Truth or Dare: Data Augmentation in the Point Source 2002 NEI

Anne Pope and Madeleine Strum
Emission Factor and Inventory Group, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711
pope.anne@epa.gov
strum.madeleine@epa.gov

Stephanie Finn
Eastern Research Group, Inc., 1600 Perimeter Park, Morrisville, NC 27560
stephanie.finn@erg.com

ABSTRACT

The Environmental Protection Agency (EPA) compiles the National Emission Inventory (NEI) for hazardous air pollutants (HAPs) and criteria air pollutants (CAPs). The NEI plays an important role in air quality management activities such as emission trends, rule and policy development and risk assessment. To support many of the these functions, the NEI must contain the data necessary for air quality modeling.

The NEI contains estimates of facility-specific HAP and CAP emissions and their source-specific parameters necessary for modeling such as location and facility characteristics (stack height, exit velocity, temperature, etc.). Complete source category coverage is needed, and the NEI contains estimates of emissions from stationary point and nonpoint (i.e., stationary sources such as residential heating that are inventoried at the county level) and mobile source categories. This paper focuses on a particular aspect of the development of the point source inventory. Point source categories include major and area sources as defined in section 112 of the CAA.

Key processing activities for point source data include submittal of 2002 inventory data by state and local agencies, tribes, EPA, and industry; blending/merging of data from multiple data sources; augmentation of data for missing data elements; QC/QA of the data; preparation of draft NEI for external review; incorporation of external review comments; and preparation of final NEI. An important step in preparing point source files is augmentation of data. The EPA conducts a variety of QA steps to identify any missing or out-of-range parameters which are needed for air quality and exposure modeling. After conducting QA, the EPA augments data using a published methodology. The NEI data files identify all records that have defaulted parameters and the methodology used to default data fields.

This paper discusses the methodology EPA employs to identify and augment point source data with missing or out-of-range values.

INTRODUCTION

The Emission Factor and Inventory Group (EFIG) in the Environmental Protection Agency (EPA) compiles the National Emission Inventory (NEI) for hazardous air pollutants (HAPs) and criteria air pollutants (CAPs). The NEI for HAPs is compiled in order to support air the EPA air toxics programs and to quantify the the success of the Clean Air Act (CAA) programs in reducing emissions and human health and environmental risk due to HAPs emissions. Title I, Section 110 of the CAA requires states to submit emission inventories for CAPs as part of their State Implementation Plans.

The NEI contains estimates of facility-specific HAP and CAP emissions and their source-specific parameters necessary for modeling such as location and facility characteristics (stack height, exit velocity, temperature, etc.). Complete source category coverage is needed, and the NEI contains estimates of emissions from stationary point and nonpoint and mobile source categories. This paper focuses on the development of the point source inventory.   Point source categories include major and area sources as defined in section 112 of the CAA.

The major steps involved in compiling the 2002 point source NEI include:

- Submittal of 2002 inventory data by state and local agencies, tribes, industry, and EPA offices;
- Blending/Merging of data from multiple data sources;
- Augmentation of data for missing data elements;
- QC/QA of data;
- Preparation of draft 2002 NEI for external and internal review;
- Incorporation of external and internal review comments on the draft 2002 NEI and incorporation of new inventory data submitted during review period; and
- Preparation of final 2002 NEI.

Important steps in preparing point source files are Quality Assurance (QA) and augmentation of data. QA and data augmentation are needed to prepare point source files for use in air quality and risk and exposure modeling. In addition, the QA of data and data augmentation are needed in order for the NEI to meet recent EPA data standards and Office of Management and Budget's (OMB) Information Quality Guidelines.

The EFIG conducts a variety of QA activities to identify records with referential integrity problems, duplicate records, and records with missing or out-of-range parameters which are needed for air quality and exposure modeling. The EFIG summarizes the errors found and reports back to the data providers on the QA findings. The EFIG tracks errors, their resolution and all communications on a QA/QC form, through emails, and in a phone log as part of documentation. These tracking mechanisms help to ensure the transparency and reproducibility of the NEI. This is a requirement of OMB Information Quality Guidelines, but it also helps EFIG establish an electronic trail for each record in the NEI. The EFIG has created a QA/QC process and tracking database to provide feedback reports to data providers at regular intervals during the QA of the data. The EFIG archives all files submitted by data providers, records removed during QA, records augmented, and each iteration of the draft and final

NEI.

The EFIG first resolves records with referential integrity problems and duplicates. Then after identifying parameters and data fields with missing or out-of-range values during the QA of files, the EFIG augments or defaults the data. The NEI data files identify all fields of data that have defaulted parameters. Data augmentation occurs at different times in the compilation of the NEI. For example, augmentation of location coordinates occurs prior to blending/merging of data, while augmentation of stack parameters occurs after blending/merging of data. The EFIG has recently revised the memorandum, "NEI Quality Assurance and Data Augmentation Steps for Point Sources", June 2003, with the report, "NEI Quality Assurance and Data Augmentation for Point Sources" May 2004.[1]

Data providers submit their data to EPA using the NEI Input Format (NIF) or an extensible markup language (XML) format. Both formats are described and located at http://www.epa.gov/ttn/chief/nif/index.html. The 2002 point source data will be submitted to EPA in NIF V3.0 or NEI XMLV3.0 for incorporation into the 2002 NEI. The point source NIF V3.0 and NEI XML V3.0 contain more than 115 data fields. In addition EFIG includes additional data fields in the NEI that are not included in the NIF V3.0 Tables. These fields may be in auxiliary files or may be included in NIF tables as output data fields.

This paper discusses QA and data augmentation for Location Coordinates, Stack Parameters, and HAP Pollutant Codes. For details on how EFIG will QA and augment the remaining NIF fields, please refer to the May 2004 augmentation report.

LOCATION COORDINATES

The first step in blending and merging of point sources is to QA geocoordinates and correct erroneous coordinates. Latitude and longitude are also needed to correctly place facility emission release points and associated emissions into specific geographic domains (grid cell, census tract, etc.) for proper emissions modeling.

Location Coordinate QA and Augmentation Procedure

All UTM coordinates will first be converted to latitudes and longitudes based on 1984 World Geodetic System datum.

The longitude should be reported in units of decimal degrees with a negative sign and the latitude should be reported in units of decimal degrees with a positive sign in NIF. We will QA the reported latitude and longitude to determine if units are reported in degrees-minutes-second. If the reported value does not have a decimal or if it does not agree with state and county FIPS and latitude and longitude within NEI County FIPS Lookup Table, we will look at the reported value and contact the data provider. If the latitude and longitude are reported in degrees-minutes-seconds, we will convert the latitude and longitude to decimal degrees using the following equations.

Latitude (decimal degrees) = Latitude degrees + Latitude minute/60.0 + Latitude second/3600.0

Longitude (decimal degrees) = Longitude degrees + Longitude minute/60.0 + Longitude second/3600.0

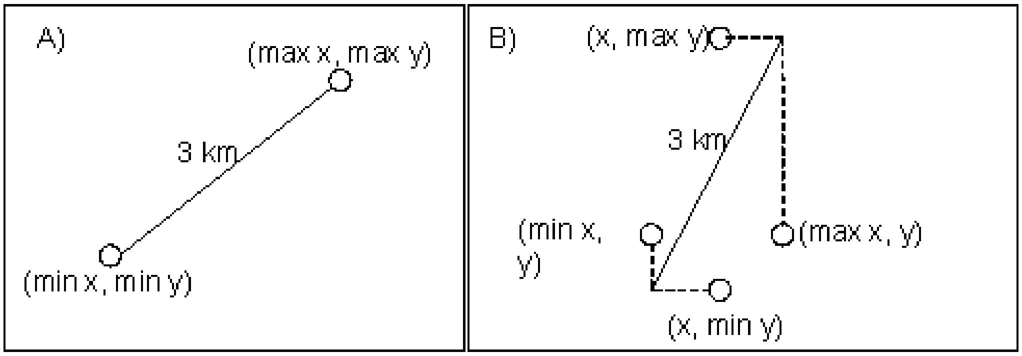
If latitude and longitude are reversed, a common error in the 1999 NEI submittals, we will revise the latitude and longitude.

After converting UTMs to latitude and longitude and correcting latitudes and longitudes reported in wrong units, we will use a routine to assess the validity of the latitude and longitude values, to replace values if necessary, and to fill-in missing data points. The QA of location coordinates is a multi-step process. The first step is to make sure that all of the emission release points within a facility are within a reasonable distance of one another. The second step uses geographic information system (GIS) overlays to evaluate each coordinate pair with respect to its county boundary. The stages of the routine are described below.

Step 1. Find and replace the latitude/longitude of emission release points within a facility that are located at distances greater than 3.0 km of other release points in the facility.
This step includes determining if an individual release point within a facility is within 3.0 km of other emission release points. (i.e., to ensure no stacks within a single facility are located miles apart).

If there is only one emission release point, the process is complete and we proceed to Step 2.

If there is more than one emission release point, we identify the emission release points with the highest and the lowest latitudes and emission release points with the highest and lowest longitudes at the site. We calculate the distances between the highest and lowest latitudes and between the highest and lowest longitudes. Please refer to the diagram below for details on how we conduct this analysis.



This method may but does not necessarily measure the true distance between emission release points. For example, if a facility has 2 emission release points, the true distance is measured by examining minimum and maximum latitudes/longitudes at all emission release points described by Example A in the diagram. If, for example, a facility has 4 emission release points as in Example B of the diagram, the distance calculated does not represent the distance between any two specific emission release points, but rather defines the maximum distance that could exist between any two emission release points given all of the points in the set. This analysis identifies sites whose emission release points are potentially far apart and whose coordinates need correction, but it is not designed to identify the specific outlying emission release points.

A. If the greatest distance between latitudes and between longitudes is less than 3.0 km, the process is complete and we proceed to Step 2.

B. If the distances between the highest and lowest latitudes or the highest and lowest longitudes are greater than 3.0 km, the latitudes and longitudes of all emission release points within a facility are evaluated. If there are records whose latitudes and longitudes are found to be at a distance of more than 3.0 km, the SIC or NAICS codes are examined to determine if the distance between emission release points is technically correct.
   - For a source category, if it is acceptable that the distance between emission release points is greater than 3.0 km, the process is complete and we proceed to Step 2.
   - For a source category, if it is determined that the distance between emission release points should not be more than 3.0 km, then the following steps are conducted.
     1. The distances between all emission release points within the facility is calculated and outlier(s) are identified. For example if four emission release points are present (A, B, C and D), then the distances between the latitudes and longitudes of A & B, A & C, A & D, B & C, B & D, and C & D are calculated.
     2. An average site latitude/longitude is calculated using only the acceptable coordinates. This average site latitude/longitude is then used to replace the inaccurate latitude/longitude values. After this step, the process is complete and all the latitude/longitude data are assumed to be correct. We then proceed to Step 2.

Step 2. Find and replace latitude/longitude of emission release points that are out-of-county boundary or missing.

This step involves the use of a GIS overlay to plot each latitude/longitude value and compare the plot to the physical boundaries of the FIPS county to which the value is associated (i.e., to ensure no stacks are located in the oceans or in far away states). Detailed county boundaries using a scale of 1 to 100,000 or better are used in the GIS overlay.

If the plotted release point is within 10 km of the county, the point is assumed to be valid and neither latitude/longitude nor county FIPS code are corrected. The process is complete.

If the plotted release point is found to exist more than 10 km outside of the county or if the latitude/longitude is missing, then the latitude/longitude of each emission release point is replaced using the following hierarchy. Only one method is used for missing or out-of-boundary latitude/longitude within a facility, i.e., method A is not used for one emission release point at a facility and method D for another emission release point at the same facility.

A. *Use Facility Specific Data* - A check is first completed to see if there are any other emission release points at the facility that exist within the10 km zone. The county FIPS code for the emission release points that are outside the 10 km boundary is compared with county FIPs codes of other release points within the facility that exist within the 10 km boundary. If the other valid emission release points within a facility have a different county FIPS code from the emission release points that are in question, then the county FIPS code is changed and the latitude/longitude of the emission release points rechecked to see if the latitude/longitude of emissions release points in question are now valid.

If any emission release points in the facility exist within the 10 km boundary, an average site location is estimated using the latitude/longitude of the group of emission release points located within the 10 km zone. The county FIPS code of the group of emission release points within the 10 km zone and the average site latitude/longitude are used to replace those latitude/longitude values and county FIPS code found to exist at emission release points outside of the 10 km boundary. The process is complete.

B. *Use Geocoding software* – If none of the reported emission release point latitude/ longitude values exist within the ten kilometer boundary of the county according to the facility's FIPS code, geocoding software is used. More information on Tele Atlas North America "EZ-Locate" geocoding software is located at www.geocode.com/.[2]

The first step in using Geocoding software is to check the quality of the zip code provided in the inventory with a zip code QA file. The zip code file can be found at the following address, www.epa.gov/ttn/emch/invent/. The source of the zip code file is ESRI and the file uses zip code information from the US Postal Service and FIPS code information from the US Department of Commerce, National Institute of Information and Technology. Zip codes and FIPS codes reported in the Site records are compared with zip codes and FIPS codes in the zip code file. If the reported FIPS Code does not match the FIPS Code in zip code QA file, then we would not use the reported Zip Code in the Geocoding process. If Zip Codes are incorrect, we will first use the NEI Standardized Zip Code in the NEI Historical Facility Table to default the Zip Code. If global errors such as dropping leading zeroes for a state or transposition errors exist, then we will correct the Zip Codes. If we are unsure of how to correct errors (e.g., zip code does not match town or latitude/longitude), then we will contact the data providers.

After correcting zip codes, a list of sites that have emission releases points with missing or erroneous latitude/ longitude is compiled. The file contains site name, the physical addresses of the sites, state FIPS code and county FIPS code. The file is submitted to the geocoding software. Using the inventory record's street address, the geocoding software matches using the following hierarchy.

1. First the software standardizes the street address and looks for an exact address match.
2. If an exact match cannot be found, the software then tries to match at the single street block, as defined by Geocoder's documentation. The latitude/longitude are located at the centroid of the single street block.
3. If a single street block match cannot be found, then the software tries to match on the 5-digit zip code plus 2 digits. The latitude/longitude are located at the centroid of the 5 digit zip code plus 2 digits.
4. If a match to 5-digit zip code plus 2 digits cannot be found, then the software tries to match to the 5-digit zip code alone. The latitude/longitude are located at the centroid of the 5-digit zip code.
5. If a match to 5-digit zip code cannot be found, then the software tries to match to 3-digit zip code. The latitude/longitude are located at the centroid of the 3- digit zip code.
6. If a match to a 3-digit zip code cannot be found, then the software provides an "ambiguous" match which is a match to multiple non-standardized street segments.

If the geocoding software finds valid latitude/longitude data, the process is considered to be complete and all latitude/longitude pairs are assumed to be valid. The county FIPS code will be changed if necessary to match the geocoded county FIPS code only after we contact the data provider and verify the county FIPS code. If the data provider indicates the geocoded FIPS code is incorrect, we will not use the geocoded latitude/longitude. If geocoded latitude/longitude are used, the process is considered complete.

C. *Use NEI Historical Facility Table* - If Geocoder cannot find a valid latitude/longitude for the facility, then the NEI Historical Facility Table will be used in the 2002 NEI. If the identified county in NEI Historical Facility Table for a facility does not match the reported county in the inventory, the NEI Historical Facility Table data will not be used. If the NEI Historical Facility Table contains valid latitude/longitude data with corresponding valid county, the process is considered complete and all the latitude/longitude data are assumed to be valid.

D. *Assign Site Release Point at County Centroid* - If, after each of these stages, an emissions release point latitude/longitude data set is still found to be missing or invalid, site the emission release point to the county centroid.

All defaulted latitudes and longitudes will be identified in the NEI database. Default flags are also included for coordinate data in the Emission Release Point record. We will use the defaults for coordinates shown in Table 1.

Latitude/Longitude EPA Data Standard Elements

The EPA's Latitude/Longitude Data Standard consists of the group of data elements used for recording horizontal and vertical coordinates and associated metadata that define a point on earth. Table 2 summarizes these elements. This standard will help users gauge the accuracy and reliability of a given set of coordinates. The primary responsibility for populating these fields lies with the data submitters, as it is difficult if not impossible to discern the origin of a latitude/longitude without being the primary author of the data.

Horizontal Collection Method Code
If the Horizontal Collection Method code is missing, we will populate this field code as "027" (The information is not known) when latitude and longitude coordinates that are not defaulted are retained in the 2002 NEI. When we default latitude and longitude using Geocoding software, we will only be able to populate these fields whenever latitude/longitudes were obtained from the TeleAtlas Geocoding EZ Locator Service (http://geocode.com). Table 3 presents default flags from Geocoder software and default values for the latitude and longitude data standards.

Horizontal Reference Datum Code*, *Horizontal Accuracy Measure
We will only populate these fields with codes in Table 3 when latitude/longitudes are obtained from the TeleAtlas Geocoding EZ Locator Service (*http://geocode.com*). If these fields are not reported by data submitters and we do not default latitude and longitude using Geocoder software, we will not be able to populate these fields.

Coordinate Data Source Code
If the Coordinate Data Source Code is missing, we will populate this data field based on the data source using the codes in the NIF code tables.

Reference Point Code
If the Reference Point Code is missing, we will populate this field code as "108" (Points not represented by 101-107) when latitude and longitude coordinates that are not defaulted are retained in the 2002 NEI. We will populate the Reference Point Code with codes in Table 3 when latitude/longitudes are obtained from the TeleAtlas Geocoding EZ Locator Service (*http://geocode.com*).

Source Map Scale Number
This data field is only applicable when a map has been used to determine latitude and longitude. We will only populate the Source Map Scale Number with codes in Table 3 when latitude/longitudes are obtained from the TeleAtlas Geocoding EZ Locator Service (*http://geocode.com*). If the Source Map Scale Number is not reported by data submitters and we do not default latitude and longitude, we will not be able to populate this field.

STACK PARAMETERS
In preparing emissions for grid modeling, valid parameters for the physical characteristics of each release point (stack height, diameter, temperature, velocity, and flow) are necessary to correctly place facility release points and associated emissions into vertical layers for proper air quality modeling. Gaussian dispersion models need stack parameters to characterize the plume, which is needed to estimate proper concentrations from these models. The first step is to QA the Emission Release Point

Type. After we augment for any invalid or missing Emission Release Point Types, we use a routine to assess the validity of the stack parameters, to replace values if necessary, and to fill-in missing data points.

Emission Release Point Type
The Emission Release Point Type identifies whether emissions are released as a stack or fugitive emissions. The valid NIF codes for the Emission Release Point Type are:

01 - Fugitive,
02 - Vertical,
03 - Horizontal,
04 - Goose Neck,
05 - Vertical with Rain Cap, and
06 - Downward Facing Vent

The Emission Release Point Type is needed to determine how to augment missing or out-of-range stack parameters.
We will evaluate the SCCs of Emission Release Point Types to determine if the Emission Release Point Type is reported correctly as a fugitive or a stack by using the SCC/Emission Release Point Type Crosswalk. If the Emission Release Point Type is found to be inconsistent with the reported SCC, we will revise the Emissions Release Point Type. When we change a non-fugitive(stack) value to fugitive, we will use NIF Code 01 for Emissions Release Point Type. When we change a fugitive value to non-fugitive, we will use NIF Code 02 for a vertical stack.

Stack Parameters QA and Default Procedure
Stack parameters include stack height, stack diameter, exit gas temperature, exit gas velocity, and exit gas flow rate.

We will employ a routine that compares each emission release point parameter to a minimum and maximum range of values and when that parameter is missing or is found to exist outside of that range, we will augment the parameter. We will also check non-fugitive stack parameters for internal consistency between:

- stack height and diameter, and
- stack diameter, exit gas velocity, and exit gas flow rate.

When internal consistency is not met, we will replace the parameters.

The following steps summarize the process of finding and replacing missing, out-of-range, or internally inconsistent stack parameters.

Step 1: <u>For fugitive emission release points, replace stack parameters</u>
For fugitive emission release points, we will first compare the existing height against the following range thought to be representative of the minimum and maximum values allowable for most fugitive emission release points.

Fugitive Release Height:	0.1 to 100 ft

If the height is valid, we will keep the height and replace all other stack parameters with the defaulted values listed below. If the height is invalid, we will replace all stack parameters with the defaulted values.

Stack Height:	10 ft
Stack Temperature:	72 °F
Stack Diameter:	0.003ft
Stack Velocity:	0.0003 ft/sec
Stack Flow:	0 cu ft/sec

Step 2: <u>For non-fugitive emission release points, find out-of-range or missing stack parameters</u>
For non-fugitive emission release points, we will first compare existing stack parameters against a set of the following ranges thought to be representative of the minimum and maximum values allowable for most emission release points.

Stack Height:	0.1 to 1200 ft
Stack Temperature*	50 to 1,800 °F
Stack Diameter	0.1 to 50 ft
Stack Velocity	0.1 to 100 ft/sec

*Stack Temperature must be greater than 250 °F for the following SCCs and MACT Codes.

- SCCs: 10xxxxxx, 20xxxxxx, 501001xx, 501005xx, 502001xx, 502005xx, 503001xx, and 503005xx
- MACT codes: 0105, 0107, 0107-1, 0107-2, 0107-3, 0107-4, 0108, 0801, 0801-1, 0801-2, 0801-3, 0801-4, 1801, 1802, 1808-1, 1808-2, 1808-3

After we identify missing or out-of range parameters, we will evaluate the source category to determine if out-of-range parameters may be plausible. If any parameter is missing or out-of range, the parameter will be replaced using the procedures described in Step 4. If all parameters are found to exist within the bounds of the emission release point ranges, we will proceed to Step 3.

Step 3: <u>For non-fugitive emission release points, find inconsistencies in stack parameters</u>
We will determine any inconsistencies in stack parameters by conducting the following two steps.

A. For stack diameter, we will compare the stack diameter to the stack height. For non-fugitive emission release points, the stack height may not be less than stack diameter.

B. We will determine the internal consistency between diameter, velocity and flow rate using the following equation.

Stack Flow [cu ft/sec] = (Π [Pi] * (Stack Diameter [ft] / 2) ^ 2) * Stack Velocity [ft/sec]

If the calculated flow and the reported flow are within 10 % of one another, then internal consistency is assumed to be valid.

If all parameters are found to exist within the bounds of the emission release point ranges in Step 2, and the consistency checks (A) and (B) in Step 3 are satisfied, no additional steps will be taken. If any parameter is missing or out-of range, or if the parameters fail the internal consistency tests, the parameter will be replaced using the procedures described in Step 4.

Step 4: Replace stack parameters for non-fugitive emission release points
The first step in replacing stack parameters is to determine if there are problems with stack height or diameter. Because stack height and diameter are the physical parameters that are most easily measured or estimated, when there are problems with these parameters, then the entire set of stack parameters are deemed questionable. If either height or diameter is missing or out-of range, or if the stack diameter is greater than stack height, then all 5 parameters will be defaulted using national default sets of physical parameter data. No additional steps are taken once all 5 parameters are defaulted.

If stack height and diameter do not need replacement, then velocity and flow rate are evaluated next. If velocity and flow rate are not internally consistent, we will QA the flow rate to determine if it was reported in cubic feet per minute rather than cubic feet per second as required in the NIF. In the 1999 NEI, we found that several data submitters reported flow rate in cubic feet per minute rather than cubic feet per second. We will correct flow rates reported in cubic feet per minute to cubic feet per second and than evaluate the flow rate and velocity for internal consistency.

If the internal consistency is not met for velocity, flow rate, and diameter, Table 4 provides instructions on how we will replace missing, out-of-range values, or internally inconsistent values for velocity and flow rate based on different reported scenarios. Velocity and flow rate are augmented either by calculation or the use of national defaults.

Finally, in cases where all 5 parameters have not been defaulted, and velocity and flow rate have been evaluated and replaced if necessary, temperature is evaluated. If temperature is missing or out-of-range, then the temperature is defaulted using national default sets of physical parameter data.

National default sets of physical parameter data organized by SCC and SIC codes are used to replace all 5 parameters using the following hierarchy.

1. SCC match
2. facility level SIC code match
3. national default for release points, if no SCC or SIC code match is possible

From Version 3 of the 1999 NEI for HAPs and 1999 NEI for CAPs, we generated default look-up tables by SCC, MACT Code and SIC code to report the average value calculated for stack height, diameter, temperature and velocity for all emission release point types that are coded as stacks. Only valid values were used in the averaging. The following records were removed from the 1999 NEI files prior to preparing SCC, MACT Code, and SIC stack parameter default look-up tables.

- Records with SCCs incorrectly reported as stack
- Records with national defaults
- Records with velocity greater than 100 ft/sec

- Records with temperatures < 250 °F with SCC 10xxxxxx, 20xxxxxx, 501001xx, 501005xx, 502001xx, 502005xx, 503001xx, and 503005xx and MACT codes 0105, 0107, 0107-1, 0107-2, 0107-3, 0107-4, 0108, 0801, 0801-1, 0801-2, 0801-3, 0801-4, 1801, 1802, 1808-1, 1808-2, and 1808-3

The flow rate was calculated by using the average diameter and average velocity and the equation in Step 3. Default stack parameters are available for more than 3,600 SCCs, 125 MACT codes and more than 800 SIC codes. Separate look-up tables are prepared for SCCs, MACT Codes, and SIC Codes. These files can be found at the following address. www.epa.gov/ttn/chief/emch/invent/ .

When an out-of-range parameter value is found for a specific stack, all of the SCCs, MACT Codes and SIC Codes characterizing emissions through that stack are first determined. In the case of multiple SCCs, MACT Codes or SIC codes, a match is done for each in the respective look-up table to find a default replacement value for the out-of-range parameter. When there are multiple default replacement values by SCC /MACT Codes/SIC Code possible for a specific stack, we will use the SCC defaults first unless the reported SCC is 39999999. If no SCC match is possible or a SCC of 39999999 is reported, we will use MACT Code defaults. If no SCC or MACT Code match is possible, we will use SIC Code defaults. If multiple SCCs or MACT Codes or SIC Codes are available for a single Emission Release Point, we will use the default record having the lowest stack height to modify and replace that out-of-range release parameter.

If no SCC, MACT Code, or SIC code match is possible, we will use the following national default values for the stack parameters.

Stack Height:	10 ft
Stack Temperature:	72 °F
Stack Diameter:	1 ft
Stack Velocity:	15 ft/sec
Stack Flow:	12 cu ft/sec

Stack Parameter Default Flag

All defaulted stack parameters will be identified in the Emission Release Point record. We will use the following coding system to identify the source of default stack parameters.

0 = Original value (not a default)
1 = SCC default
2 = SIC code default
3 = National default
4 = Calculated value
5=MACT Code default

A single default field will be used to represent the source of all five stack parameters. The codes will be presented in this field in the following order: stack height, exit gas temperature, stack diameter, exit gas velocity, exit gas flow rate. Thus, the code “00014" indicates that stack height, exit gas temperature, and stack diameter are original values, exit gas velocity is SCC defaults, and exit gas flow rate is calculated based on the stack diameter and exit gas velocity values.

HAP POLLUTANT CODES

Section 112(b) of the CAA contains a list of 188 HAPs. HAPs are generally defined as those pollutants that are known or suspected to cause serious health problems, including cancer. Section 112(b) of the Clean Air Act currently identifies a list of 188 pollutants as HAPs, www.epa.gov/ttn/atw/orig189.html. EPA's ATW web site presents more information on HAPs, their effects, and EPA's programs to reduce HAPs. (www.epa.gov/ttn/atw/basicfac.html)

The NEI includes emissions data for all 188 HAPs. In addition to numerous specific chemical species and compounds, the list of 188 HAPs includes several compound groups (e.g., individual metals and their compounds, polycyclic organic matter (POM), etc). Many of the uses of the NEI depend upon data for individual compounds within these groups rather than aggregated data on each group as a whole. For risk assessments, individual speciated HAPs are needed because the toxicity associated with an individual compound within a compound group varies widely. For modeling files generated from the NEI, we report speciated pollutants using the NIF Code Table. For trends analyses and summary NEI data, we prepare files using the Section 112b HAPs and aggregate speciated NEI pollutants using the HAP Category Name. The HAP Category Name is the same name for HAPs that are not listed as compound groups in Section 112b, for example, formaldehyde. For a pollutant such as methylmercury, the HAP Category Name is mercury and compounds.

We have encouraged data providers to report emissions of individual compounds and to identify HAP compounds by Chemical Abstract Services (CAS) number. We will QA the reported Pollutant Codes. If the reported codes are incorrect, we will contact the data provider for a crosswalk of HAPs reported to CAS Numbers and pollutant names.

The EPA's Chemical Data Standard provides for the use of common identifiers throughout the Agency for all chemical substances regulated or monitored by EPA environmental programs.

This standard provides unique, unambiguous, chemically correct common names for all chemicals substances and groupings in EPA's system and are provided here as they enable users to search for chemical substances across EPA programs and their databases. The Chemical Identification Standards Database is an excellent source of CAS numbers for HAPs. The Chemical Identification Standard consists of the following data elements:

- ***Chemical Abstracts Service Registry Number*** -unique number assigned by CAS to a chemical substance;
- ***Chemical Substance Systematic Name (9th Collective Index Name)*** - the name assigned to a chemical substance that describes it in terms of its molecular composition;
- ***EPA Chemical Identifier*** - unique number assigned when CAS number not available; and
- ***EPA Chemical Registry Name*** - the name EPA has selected as the name to be commonly used by EPA in referring to a chemical substance.

We will compile a Chemical Data Standards Table (Table 5) to meet EPA's Chemical Data Standards. The Chemical Data Standards Table contains each of these elements for every NEI pollutant code and for the corresponding NEI HAP Category.

CONCLUSIONS

QA and data augmentation are key steps in improving the quality of data in the NEI for modeling. Modeling requires the QA and augmentation of key data fields prior to model preprocessing the NEI. In order to improve transparency of data, EPA identifies all data fields that are augmented. For data fields that can have more than one default procedure, EPA identifies the methodology used to augment data fields. EPA recognizes that resources may limit the ability of data providers to provide an inventory with all data fields completed. However, EPA expects data providers to QC data prior to data submission. Although EPA can augment most of the 115 data fields in the point source NEI, EPA prefers to use source specific data from data providers that meet acceptable levels of quality. It is hoped that by understanding the consequences of submitting files with missing or out-of-range parameters, data providers can better plan and prioritize their emissions inventory development activities.

REFERENCES

1. Pope, Anne. "NEI Quality Assurance and Data Augmentation Steps for Point Sources"; U.S. Environmental Protection Agency (EPA). http://www.epa.gov/ttn/chief/emch/invent.

2. Tele Atlas North America "EZ-Locate" geocoding software. http://www.gecode.com

KEY WORDS

Point Source Inventory
National Emissions Inventory
Criteria Pollutant
Hazardous Air Pollutant
QA (Quality Assurance)
Data Augmentation

Table 1. Coordinate Default Flags

Code	Code Description
Exact	Match is to within a unique intersection or within a single side of a single street block.
Near	Match is to a single street block but the correct placement within block is unknown.
Zip code+2	Match to a 5-digit zip code, plus the first two digits of the 4-digit extension.
Zipcode5	Match to a 5-digit zip code
Zipcode3	Match to multiple 3-digit zip codes based on postal service Sectional Center Facility (SCF).
Ambig	Match is to multiple street segments.
Cntycent	County centroid.
NEI Fac Table	Coordinate found in the NEI Historical Facility Table

Code	Code Description
Site-Avg.	Average of accurate coordinates of other emission release points at the same site

Table 2. EPA Latitude/Longitude Data Standards.

Latitude/Longitude Standard	Description	Comments
Latitude Measure	Y Coordinate - The measure of the angular distance on a meridian north or south of the equator.	+78.123456 The number of decimal positions recorded is determined by the precision of the measurement.
Longitude Measure	X Coordinate - The measure of the angular distance on a meridian east or west of the prime meridian.	-123.234561 The number of decimal positions recorded is determined by the precision of the measurement
Source Map Scale Number	The number that represents the proportional distance on the ground for one unit of measure on the map or photo.	Only used when a map has been used to determine latitude/longitude. e.g. 125,000

Latitude/Longitude Standard	Description	Comments
Horizontal Collection Method Code	Method used to determine the latitude and longitude coordinates for a point on the earth.	e.g., 001 = address-matching house number, 018 on interpolation map, 028 = Global Positioning Method, with unspecified parameters.
Horizontal Accuracy Measure	The measure of the accuracy (in meters) of the latitude and longitude coordinates.	
Horizontal Reference Datum Code	The code that represents the reference datum used in determining latitude and longitude coordinates.	001 = North American Datum of 1927 002 = North American Datum of 1983 003 = World Geodetic System of 1984
Reference Point Code	The code that represents the place for which geographic coordinates were established.	e.g. 101 = Entrance point of a facility or station.; 105 = Point where substance is processed, treated, settled, or stored.; 106 = Point where a substance is released.
Coordinate Data Source Code	The code that represents the party responsible for providing the latitude and longitude coordinates	e.g. EPA Headquarters, a state agency, tribal organization, EPA regional office etc.

Table 3. Geocoder Default Flags and Default Values for Latitude/Longitude Data Standard

Code	Description	Source Map Scale	Horizontal Collection Method Code & Description	Horizontal Reference Datum	I
Exact	Match is to within a unique intersection or within a single side of a single street block.	24000	002 - Determination method based on address matching-block face.	001 - North American Datum of 1927	
Near	Match is to a single street block but the correct placement within block is unknown.	24000	003 - Determination method based on address matching-street centerline.	001 - North American Datum of 1927	
Zipcode+2	Match to a 5-digit zip code, plus the first two digits of the 4-digit extension.	24000	038 - Determination method based the center of an area defined by the 5-digit ZIP code and its 2-digit geographic segment extension.	001 - North American Datum of 1927	
Zipcode5	Match to a 5-digit zip code.	24000	026 - Determination method based on zipcode-centroid.	001 - North American Datum of 1927	
Zipcode3	Match is to a 3-digit zip code.	24000	021 - Determination method based on interpolation-other.	001 - North American Datum of 1927	
SCF3	Match to multiple 3-digit zip codes based on postal service Sectional Center Facility (SCF).	24000	021 - Determination method based on interpolation-other.	001 - North American Datum of 1927	
Ambig	Match is to multiple street segments.	24000	007 - Determination method based on address matching-other.	001 - North American Datum of 1927 001	
Countycent	County centroid.	24000	021 - Determination method based on interpolation-other; 030 - based on a digital map source (TIGER).	001 - North American Datum of 1927	

* Coordinates are derived from USPS, Census Bureau Tiger server, or Eagle's TeleAtlas. These correspond to codes 080 (government agency) and 084 (contracting organization).

Table 4. Stack Parameter Data Replacement Matrix (X = Data value present)

Diameter	Velocity	Flow Rate	Action
X	X	X	1. Check that velocity is within range. A. If velocity is within range and flow rate does not meet internal consistency for diameter, velocity and flow rate, then: ▸ Calculate flow rate using internal consistency formula. B If velocity is not within range, then: ▸ Calculate velocity using internal consistency formula. ▸ Check that calculated velocity is within range. If so, then default to calculated velocity. ▸ If calculated velocity is not within range, then default all 5 parameters using national default set.
X	-	X	1. Calculate velocity using internal consistency formula. 2. Check that calculated velocity is within range. A. If calculated velocity is not within range, then: ▸ Default all 5 parameters using national default sets.
X	X	-	1. Check that velocity is within range. A. If velocity is within range, then: ▸ Calculate flow rate using internal consistency formula. B. If velocity is not within range, then: ▸ Default all 5 parameters using national default sets.
X	-	-	1. Default velocity using national default sets. 2. Calculate flow rate using internal consistency formula.
-	X	X	1. Default all 5 parameters using national default sets.

Table 5 NEI Chemical Data Standards Table

Field Name	Description
NEI Pollutant Code	Unique code assigned to NEI pollutant
NEI Pollutant Name	HAP name for NEI pollutant
NEI HAP Category	Grouping of related NEI pollutants
CASRN	Chemical Abstracts Service Registry Number
CASRN_compact	Chemical Abstracts Service Registry Number without dashes
ChemSystematicName	Chemical Substance Systematic Name (9th Collective Index Name)
EPAChemicalID	EPA Chemical Identifier
EPAChemRegistryName	EPA Chemical Registry Name
HAPCAT_CASRN	Chemical Abstracts Service Registry Number (assigned to HAP category)
HAPCAT_CASRN_compact	Chemical Abstracts Service Registry Number without dashes (assigned to HAP category)
HAPCAT_ChemSystematic Name	Chemical Substance Systematic Name (9th Collective Index Name)(assigned to HAP category)
HAPCAT_EPAChemcialID	EPA Chemical Identifier (assigned to HAP category)
HAPCAT_EPAChem RegistryName	EPA Chemical Registry Name (assigned to HAP category)