Introduction to Data Validation

Hilary Hafner
Sonoma Technology, Inc.
Petaluma, CA
for
National Ambient Air Monitoring Conference
St. Louis, MO
August 8, 2016
VOC and PM Speciation Data

• Differences from one measurement (such as ozone or PM mass)
  – More complex instruments (more to go wrong?)
  – Many species per sample
  – Data overload

• Opportunity for intercomparison
Why You Should Validate Your Data (1)

- It is the monitoring agency’s responsibility to prevent, identify, correct, and define the consequences of monitoring difficulties that might affect the precision and accuracy, and/or the validity, of the measurements.
- Serious errors in data analysis and modeling (and subsequent policy development) can be caused by erroneous data values.
- Accurate information helps you respond to community concerns.

Validate data as soon after collection as practical – it reduces effort and minimizes data loss.
Why You Should Validate Your Data (2)

• Criteria pollutant data quality issues are important to national air quality management actions, including
  – Attainment/nonattainment designations
  – Clean data determinations
  – Petitions to EPA for reconsideration

• Air quality data are very closely reviewed by stakeholders
  – Do data collection efforts meet all CFR requirements?
  – Have procedures outlined in the QA handbook or project-specific QA plans been followed?
  – Are agency logbooks complete and up-to-date?

• Deviations are subject to potential litigation

Data Validation Process Changes

- More data being collected
- New instruments
- Better computing
- Better tools (e.g., visualization)
- Improved data handling and access allow for more frequent review

Provides ability to assemble data and metadata all in one place and allows a more efficient validation and review process.
Data Validation Levels

- **Level 0 – Routine checks**
  - Field and laboratory operations, data processing, reporting conducted in accordance with SOPs
  - Proper data file identification; review of unusual events, field data sheets, and result reports; instrument performance checks

- **Level I – Internal consistency tests**
  - Identify values that appear atypical when compared to values of the entire dataset

- **Level II/III – External consistency tests**
  - Identify values in the data that appear atypical when compared to other datasets
  - Continued evaluation of the data as part of the data interpretation process

Sidebar: Outliers

• Definition: a value that lies outside most of the other values in a set of data.

• Identification: statistically, ideas include
  – >95th percentile (from exceptional event documentation)
  – 3 to 4 standard deviations above the mean

• Treatment: valid/suspect until proven invalid...

“The first assumption upon finding a measurement that is inconsistent with physical expectations is that the unusual value is due to a measurement error. If, upon tracing the path of the measurement, nothing unusual is found, the value can be assumed to be a valid result of an environmental cause.”

*Judy Chow, Desert Research Institute*
General Approach to Data Validation

• Look at and manipulate your data—sort it, graph it, map it—so that it begins to tell a story.

• Examples
  – Scatter, time-series, and fingerprint plots
  – Summary statistics
  – Box-whisker plots
  – Wind, pollution roses

• Important issues or errors with data may become apparent only after someone begins to use the data for something
Approach/Tips

• Apply screening criteria to help focus validation efforts
• Inspect every species, even to confirm expectation that the species would normally be below the method detection limit
• Apply flags to data
• Document changes

Proceed from the big picture to the details
Considerations in Evaluating Your Data

- Levels of other pollutants
- Time of day/year
- Observations at other sites
- Audits and inter-laboratory comparisons
- Instrument performance history
- Calibration drift
- Site characteristics
- Meteorology
- Exceptional events
Screening Criteria

- Range
- Sticking
- Buddy site
- Temporal consistency
- Rate of change or spike
- Abundant species
- Chemical consistency
- Co-pollutants

Automated checks are helpful to focus efforts on the data that need the most attention.
## PAMS Auto-Validation: Screening

<table>
<thead>
<tr>
<th>Check</th>
<th>Fails If ...</th>
<th>DART Smarts Action If Check Fails</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abundant Species</td>
<td>Any of Benzene, Propane, N-Butane, Isoprene, N-Hexane, Ethylbenzene are missing or 0</td>
<td>If two or more species missing or =0, flag sample with code “AQ”</td>
</tr>
</tbody>
</table>
| TNMOC                  | - TNMOC is missing or 0; or - Unidentified exceeds 50% of TNMOC; or - Sum of PAMS exceeds TNMOC | - Flag TNMOC and unidentified with code “AN”  
- Flag Unidentified with code “DA”  
- Flag TNMOC and Sum of PAMS with code “DA” |
| Variability            | Species concentration exceeds the mean + 4x standard deviation                | None                                                                                             |
| Sticking               | Species has same non-zero value for 3 or more consecutive samples            | Flag species with code “DA”                                                                      |
| Benzene : Toluene      | Benzene exceeds 0.2 and exceeds Toluene                                      | Flag Benzene and Toluene with code “DA”                                                          |
| Ethylene : Ethane      | Ethylene exceeds 0.5 and exceeds Ethane                                      | Flag Ethylene and Ethane with code “DA”                                                          |
| Propylene : Propane    | Propylene exceeds 0.5 and exceeds Propane                                    | Flag Propylene and Propane with code “DA”                                                         |

^ All checks done in ppbC. AQ = collection error; AN = machine malfunction; DA = aberrant data; TNMOC = total nonmethane organic compounds
PM$_{2.5}$ concentrations (µg/m$^3$) gradually increased over a period of days, but there were no known local major PM sources or regional build-up expected to affect the site. PM concentrations were not high enough to trigger auto-QC checks. The agency responsible for the monitor noted “a communication error between [the monitor] and the data logger.”
Visual Data Review: Time Series

• Look for
  – Jumps, dips
  – Periodicity of peaks
  – Calibration gas, carryover
  – Expected diurnal pattern
  – Expected relationships
  – High concentrations of less abundant species or low concentrations of more abundant species
Visual Data Review: Time Series

- Stuck values
- Wild swings
- Data gap

Graph showing PM$_{10}$ (µg/m$^3$) over time from 6/25/2014 to 7/10/2014.
Visual Data Review: Scatter Plots

SCATTER PLOT

Data Range

12/09/2011 00 to 12/10/2011 00
Visual Data Review: Fingerprint Plots

“Typical” VOC Fingerprint
"Typical" and Precision Test VOC Fingerprints
“Typical,” Precision Test, and Zero Air VOC Fingerprints
Putting Your Data in Perspective

- National averages
- Trends over time
- Comparison to nearby sites, similar areas
- Detection limits

A statistically significant decreasing benzene trend

\[ y = -0.16x + 314.62 \]

\[ R^2 = 0.90 \]
Using the Validated Data

- Health effects research
- Model validation
- Emissions inventory evaluation
- Trends analysis
- Control strategy development and effectiveness
- Supporting other programs (e.g., air toxics)
- Comparisons to other similar cities/areas
What’s Coming Up Next?

• This session
  – UC Davis Data Validation Procedures
  – DART for CSN and PAMS Data Validation

• Wednesday
  – PAMS Session
Contact

Hilary Hafner
Manager, Environmental Data Analysis
hilary@sonomatech.com
707.665.9900