Level 2 Data Validation: Whoops! What the Grass Roots Level Missed

Prepared by:
Hilary R. Hafner
Sonoma Technology, Inc.
Petaluma, CA

Kevin Cavender
U.S. EPA OAQPS

Presented to:
2006 National Air Monitoring Conference
Las Vegas, NV
November 6-9, 2006
Introduction and Overview

- Importance of Data Validation
- Data Validation Levels
- General Approach to Data Validation
- Examples
- Resources
Why Should You Validate Your Data?

- 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.
Examples of Problems in Criteria Pollutant Databases (and Validation Actions)

- Air quality data reported during calibration runs. For example, ozone data with values of 0 ppb (or the calibration gas level) reported during hours when instruments are known to be automatically calibrated. *Data were flagged as calibration.*

- Nitrogen oxides data found to have a constant offset based on comparisons of $\text{NO}_x$ to NO+NO$_2$. *Data were adjusted.*

- Ozone concentrations “capped” at 100 ppb. Investigation showed that the instrument maximum concentration setting was incorrect. *Data at 100 ppb were flagged as suspect low, and the instrument settings were adjusted.*
Data Validation Levels: Summary of Types of Checks

- Level I
  - Routine checks during the initial data processing and generation of data including proper data file identification; review of unusual events, field data sheets, and result reports; and instrument performance checks.

- Level II
  - Internal consistency tests to identify values in the data that appear atypical when compared to values from the entire data set.
  - Comparisons of current data with historical data to verify consistency over time.
  - Parallel consistency tests with data sets from the same population (e.g., region, period of time, air mass) to identify systematic bias.

U.S. EPA, 1999a
Level II: Internal Consistency Checks

- Inspect time series. Are concentrations consistent with time of day, day of week, and season?
- Compare pollutant concentrations. Are expected relationships observed?
- Identify and flag unusual values including
  - Values that normally follow a qualitatively predictable spatial or temporal pattern
  - Values that normally track the values of other variables in a time series
  - Extreme values, outliers

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.

Chow et al., 1996
Level II+: Comparisons to Other Data Sets

- Compare collocated measurements.
- Compare relationships (e.g., temporal, among species) observed in the current data set to relationships observed at other sites or in previous years.
- Compare pollutant concentrations to meteorology.
General Approach to Data Validation

- Look at your data.
- Manipulate your data—sort it, graph it, map it—so that it begins to tell a story.
- Often, important issues or errors with the data will become apparent only after someone begins to use the data for something.

Examples
- Scatter plots
- Time series plots
- Fingerprint plots
- Box whisker plots
- Summary statistics
Example Validation Steps

- Assemble the database.
- Place data in a common data format with descriptive information concerning variables, validation level, QC codes, detection limits, time standard, standard units, and metadata (site information, etc.).
- Ensure that results of and suggestions from all audit reports have been incorporated into the database.
Example Validation Steps  (Page 2 of 2)

- Review summary statistics for unrealistic maxima or minima and for consistency with nearby stations (data are still Level I).
- Perform spatial and temporal comparisons of the data (begin Level II).
- Perform intercomparisons of the data (e.g., from two different instruments).
As a part of validation, it is useful to prepare a summary of the monitoring network by year: summarize which sites have data and how much data for which years.

Use this summary to detect potential problems and to determine what types of analyses are possible.
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
Singling Out Odd Data

- Range checks: minimum and maximum concentrations
- Temporal consistency checks: maximum hour test
- Rate of change or spike check
- Buddy site check: comparison to nearby sites
- Sticking check: consecutive equal data values
Example Screening Criteria – Ozone

- Checks:
  - Are often site-specific
  - May be hour-specific
  - May be automated

- But, data should be graphically reviewed!

<table>
<thead>
<tr>
<th>Check</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>~170 to 225 ppb</td>
</tr>
<tr>
<td>Minimum</td>
<td>–5 ppb</td>
</tr>
<tr>
<td>Rate of change</td>
<td>&gt;50 to 60 ppb/hr</td>
</tr>
<tr>
<td>Buddy sites</td>
<td>±50 ppb up to 5 sites</td>
</tr>
<tr>
<td>Sticking check</td>
<td>@≥40 ppb for 5 hours</td>
</tr>
<tr>
<td>Co-pollutant</td>
<td>NO, NO_x</td>
</tr>
</tbody>
</table>

### Example – Ozone Screening

#### AIRNowTech

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Buddy Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ozone</td>
<td>No buddy sites selected</td>
</tr>
</tbody>
</table>

#### Quality Control Criteria

<table>
<thead>
<tr>
<th>Hour [LST]</th>
<th>Max Suspect</th>
<th>Max Severe</th>
<th>Rate Of Change</th>
<th># of Buddy Sites</th>
<th>Buddy Average</th>
<th># of Sticking Hours</th>
<th>Sticking Value</th>
<th>Minimum Drift</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td>110</td>
<td>169</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>0100</td>
<td>110</td>
<td>169</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>0200</td>
<td>110</td>
<td>169</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>0300</td>
<td>100</td>
<td>168</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>0400</td>
<td>100</td>
<td>168</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>0500</td>
<td>100</td>
<td>168</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>0600</td>
<td>100</td>
<td>168</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>0700</td>
<td>100</td>
<td>168</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>0800</td>
<td>100</td>
<td>168</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>0900</td>
<td>100</td>
<td>168</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>168</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>1100</td>
<td>140</td>
<td>210</td>
<td>60</td>
<td>3</td>
<td>60</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>1200</td>
<td>140</td>
<td>210</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>1300</td>
<td>140</td>
<td>210</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>1400</td>
<td>140</td>
<td>210</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>1500</td>
<td>140</td>
<td>210</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>1600</td>
<td>140</td>
<td>210</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>1700</td>
<td>140</td>
<td>210</td>
<td>60</td>
<td>3</td>
<td>50</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
<tr>
<td>1800</td>
<td>140</td>
<td>210</td>
<td>60</td>
<td>3</td>
<td>60</td>
<td>5</td>
<td>40</td>
<td>-5</td>
</tr>
</tbody>
</table>

**Note hour-specific screening**

Max Suspect: Still used in spatial mapping
Max Severe: Not used in maps
Compared to “buddy” sites, this site had an $[O_3]$ difference of more than 50 ppb. Note, however, that data are valid.
Example – Ozone Validation

1 year of hourly ozone data

Be aware of potential shifts in baseline concentrations. These lower limits in the measurements may be real or artificial, may display patterns, and can definitely affect analyses.
Example Screening Criteria – NO/NO$_x$/NO$_y$

- Checks:
  - Select buddy check criteria
- Collocated ozone can be used to assess NO, NO$_x$, NO$_y$
- Checks may vary depending upon instrument sensitivity

<table>
<thead>
<tr>
<th>Check</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>&gt;700 ppb urban</td>
</tr>
<tr>
<td></td>
<td>&gt;300 ppb rural</td>
</tr>
<tr>
<td>Minimum</td>
<td>–1 ppb</td>
</tr>
<tr>
<td>Rate of change</td>
<td>&gt;30 ppb/hr</td>
</tr>
<tr>
<td>Sticking check</td>
<td>5 hours</td>
</tr>
<tr>
<td>Co-pollutant</td>
<td>NO should not exceed NO$_x$ or NO$_y$</td>
</tr>
</tbody>
</table>

Hafner et al., 2003
Example – Odd Patterns

Zero calibration hour data show up every day at 0300 hrs.
Note period of lower NO\textsubscript{x}, NO concentrations. Is this real or an instrument problem?

Other pollutants are useful in identifying periods of different behavior.
Example – Ozone, NO\(_x\), NO

Likely real “blip” in NO and NO\(_x\) at Camp Logan, IL
Example Screening Criteria – CO

Checks:
- Select buddy check criteria

Checks may vary depending upon instrument sensitivity

<table>
<thead>
<tr>
<th>Check</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>&gt;15 ppm</td>
</tr>
<tr>
<td>Minimum</td>
<td>-1 ppm</td>
</tr>
<tr>
<td>Rate of change</td>
<td>&gt;10 ppm/hr</td>
</tr>
<tr>
<td>Sticking check</td>
<td>&gt; 0 ppm for 5 hours</td>
</tr>
<tr>
<td>Co-pollutant</td>
<td>NO, acetylene</td>
</tr>
</tbody>
</table>
PAMS VOC data at Camden, New Jersey, were compared to CO data for all hours and for selected hours. Some of the midday CO concentrations should be investigated further.
Example – CO Baseline

- In this example, CO data were collected with a deliberate offset of about 5 ppm so that changes in baseline concentrations could be observed.

- Note the lower baseline concentrations and diurnal variation for the indicated time period. The changes were a result of no air conditioning in the instrument shelter.
Example Screening Criteria – SO₂

- Checks:
  - Select buddy check criteria
- Checks may vary depending upon instrument sensitivity
- Regional issues
  - Rural/urban differences
  - Southeast vs. west

<table>
<thead>
<tr>
<th>Check</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>400 µg/m³ (or 150 ppb)</td>
</tr>
<tr>
<td>Minimum</td>
<td>-5 µg/m³ (or -2 ppb)</td>
</tr>
<tr>
<td>Rate of change</td>
<td>&gt;100 µg/m³/hr (or 40 ppb/hr)</td>
</tr>
<tr>
<td>Sticking check</td>
<td>&gt;0 for 5 hours</td>
</tr>
<tr>
<td>Co-pollutant</td>
<td>NOₓ</td>
</tr>
</tbody>
</table>
SO$_2$ Baseline Check

1 year of hourly data

Note shifts in baseline concentrations.
Typical Pollutant Groupings Used in Validation

- Ozone, NO, and NO$_x$ or NO$_y$
  - Also useful: particle scattering
- CO, NO, and NO$_x$
  - Also useful: TNMOC
- SO$_2$ and NO$_x$
  - Also useful: continuous mass or sulfate

TNMOC = total nonmethane organic compounds
Example Baseline Investigation Criteria

Example criteria for investigating baseline changes (especially important to exposure assessments)

- Use 1 year of hourly data
- Use the following data ranges
  - Ozone: 0 to 10 ppb
  - NO₂: 0 to 15 ppb
  - CO: 0 to 2 ppm
  - SO₂: 0 to 10 ppb
- Look for either step functions or gradual drift in the baseline resulting from improper maintenance, post-processing of the data, etc.

Alcorn, 2003
Example Screening Criteria – 1-hr PM$_{2.5}$ Mass

- **Checks**
  - Are often site-specific
  - May be hour-specific
  - May be automated

- **But, data should be graphically reviewed!**

<table>
<thead>
<tr>
<th>Check</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>$&gt;200$ $\mu$g/m$^3$</td>
</tr>
<tr>
<td>Minimum</td>
<td>$-5$ $\mu$g/m$^3$</td>
</tr>
<tr>
<td>Rate of change</td>
<td>$&gt;50$ $\mu$g/m$^3$/hr</td>
</tr>
<tr>
<td>Buddy Sites</td>
<td>± $50$ $\mu$g/m$^3$ up to 5 sites</td>
</tr>
<tr>
<td>Sticking check</td>
<td>$&gt;50$ $\mu$g/m$^3$ for 5 hours</td>
</tr>
<tr>
<td>Co-pollutant</td>
<td>PM$_{10}$</td>
</tr>
</tbody>
</table>

U.S. EPA, 1999a
Example – Erroneous Data in Tennessee
Examples – Wildfire Events
Los Angeles continuous PM$_{2.5}$ mass concentrations on 10/24/03 to 10/27/03 (raw data – USEPA AirNow)

High concentrations in the eastern part of the basin.
High concentrations are consistent with wildfire smoke as shown on this satellite photo from http://rapidfire.sci.gsfc.nasa.gov/gallery/
Example – PM$_{10}$ Baseline Changes

1 year 24-hr average PM$_{10}$ data

Note baseline concentration changes relative to site visits. Were data truncated? Was the instrument baseline physically changed?

Operator visited; adjusted the instrument.
Example – Odd Low Concentrations

Columbus, Ohio
2000

Possible interference from moisture on the TEOM
Speciated PM$_{2.5}$

Internal Consistency Checks (1 of 2)

- Check sum of chemical species versus PM$_{2.5}$ mass (multi-elements Al to U + sulfate + nitrate + ammonium ions + OC + EC)
- Check physical and chemical consistency (sulfate vs. total sulfur, soluble potassium versus total potassium, soluble chloride vs. chlorine, $b_{abs}$ versus elemental carbon)
- Balance charge (cations and anions)
- Balance ammonium

Chow, 1998
Speciated PM$_{2.5}$
Internal Consistency Checks (2 of 2)

- Investigate nitrate volatilization and adsorption of gaseous organic carbon (compare front and backup filter concentrations).
- Prepare crude mass balance (sum of geologic mass, combustion-related mass, and sulfate)
- Compare to collocated or near-collocated FRM mass.
## PM Consistency Checks and Expectations

<table>
<thead>
<tr>
<th>Consistency Check</th>
<th>Expectation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference between PM$<em>{10}$ and PM$</em>{2.5}$*</td>
<td>PM$<em>{2.5}$ ≤ PM$</em>{10}$</td>
</tr>
<tr>
<td>Sum of individual chemical species and PM$_{2.5}$</td>
<td>species sum &lt; PM$_{2.5}$</td>
</tr>
<tr>
<td>Ratio of water-soluble sulfate by IC to total sulfur by XRF</td>
<td>~ 3</td>
</tr>
<tr>
<td>Ratio of chloride by IC to chlorine by XRF</td>
<td>&lt; 1</td>
</tr>
<tr>
<td>Ratio of water-soluble potassium by AAS to total potassium by XRF</td>
<td>&lt; 1</td>
</tr>
<tr>
<td>$b_{abs}$ compared to elemental carbon</td>
<td>good correlation</td>
</tr>
</tbody>
</table>

IC = ion chromatography  
XRF = energy dispersive X-ray fluorescence  
AAS = atomic absorption spectrophotometry  

* Dichotomous data may be an exception to this check  
Chow, 1998
PAMS Data Validation
Example Screening Criteria

- Check that abundant hydrocarbons (e.g., acetylene, ethane, propane, n-butane, i-pentane, n-pentane, n-hexane, benzene, toluene, and m-&p-xylenes) are present in the same samples.
  - This check helps identify “missing” abundant hydrocarbons.
  - Set the screening concentrations sufficiently higher than the detection limit (e.g., 10 times) to limit the number of data “failing” these criteria.

- Check that the data meet expected relationships.
  - For example, n-pentane concentrations are usually less than i-pentane concentrations.
  - Other possible screens include o-xylene < m-&p-xylenes and benzene < toluene.

- Check for unusual sample compositions including
  - ethane concentration < 2 ppbC but benzene > 2 ppbC (may indicate cold trap problems in auto-GC)
  - unidentified fraction of TNMOC > 50% (the less known about a sample’s composition, the less useful the sample).
Typically, urban concentrations of 2,3,4-trimethylpentane (y-axis) and toluene (x-axis) correlate well with each other. In a few samples at this site, the toluene appears to have been misidentified.

Possible misidentified samples
N-undecane shows evidence of the slow removal of the compound after calibration. Six or seven more hours of undecane data need to be invalidated.

Decane and 1,2,3-trimethylbenzene do not show carry-over.

Calibration data removed.
Another Carryover Example

Many of these higher concentrations appear to be from calibration gas carryover
Typically, toluene concentrations are higher than benzene concentrations. These data at Gary, Indiana, are real though because there is a nearby source that emits benzene (but not toluene).
In this plot of 1-hr data, the coke oven is evident to the north of the site.
Air Toxics: Characterizing Spikes – Tetrachloroethylene

• Two sites on the Gulf of Mexico have spikes in tetrachloroethylene concentrations in 1997 and 1998.
• Are these spikes real or measurement artifacts?
• Should these data be used to understand trends or health risks?

Annual averages of tetrachloroethylene
• Time series of seasonal average concentrations of tetrachloroethylene at two sites in Texas.
• Both sites have a single spike in concentration that is at least ten times the typical seasonal average.
Tetrachloroethylene Spikes (2 of 4)

- The Brazoria County spike on June 9, 1997, was more than three orders of magnitude larger than any other in the site’s history.
- The Beaumont site had four years with concentrations at 5 μg/m^3 before decreasing in 1997. Three spikes in 1998 were larger than 50 μg/m^3. After that, concentrations were usually below 0.1 μg/m^3.
• The Brazoria County site is located within 1 km of a Dow Chemical Company industrial facility.

• The TRI documents that this site has fugitive releases of tetrachloroethylene.

• 17,000 pounds of tetrachloroethylene were released in 1997.
• The Beaumont site is located within 2 miles of liquid storage tanks to the north, east, and south.

• The NEI documents oil-tanking operations 3.5 miles from the monitoring station.

• The TRI showed that the Exxon facility two miles to the north of the monitor (circled) released 9,000 pounds of tetrachloroethylene in 1998.

Note date of imagery, over 13 years old
The Brazoria County seasonal spike was an anomaly

- No other measurement at that site was nearly that high.
- No other co-measured species were abnormally high.
- Rough calculations indicate the nearby source could account for a spike of that magnitude (3000 pounds).
- However, sampling or analytical error may also explain the spike.
- TCEQ does not have readily available records of upsets or emissions older than five years to provide additional information (per Dick Flannery, TCEQ).
- This spike should not be used for understanding trends, because it is completely atypical for this site.
Spikes Summary (2 of 3)

- The Beaumont seasonal spike is not anomalous
  - Three measurements were ten times normal concentrations. However, only tetrachloroethylene concentrations were high; no other co-measured species were abnormally high.
  - Concentrations at the site were elevated relative to typical levels from 1994-1997.
  - This seasonal average concentration should be used for both risk and trend assessment, since it likely reflects real concentrations.
Characterizing spikes requires significant work

- Identifying the spikes is straightforward using visual plots of the data (e.g., maps or time series).
- Spikes caused by analytical or sampling error may have anomalous concentrations of other species.
- Real spikes in ambient concentrations are likely due to nearby point sources.
- A combination of maps, the TRI, and local knowledge is likely required (but may not be sufficient) to explain spikes in ambient concentrations.
- Fugitive emission/upsets data are needed!
Visualization Is Key!

Carbon Tetrachloride – annual averages circa 2004
For pollutant data validation,

- Understand formation, emissions, and transport
- Establish screening criteria to identify potentially suspect data
  - Investigate suspect data
  - Invalidate data only if there is sufficient evidence

Data validation is very important!
Resources

- Operator knowledge
- Previous documentation for the site and past data validation results
- EPA guidance documents (available on AMTIC web site)
- Workbooks (e.g., PAMS and PM$_{2.5}$ Data Analysis Workbooks)
- Web sites (e.g., IMPROVE, EPA Supersite)
- Journal articles and conference presentations (e.g., *Atmospheric Environment, Environmental Science and Technology, Air and Waste Management Association*)
- Academia
Key Internet Sites

- Ambient Monitoring Technology Information Center: http://www.epa.gov/ttn/amtic/
- IMPROVE QA/QC: http://vista.cira.colostate.edu/improve/Data/QA_QC/qa_qc_Branch.htm
- EPA Quality Assurance: http://www.epa.gov/oar/oaqps/qa/index.html#back
- PAMS Data Analysis Workbook (old): http://www.epa.gov/oar/oaqps/pams/analysis/
- EPA Supersite Overview: http://www.epa.gov/ttn/amtic/supersites.html