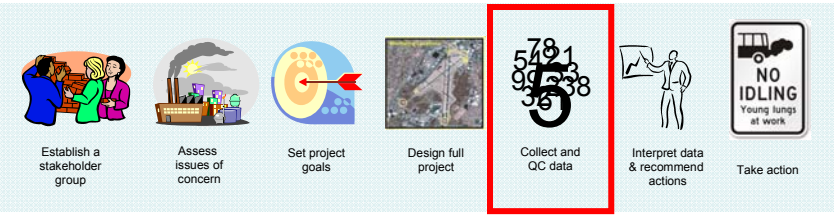
Agenda

Schedule	Topic
10:00-10:15	Introductions
10:15-12:00	1. Overview of successful air toxics monitoring projects (45 min.) 2. Getting started/Setting project goals (30 min.) 3. Monitoring strategy and design (part 1, 30 min.)
12:00-1:00	Lunch (on your own)
1:00-3:00	3. Monitoring strategy and design (part 2, 60 min.) Discussion (30 min.) 4. Collect and QC data (30 min.)
3:00-3:15	Break
3:15-5:00	5. Data analysis and interpretation (50 min.) 6. Taking action (20 min.) 7. Summary (30 min.) Wrap up

Collect and QC Data

- Ensuring high quality data
- Preparing data
- Validating data
- Lessons learned



Ensuring High-Quality Data

- Provide up-to-date SOPs and ensure that they are used consistently
- Schedule regular meetings among the monitoring staff to share ideas and lessons learned
- Require that data are reported with consistent units, naming conventions, etc.

Ensuring High-Quality Data (cont.)

- Regularly review data (e.g., use the web). Timely validation
  - minimizes the generation of additional data that may be invalid or suspect
  - maximizes the recoverable data
- Keep track of what is happening at/near the monitoring site (site log) for later use in data analysis efforts

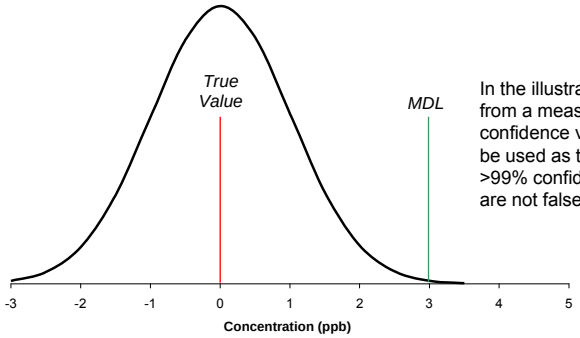
Data Preparation and Validation

- Data Quality 1 – Method blanks, spike recovery, MDLs, uncertainties, units
- Data Quality 2 – Replicates, duplicates, and collocated data
- Data Quantity – Number of samples, data completeness, fraction above MDL
- Validation
  - Level 0 – checking SOPs, lab audits, and blanks
  - Level 1 – internal consistency checks (scatter plots, time series)
  - Level 2 – historical consistency checks
  - Level 3 – spatial consistency checks
- Data Treatment – Aggregation, data below MDL

See Section 4 of the online workbook for an in-depth overview of data preparation and validation.

Method Detection Limits

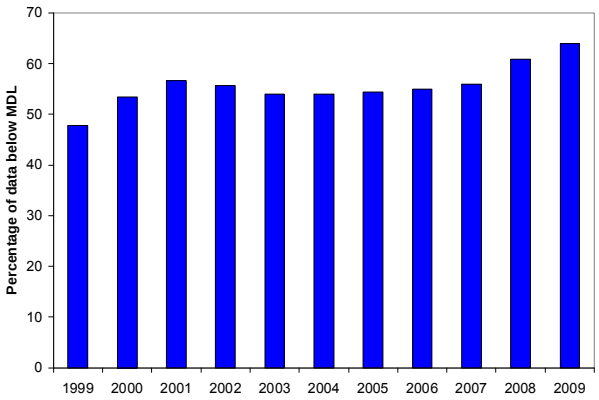
- The Code of Federal Regulations (CFR) defines the MDL as "The minimum concentration of a substance that can be measured and reported with 99% confidence that the analyte concentration is greater than zero and is determined from analysis of a sample in a given matrix containing the analyte."
- The purpose of an MDL is to discriminate against false positives. Values reported below the MDL have much higher uncertainty but can provide insight into the lower concentration distribution (i.e., are most values closer to the MDL or to zero?).



In the illustration, normally distributed results from a measured value of zero yields a 99% confidence value (3σ) at 3 ppb, which would be used as the MDL in this case. There is >99% confidence that values above 3 ppb are not false positives.

Environmental Protection Agency, 1982

Method Detection Limits (cont.)



- National 24-hr average air toxics data from AQS
- When concentrations are below MDL, summary statistics may be skewed and analysis will be complicated.

Other Performance Characteristics

- LOD – Limit of detection
  - Similar to MDL in concept and value
  - Level at which chance of a false positive is 1% (value is not zero)
- LOQ – limit of quantitation
  - ~3 times higher than MDL
  - Lowest concentration that can be measured accurately
- PQL – practical quantitation limit
  - 3 to 10 (5) times higher than MDL
  - Lowest concentration that can be measured accurately (analytical uncertainty of 10%)
- RL – Reporting limit
  - Laboratory-defined limit at which concentrations are reported and not censored
  - Often higher than MDL (e.g., LOQ or PQL)

Treating Data Below Detection

- Censored data may skew summary statistics such as means and medians used in assessing annual concentration distributions

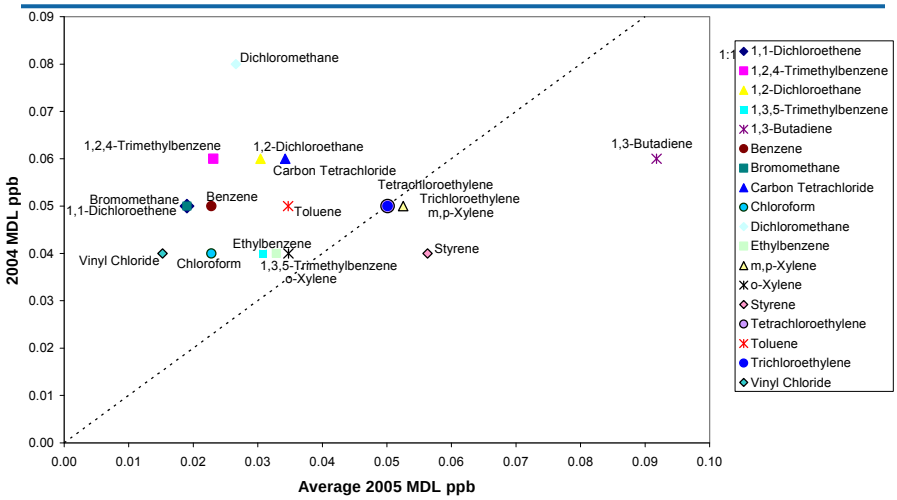
Method	Impact on Summary Stats	Drawbacks
Replace with zero	Biases concentrations low	False negatives
Replace with MDL/2	May be biased high or low	Can be false negative or positive
Replace with MDL	Biases concentrations high	False positives
Advanced stats: MLE, KM, or ROS	Best estimate for data sets when reasonable fraction is below MDL	Requires statistical software, difficult to apply

http://www.epa.gov/ttn/amtic/toxdat.html#workbook

Treating Data Below Detection (cont.)

- In a site-level analysis, in which the analyst knows how the data have been reported, multiple data treatment options exist
  - If uncensored values are reported below MDL, use the data "as is" with no substitution.
  - If uncensored values are not available, MDL/2 substitution for data at or below MDL may be useful for aggregate statistics (e.g., annual mean).
    - Bias will likely be small (10-40%) for data sets in which <70% of data are below MDL for means.
- Note at a high degree of censoring (>70% censored data), no technique will produce good estimates of summary statistics.
- More advanced statistical techniques are described in the data analysis workbook and are available in R statistical software packages; improvements to real data sets are often small.

MDL Example: JATAP



Average MDLs were lower in 2005 than in 2004 for most species.

Collocated Data

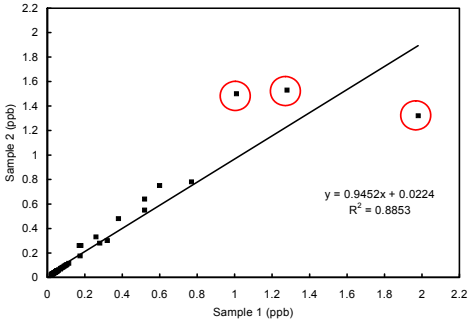
- Differences between replicate, duplicate, and collocated measurements
  - A **replicate sample** is a single sample that is chemically analyzed multiple times.
  - A **duplicate sample** is a single sample that is chemically analyzed twice.

    These samples provide a measure of the precision of the chemical analysis, but do not provide any error estimates for the sample collection method.
  - In contrast, **collocated samples** are two samples collected at the same location and time by equivalent samplers and chemically analyzed by the same method.

    These samples provide a measure of the precision of both sample collection and chemical analysis.
- EPA's National Air Toxics Trend Sites (NATTS) program proposed the following collocated data standards:
  - Less than 25% bias between collocated samples
  - Less than 15% coefficient of variation for each pollutant

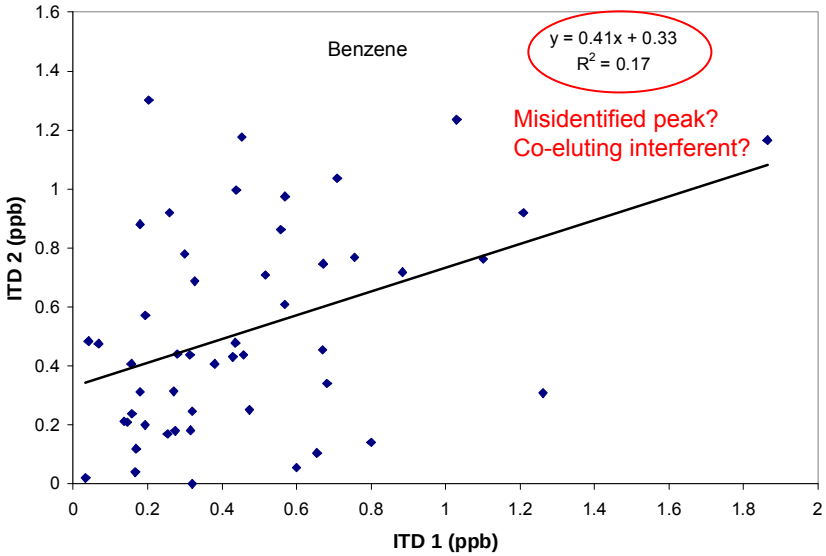
Handling Collocated Data

- If collocated data agree,
  - slope will be close to 1
  - intercept will be close to 0
  - R^2 value will be close to 1
- The graph shows three species identified as suspect because they failed to meet the NATTS criteria.
  - Confidence in the measurements of all species was reduced for this example.



Scatter plot of collocated measurements for multiple species collected at an urban southwestern site. Circled measurements (acetylene, toluene, and methyl ethyl ketone) were identified as suspect.

Collocated Data – Poor Quality



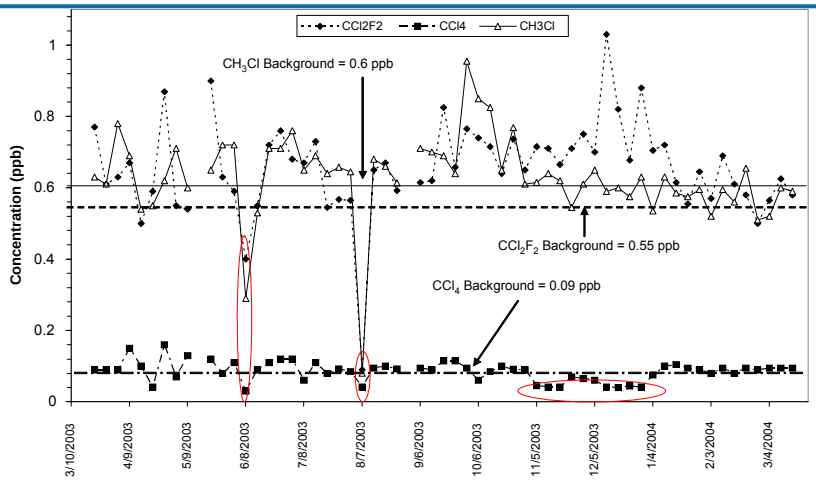
Data Validation

- Data validation is defined as the process of determining the quality and validity of observations.
- The purpose of data validation is to detect and verify any data values that may not represent the actual physical and chemical conditions at the sampling station before the data are used in analysis.
- The primary objective is to produce a database with values that are of a known quality, an acceptable quality, or a level of uncertainty given the analyses intended to be conducted.

Data Validation (cont.)

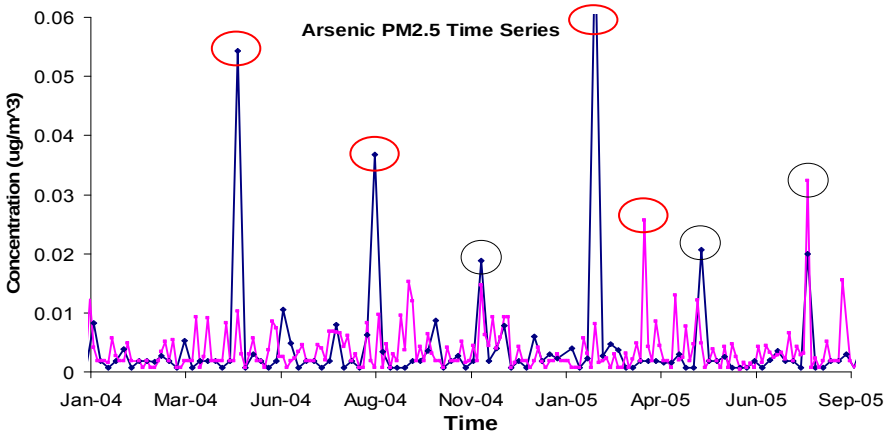
- Look at your data—visual inspection is vital.
- Manipulate your data—sort it, graph it, map it—so that it begins to tell a story. Several checks may be made during the beginning stages of data validation to single out odd data
  - Range checks: check minimum and maximum concentrations for anomalous values.
  - Buddy site check: compare concentrations at one site to nearby sites to identify anomalous differences.
  - Sticking check: check data for consecutive equal data values which indicate the possibility of censored data not appropriately flagged.
  - Comparison to remote background concentrations: urban air toxics concentrations should not be lower than remote background concentrations.

Screening Data Using Remote Background Concentrations



Concentrations (ppb) of carbon tetrachloride (CCl4), dichlorodifluoromethane (CCl2F2), and methyl chloride (CH3Cl) from 2003 and 2004. Northern hemisphere background concentrations of each species were plotted as a line. Concentration dips well below background concentrations are circled.

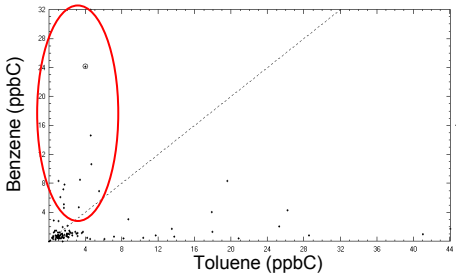
Data Validation Example: Buddy Check



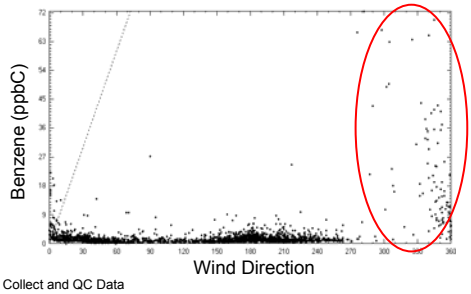
Sample time series of 24-hr arsenic $PM_{2.5}$ measurements at two sites about five miles apart. Both sites show above-average arsenic concentrations and are located near a major emissions source. The figure was created in Microsoft Excel.

Data Validation Example: Investigating Suspect Data

Initial Analysis: Typically, toluene concentrations are higher than benzene concentrations. The pattern shown in the graphic is unexpected; further investigation of the data is needed.



Advanced Analysis: Wind direction data were used to identify possible reasons for the high benzene concentrations. The highest benzene concentrations are typically coming from north of the site. Site and emission inventory inspection showed a source of coke oven emissions, which include benzene but not toluene, to the north—providing a reasonable explanation for these data (and helping prove their validity).



Things to Consider When Evaluating Your Data

- *Levels of other pollutants*
  A high concentration of benzene may be valid when concentrations of all mobile source air toxics in the sample are also elevated.
- *Time of day/year*
  Higher concentrations of some air toxics are expected in the summer (such as formaldehyde) than in the winter and vice-versa for benzene.
- *Observations at other sites*
  High concentrations of a pollutant at several sites in an area on the same date may indicate a real emission event.
- *Audits and inter-laboratory comparisons*
  If data are from differing sources, how well did the concentrations compare between labs? Did audits show some specific "problem" pollutants?
- *Site characteristics*
  High concentrations may be expected for a pollutant emitted by a nearby source.
- *Unique events (e.g., holiday fireworks)*
  High concentrations of trace metals associated with fireworks are seen around the Fourth of July and New Year's Day at many sites.

Investigating Outliers

- Use wind direction data (e.g., Do outliers occur from a consistent wind direction?)
- Use subsets of data (e.g., inspect high concentration days vs. other days for differences in meteorology or emissions)
- Investigate industrial or agricultural operating schedules, unusual events, etc. (e.g., Were high metals data associated with a dust event?)
- Determine local traffic patterns (e.g., When does peak traffic occur? Is there a recreational area or event venue nearby?).
- If no explanation is forthcoming, try contacting the agency that collected the data; they may have realized a problem too recently to report it, or your question may alert them to a problem with data collection, analysis, or reporting.

Lessons Learned in Data Collection and Validation

- There is no substitute for local knowledge about monitoring sites; operators or those who have extensive knowledge of the area are a unique resource for data analysts.
- Look at your data early and often to correct problems before the study is over.
- Always validate your data!

Lessons Learned in Data Collection and Validation

- Inter laboratory precision data indicated that laboratory selection could be a major factor influencing data comparability nationwide… but, comparability between laboratories is improving as a result of the performance evaluation program.