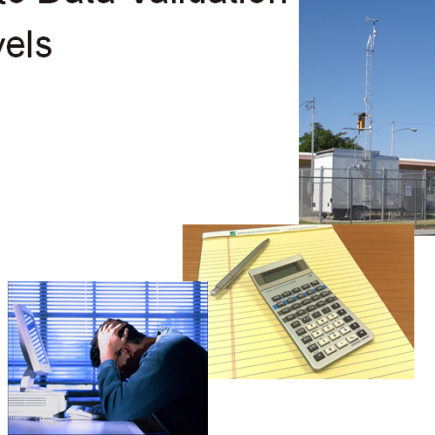


Session 1: Introduction and Overview

- I. Data Validation Overview
- II. General Approach to Data Validation
- III. Data Validation Levels
- IV. Examples
- V. Resources
- VI. Key Internet Sites

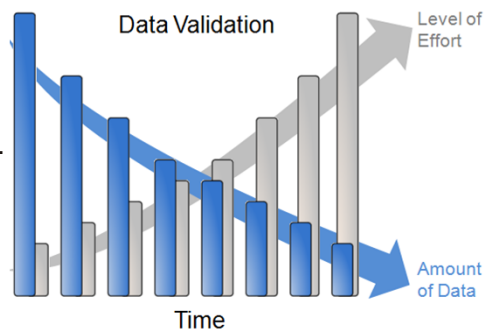
October 2011

STI
Sonoma Technology, Inc.



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.



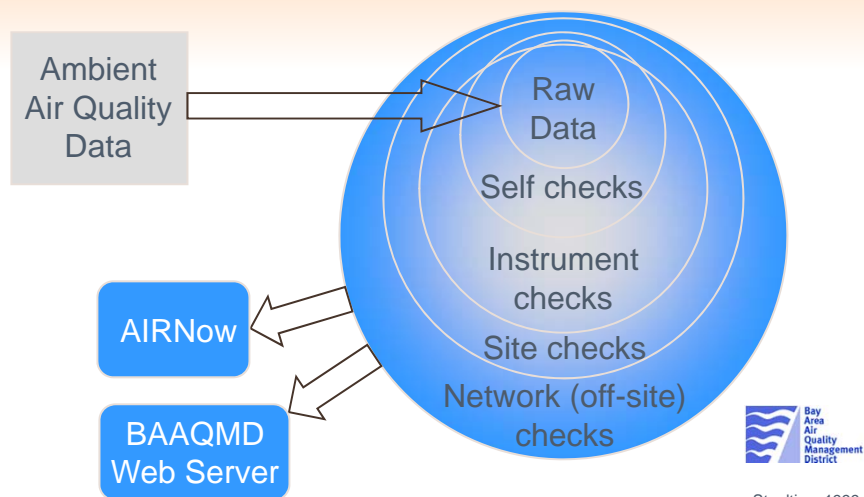
How Has the Data Validation Process Changed?

- More data being collected
- New instruments
- Better computing
- Better tools (e.g., visualization)
- Improved communication (allows remote access and frequent review)



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

Automated Quality Assurance Checks



Stoelting, 1999

Data: Analog vs. Digital

- Instrument data are generally collected in one of two forms: digital or analog.
- Capturing digital output from the monitors is necessary to poll and store all of the diagnostic data associated with the measurements.
- Digital output are useful for many of the data validation procedures. Instrument diagnostic data should be monitored in near real time to increase data quality and data capture.
- Digital output, particularly for the trace gas instruments, is encouraged.



NCore Trace Gas Instruments

- Required trace-level monitors operate on lower ranges, so calibration and audit points must reach lower concentrations
 - Old CO range 0 – 50 ppm for NAAQS compliance
 - Typical NCore CO range 0 – 5 ppm for urban sites and 0 – 2 ppm for rural
- More frequent quality control (QC) checks are highly advised, so calibration automation is important
- Integration of data systems and calibrators is important for QC validation and timely reporting of problems
- Zero air purity is more critical than ever



Data Review, Verification, & Validation

- Data review, verification and validation are techniques used to accept, reject, or qualify data in an objective and consistent manner
- **Verification** can be defined as confirmation, through provision of objective evidence, that *specified requirements* have been fulfilled
- **Review and validation** can be defined as confirmation, through provision of objective evidence, that the particular requirements for a specific *intended use* are fulfilled

U.S. EPA, 1984

Data Verification

- Data verification is checking that specific requirements, such as those specified in an agency's quality assurance project plans (QAPPs) and standard operating procedures (SOPs) have been fulfilled; may be done as part of routine audits
- Includes
 - Verifying that timestamps are correct, and operations occurred at correct times;
 - Leak/flow checks were acceptable; and
 - Manual changes to operations/data were logged and appropriately flagged.



U.S. EPA, 1984

Data Review

The flow of data from the field environmental data operations to storage in a database requires several distinct and separate steps:

- initial selection of hardware and software for the acquisition, storage, retrieval, and transmittal of data
- organization and the control of the data flow from the field sites and the analytical laboratory
- input and validation of the data
- manipulation, analysis, and archiving of the data
- submittal of the data into the EPA's Air Quality System (AQS) database

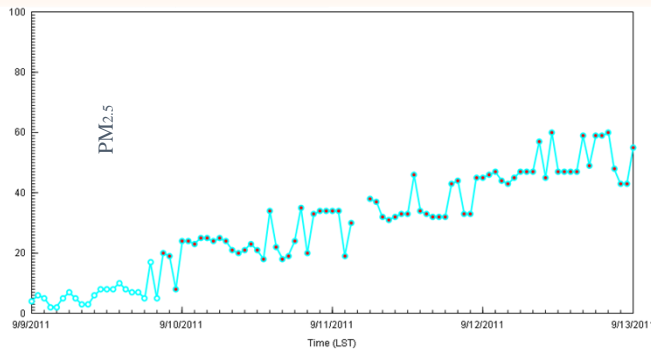
U.S. EPA, 1984

Data Review and Validation

- Data review includes
 - Checking and correcting baseline drift;
 - Identifying extreme min/max values;
 - Performing sticking checks; and
 - Ensuring QC flags from field operators are carried into the final dataset.
- Data validation includes
 - Inspecting data for instances not likely to be caught by automatic checks or instrument flags;
 - Ensuring that automatic checks *correctly* identified suspect/invalid data; and
 - Visually reviewing data to catch instances that are easy for a human eye to see, but not a program.

Data Validation: Human Eyes Needed!

$PM_{2.5}$ concentrations ($\mu\text{g}/\text{m}^3$) were gradually increasing over a period of days, but there were no known local major PM sources expected to affect the site.



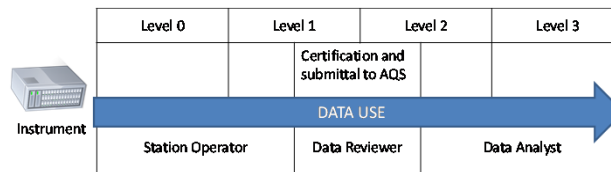
The PM concentration was not high enough to trigger auto-QC checks though, so data had to be manually invalidated. According to the agency responsible for the monitor, there was “a communication error between [the monitor] and the data logger.”

Data Validation Process

- Data review and validation is an ongoing process that is performed by the station operators (SOs) and data reviewers (DRs).
- At a minimum, a cursory review of data should be performed daily, preferably in the morning, to provide a status of the data and instrument performance.
- Detailed analysis is facilitated by access to the logbook entries, notes, and calibration information that the SOs provide for the data review and processing team for their review of the raw data.

Data Validation Process

- Once data have been verified, data are typically entered into EPA's AQS. Data analysts (DAs) typically review data retrieved from AQS.
- Generally, the best approach is to perform low-level data validation procedures on a very frequent basis (or automatically), and then apply more detailed validation procedures to a block of data, such as all data collected during a month.

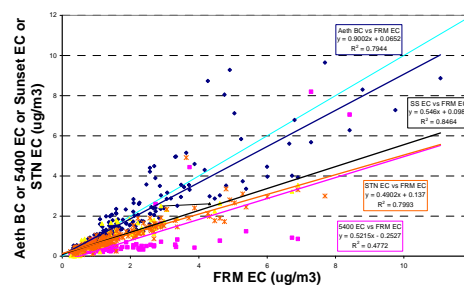


General Approach to Data Validation

- 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



Data Validation Levels: Summary of Types of Checks

- Level 0
 - Routine checks that the field and laboratory operations were conducted in accordance with the SOPs, and the initial data processing and reporting were performed in accordance with the SOP (includes proper data file identification; review of unusual events, field data sheets, and result reports; instrument performance checks).
- Level I
 - Internal consistency tests to identify values in the data that appear atypical when compared to values of the entire dataset.
- Level II/III
 - External consistency tests to identify values in the data that appear atypical when compared to other datasets (e.g., historical data at the same site, collocated data).
 - Continued evaluation of the data as part of the data interpretation process, such as comparison to other datasets with possibly similar characteristics (e.g., the same region, period of time, background values, air mass) to identify systematic bias.

U.S. EPA, 1999a

Level 0: Field and Laboratory Checks

- Verify computer file entries against data sheets.
- Flag samples when significant deviations from measurement assumptions have occurred.
- Eliminate values for measurements that are known to be invalid because of instrument malfunctions.
- Replace data from a backup data acquisition system in the event of failure of the primary system.
- Adjust measurement values of quantifiable calibration or interference bias.

Chow et al., 1996

Steps in Data Validation – Site Operator

1. As data are collected, apply automated screening procedures, such as flags for automatic calibrations, and check for expected value ranges, rate of change, sticking values, and instrument alarm codes.
2. Review zero, span, and one-point QC verification information and calibration results to ensure that they meet acceptance criteria. If significant differences are observed, determine what corrective action steps are required and document the effect on the data.
3. Review automatic screening results. Investigate and correct flagging as necessary.



Steps in Data Validation – Site Operator

4. Review hourly averaged data and any flags that could affect data, marking any notations of invalidations. Record any information on the daily summaries that might be vital in a later review of the data.
5. Review analog and digital instrument diagnostic information, maintenance records, graphic displays of metadata, and site log notes. Ensure the notes are clear, up-to-date, and complete.
6. Provide the data, site notes, monthly maintenance sheets, metadata charts, and daily summaries for ready access by the data review and processing team.



Level I: 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.”

Judy Chow, Desert Research Institute

Level II/III: Comparisons to Other Data Sets

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



Steps in Data Validation – Data Reviewer

1. Assemble dataset: 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.).
2. Ensure that results of and suggestions from all audit reports have been incorporated into the database (data are still Level I).
3. Apply general screening criteria.
4. Prepare and inspect summary statistics for unrealistic maxima and minima and other factors.
5. Investigate internal consistency.
6. Flag data and document data modifications.
7. Perform spatial and temporal comparisons, compare data from different instruments (i.e., begin Level II).
8. Perform intercomparisons of the data (e.g., from two different instruments). Data are now Level III.

Example Data Overview

Parameter	% Missing/Below Detection	Parameter	% Missing/Below Detection	Parameter	% Missing/Below Detection
Bromide	100	Sc	93	Cl	21
Mo	100	Zr	91	Cu	19
Nb	100	Ir	90	OC Improve	19
Propionate	100	Rb	90	TC Improve	16
Sc	100	Cr	89	Si	12
Cs	99	W	89	Na+	7
Hf	98	Ag	88	Mn	7
Co	97	Au	88	Ca	3
Cd	96	Sr	85	Fe	3
Sn	96	Ni	82	K	3
Ba	96	Mg	80	S	3
In	96	Hg	75	Zn	3
La	96	Eu	69	Levogluconan	2

Sample
summary of
speciated
PM_{2.5} data

- As a part of Level 1 validation, it is useful to prepare a summary of the monitoring network by year: summarize what sites have data and how much data for which years.
- Use this summary to detect potential problems (such as missing data) and to determine what types of analyses are possible.
- Make sure that aerosol measurements are split into different size and analytical groups.

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: *Singling Out Unusual Data*

- Range checks: minimum and maximum concentrations
 - Extremes may change over time, e.g., by decade
- Temporal consistency checks: maximum hour test
 - Are data fitting the expected diurnal profile?
- Rate of change or spike check
- Sticking check: consecutive equal data values
- Buddy site check: comparison to nearby sites



Known Issues Often Needing Investigation

- High value during known auto-calibration hour(s)
- First observations after several hours with zeroes
- First observations after an extended missing data period
- Measurements made in complex terrain
- High-elevation monitor data (especially night-time)
- Unique meteorological conditions (thunderstorms, bay/land breeze, stagnation, subsidence, etc.)

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

- Air quality data reported during calibrations or spans.
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 or span and values changed.*
- Nitrogen oxides data found to have a constant offset based on comparisons of NO_x to $\text{NO} + \text{NO}_2$. *Data were adjusted and flagged as adjusted.*
- Ozone concentrations “capped” at 100 ppb. The instrument maximum concentration setting was found to be incorrect. *Data at 100 ppb were flagged as suspect low, and the instrument settings were adjusted.*



Examples of Problems in Meteorological Databases (and Validation Actions)

- Data that were physically consistent (i.e., reasonable values) and thus passed statistical checks, but were spatially inconsistent. For example, calm winds observed at a site when all nearby sites measured strong winds; *calm winds were flagged as “suspect.”*
- Wind direction did not vary while wind speeds did. *Stuck vane was identified (and repaired) and data were invalidated.*
- Ground clutter, migrating birds, and precipitation affected radar profiler measurements. *Affected measurements were invalidated.*



Data Validation Summary

For pollutant data validation,

- Understand formation, emissions, and transport
- Establish and apply screening criteria to identify potentially suspect data
- Investigate suspect data
- Invalidate data only if there is sufficient evidence
- Document invalid data (so others can learn)

Data validation is very important!

Resources

- Operator knowledge
- Previous documentation for the site and past data validation results
- EPA guidance documents, webinars, and training videos (available on AMTIC website)
- Workbooks (e.g., PAMS, PM_{2.5}, Air Toxics Data Analysis Workbooks)
- Web sites (next slide)



Key Internet Sites

- Ambient Monitoring Technology Information Center
<http://www.epa.gov/ttn/amtic/>
- EPA Quality Assurance
<http://www.epa.gov/oar/oaqps/qa/index.html>
- EPA QA handbook
<http://www.epa.gov/ttn/amtic/files/ambient/pm25/qa/QA-Handbook-Vol-II.pdf>
- NCore technical support
<http://www.epa.gov/ttn/amtic/ncore/guidance.html>

