combinations of these.
	2. Response variables: CO2 rate, NOx rate, SO2 rate, heat rate.
	3. All these were calculated based on the EIA & EPA matched data
2. We ran multiple linear regressions to predict emission and heat rates for all units
	1. Regressions for each of the response variables on a subset of the explanatory variables
	2. Regressions ran separately for each “type” of generator. We defined types by using all combinations of prime mover and energy source.
	3. For generator types with fewer than 30 observations (but greater than 0), regressions were not ran. Instead they were aggregated into an “other” category. Once all “other” generators are compiled a single regression is ran on them.
	4. For generator types 0 observations, emission rates were set to 0, heat rates to an average value.
3. We used the result from the above to update our EIA-EPA dataset with emission rates and heat rates for all units.
	1. If unit had actual data, we used the actual emission rates
	2. Otherwise we used the estimated rate from the regressions
	3. The code checks for extreme values. Maximum and minimum emission/heat rates are compiled for each generator type. If the estimated values are lower than the minimum, they are set to the minimum value. If the estimated values are higher than the maximum, they are set to the maximum value.

### Inputs needed for the programs:

1. EIA-EPA dataset, csv format (see section 4)

### Output from the programs:

1. The EIA-EV-emissions\_Gens.csv file, with all EIA-EPA units (and emission/heat rates for all units)