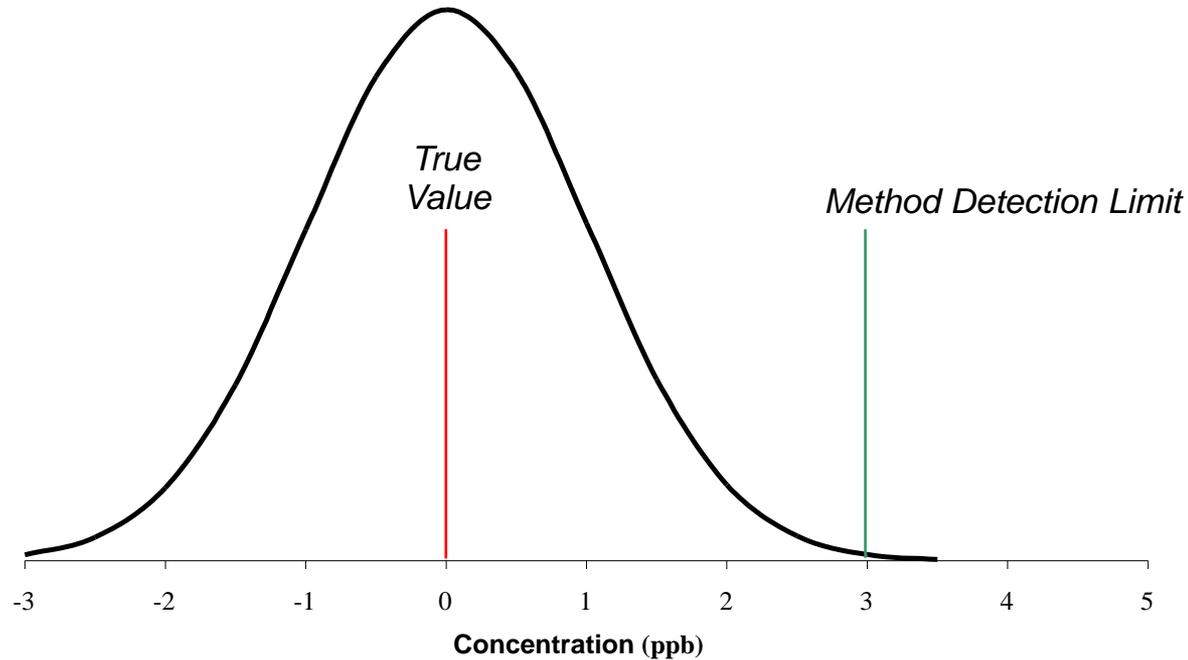


# Preparing Data for Analysis



# Analysis Flow Chart –Theory

**Set Project Goals**

**Design Monitoring Study**

**Data Quality Objectives (DQOs)**

**Perform Monitoring and Analysis**

**Data Preparation and Validation**

**Data Analysis**

**Communication and Action**

# Data Preparation Thought Process

Questions	Examples of Answers
What is the analysis objective?	Determine whether concentrations are above health benchmark, determine whether concentrations are changing, characterize spatial patterns, identify emissions sources
What data processing might be needed to prepare the data for analysis?	Unit conversion, method blank correction, data aggregation, outlier removal, removing incomplete data, MDL data substitution
What data validation can be done to ensure data are of adequate quality for analysis?	Review collocated data, inspect summary statistics and concentration ranges, review time series plots of concentrations and detection limits, identify censored data, compare to historical data, compare to nearby data, compare to national data
What data quality objectives may not be met and might confound/derail the analysis?	Data completeness, temporal representativeness, data above detection limits, data meeting analytical QA criteria, sufficient numbers of samples

# Data Acquisition

- EPA's Air Quality System (AQS)
  - IMPROVE speciated  $PM_{2.5}^*$  data (VIEWS website)
  - SEARCH speciated  $PM_{2.5}$  data (Atmospheric Research Analysis website)
- Local, state, and tribal air quality agency databases (i.e., some data are not yet submitted to AQS)
- Raw data reported from analytical laboratory or the field

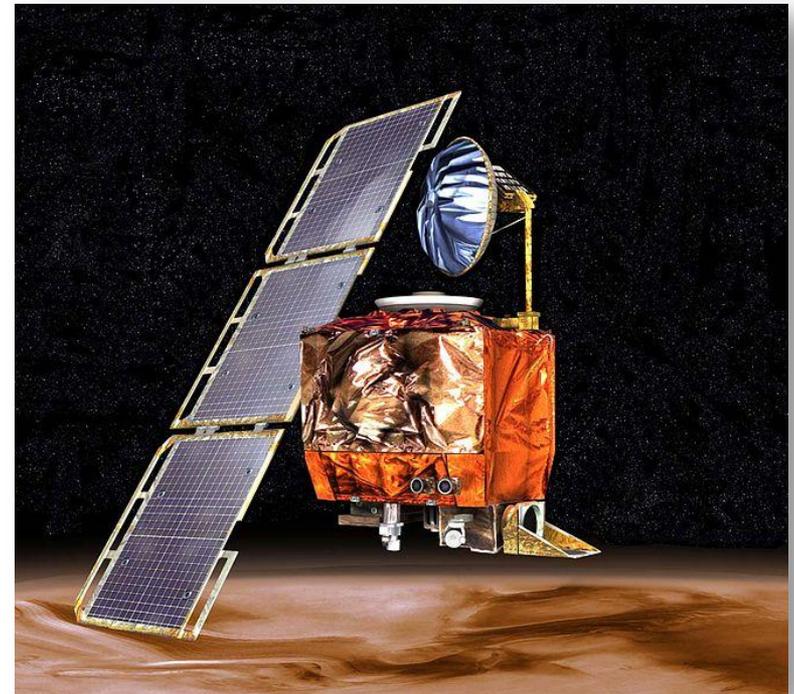
**Pro tip:** Check data to make sure it looks like you expected it to before moving on to data preparation.

\*  $PM_{2.5}$  = particulate matter with aerodynamic diameter less than 2.5 microns

# Data Preparation

- Unit conversion
- Blank correction
- MDL substitution
- Data aggregation
  - Quarterly averages
  - Annual averages
- Data removal
  - Invalid data
  - Outliers
  - Exceptional event data
  - Incomplete data

Mars Climate Orbiter – expensive unit conversion example



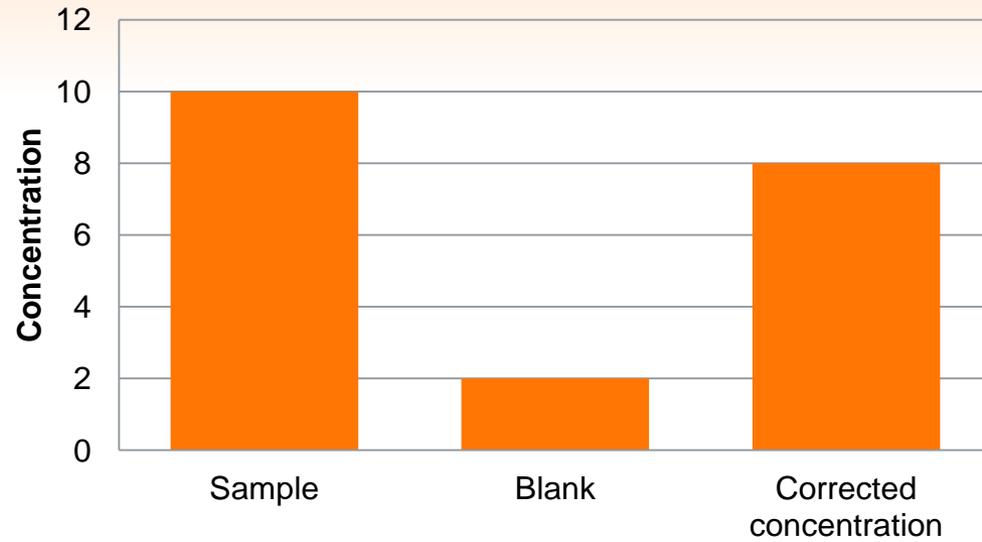
# Unit Conversion

- Make sure your units are consistent and applicable to the analysis of interest
  - Make sure units are changed for accompanying metadata information, too (e.g., MDLs)
  - Ensure that data aggregation is occurring on pollutants all in the same unit
- Useful unit conversions for some gas phase species
  - $[\text{conc. in } \mu\text{g}/\text{m}^3] = ([\text{conc. in ppb}] * \text{MW} * 298 * P) / (24.45 * T * 760)$
  - $[\text{conc. in ppb}] = ([\text{conc. in } \mu\text{g}/\text{m}^3] * 24.45 * T * 760) / (\text{MW} * 298 * P)$
  - $\text{ppbC} = \text{ppb} \times (\# \text{ of carbons in the molecule})$

where:

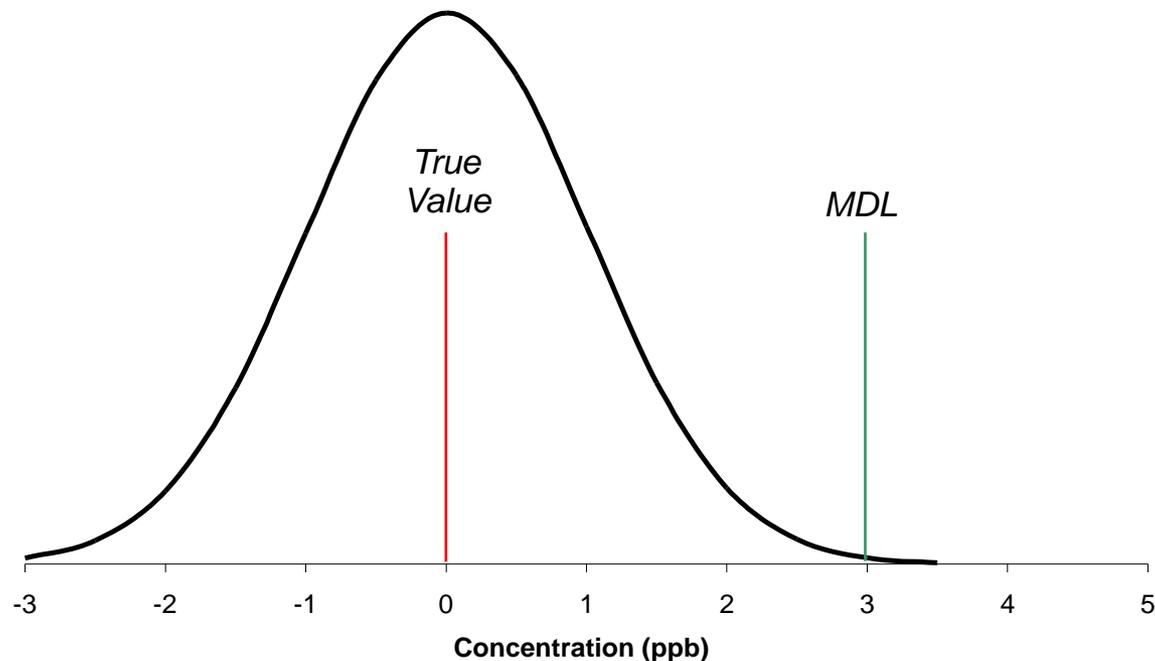
- MW = molecular weight of compound [g/mol]
  - P = absolute pressure of air [mm Hg]; 1 atm = 760 mm Hg
  - T = temperature of air [K]; 298 K is standard
- See the toxics data analysis workbook for examples

# Blank Correction



# Method Detection Limits

MDL – Method performance characteristic specifying a concentration level that is ~99% likely to originate from non-zero analyte concentrations.



# Example: MDL Matters!

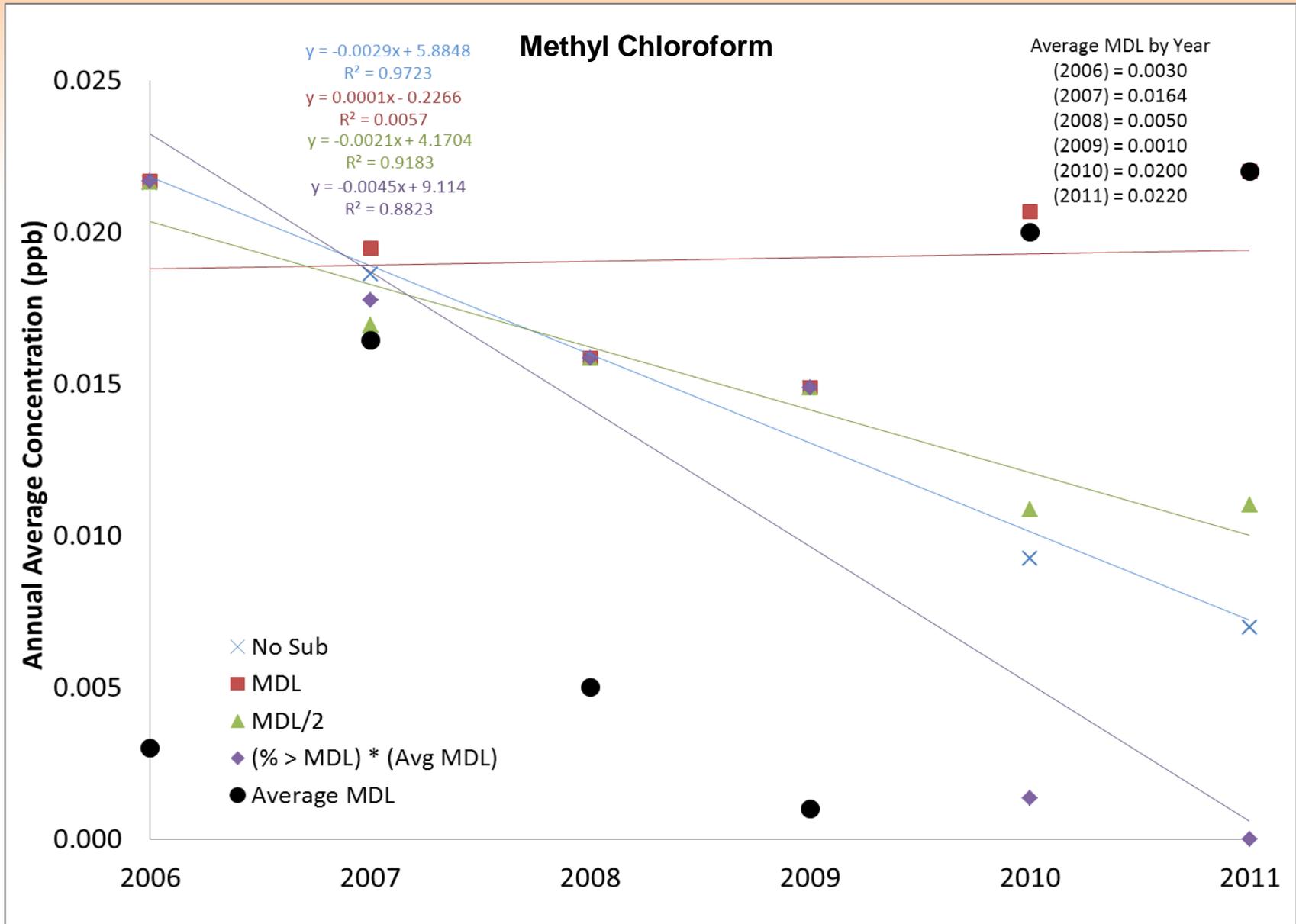


- National 24-hr average air toxics data from AQS show more than 50% of data are below MDL
- When concentrations are below MDL, summary statistics may be skewed and analysis will be complicated

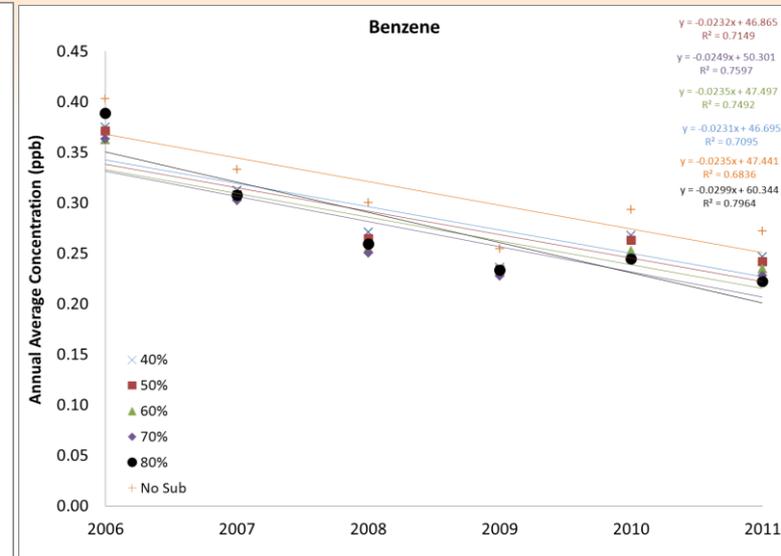
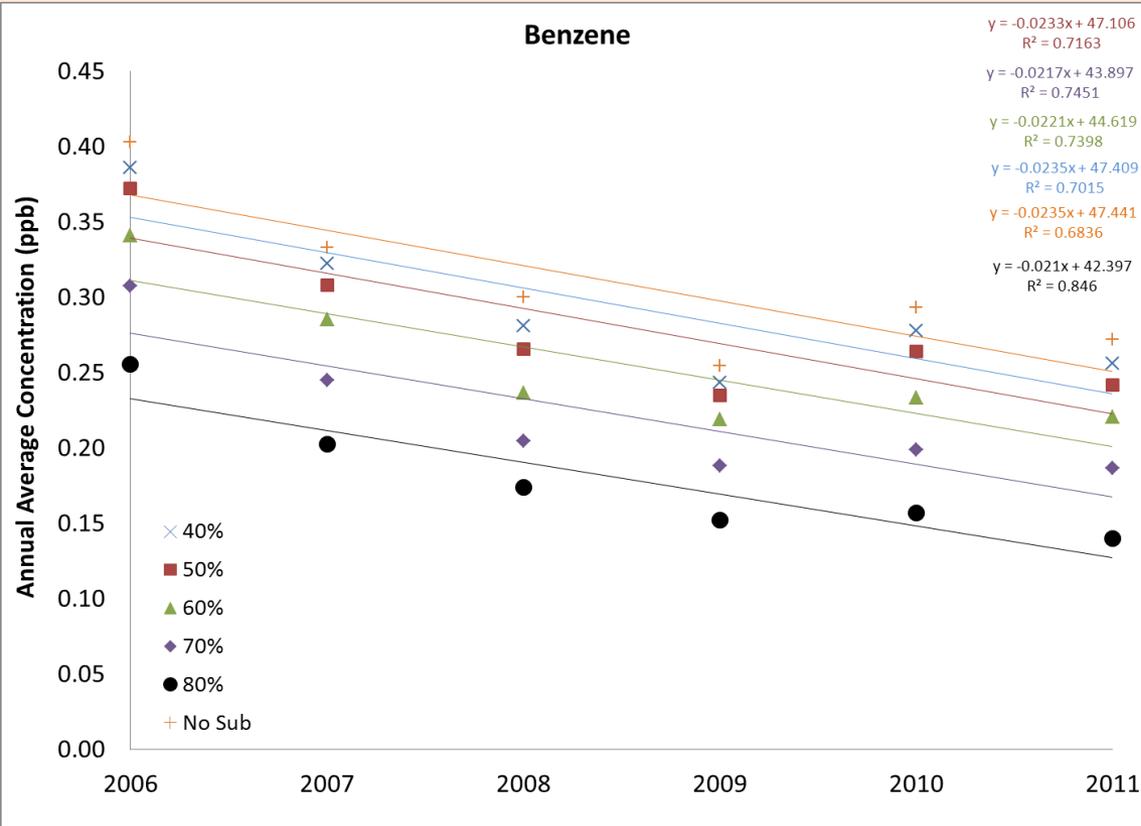
# MDL Censoring

- EPA currently recommends that all instrument values be reported
- However, some current data and much historical data are/were censored when concentrations were below MDL
  - Zero
  - MDL
  - MDL/2
- If data at or below MDL are censored, this will potentially bias future analyses
  - Data replaced with zero are likely biased low
  - Data replaced with MDL are likely biased high
  - Data replaced with MDL/2 may be high or low

# MDL Censoring Example 1



# MDL Censoring Example 2



# Advanced Statistical Techniques

- Advanced statistical techniques can be used to estimate summary statistics (e.g., mean) but require statistical software and large sample size
  - MLE (maximum likelihood estimation)
  - KM (Kaplan-Meier) aka survival analysis
  - ROS (Robust Regression on Order Statistics)
- See data analysis workbook for details

# Data Treatment Methods (Summary)

EPA's current recommendations for treating data below MDL are provided in the table below; EPA is developing more definitive guidance.

Use	Small # of Samples	Large # of Samples	Very Large # of Samples
Exploratory Use	MDL/2 <i>(if only a few samples are &lt; MDL)</i>	MDL/2 <i>(if &lt; 15% of samples are &lt; MDL)</i>	Cohen <i>(normal distribution)</i> Kaplan Meier <i>(other than normal)</i>
Publication Use	Kaplan Meier	Kaplan Meier Cohen <i>(if approx. normal distribution)</i>	Cohen <i>(normal distribution)</i> Kaplan Meier <i>(other than normal)</i>
Regulatory Use	Kaplan Meier	Kaplan Meier	Kaplan Meier

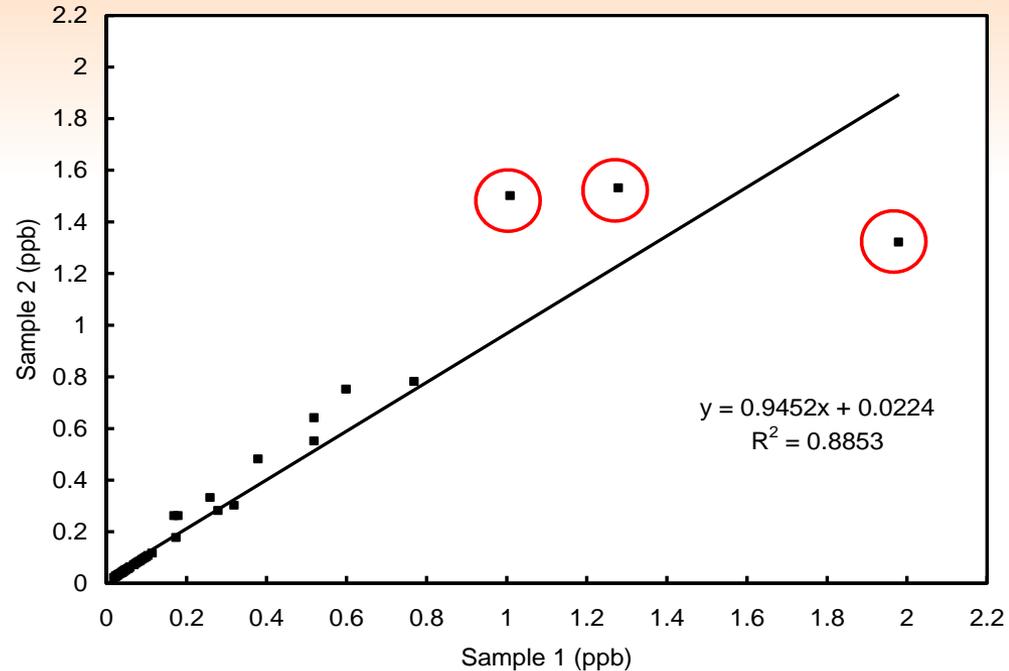
Warren and Nussbaum, 2009

# 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

- Investigate agreement between collocated data using scatter plots and linear regression lines. If collocated data agree,
  - Slope will be close to 1
  - Intercept will be close to 0
  - $R^2$  value will be close to 1
- Example
  - Three species circled in the figure were 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.

# Data Aggregation

## *Creating Valid Quarterly and Annual Averages*

- It is suggested that data meet the 75% completeness criteria as determined by sample frequency, assuming an average of 30 days in a month. Note that low sample frequency data may not adequately represent quarterly values with certainty.
- Sampling frequency may not be provided with data; frequency can usually be inferred by visual inspection.
- Annual averages just require three of four valid quarterly averages.

Frequency	75% Monthly Completeness Cutoff
Daily	68
Every 3 <sup>rd</sup> day	24
Every 6 <sup>th</sup> day	12
Every 12 <sup>th</sup> day	6

# Data Validation

The identification of outliers, errors, or biases is typically carried out in several stages or validation levels.\*

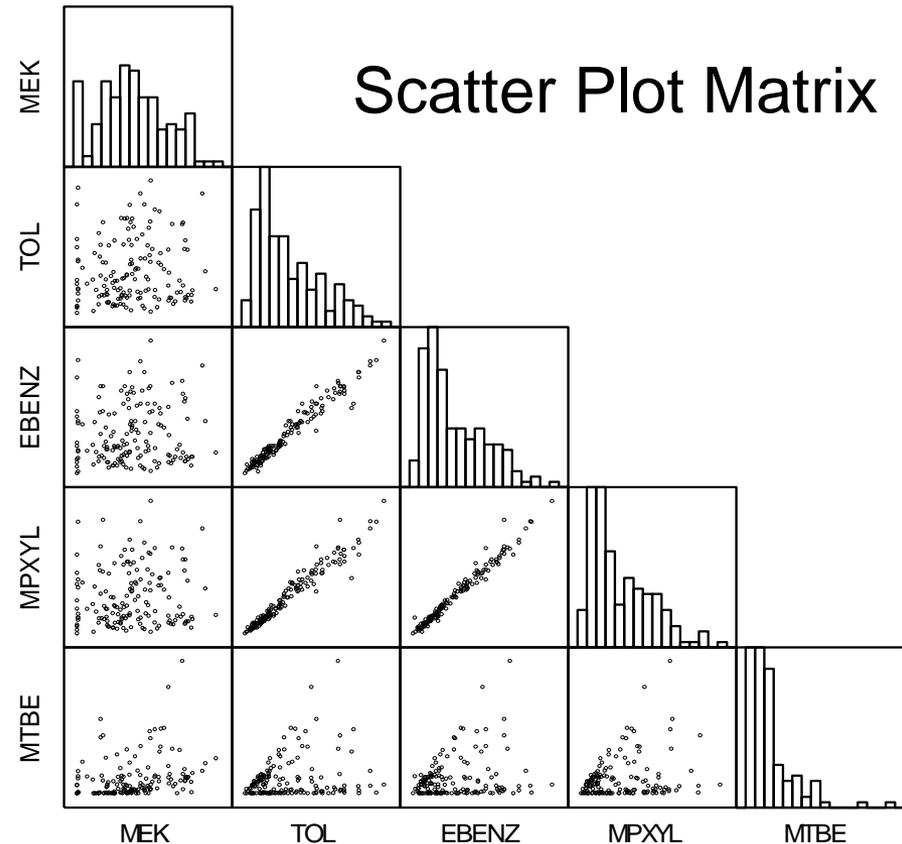
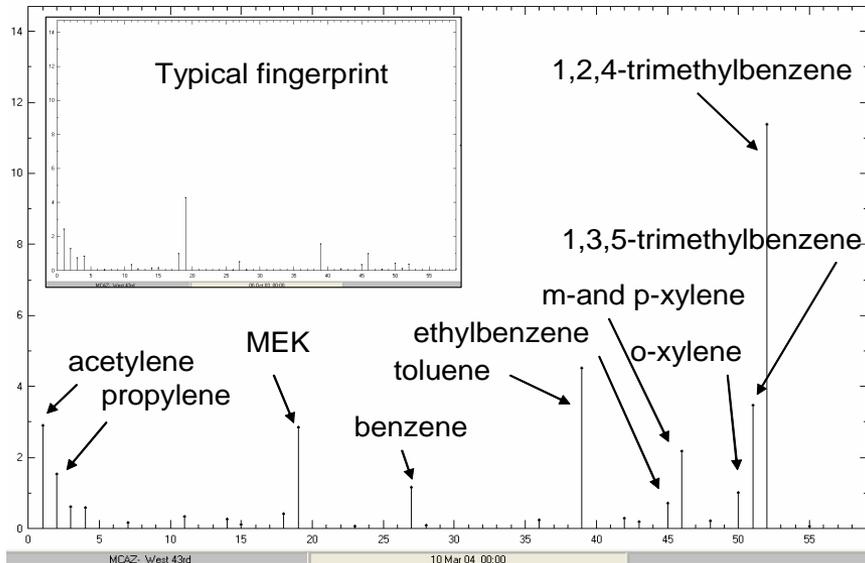
- Level 0: Routine verification that field and laboratory operations were conducted in accordance with standard operating procedures (SOPs) and that initial data processing and reporting were performed in accordance with the SOP (*typically the monitoring entity performs this step*).
- Level I: Internal consistency tests to identify values in the data that appear atypical when compared to values in the entire data set.
- Level II: Comparisons of current data with historical data (from the same site) to verify consistency over time.
- Level III: Parallel consistency tests with other data sets with possibly similar characteristics (e.g., the same region, period of time, background values, air mass) to identify systematic bias.

\* U.S. Environmental Protection Agency, 1999

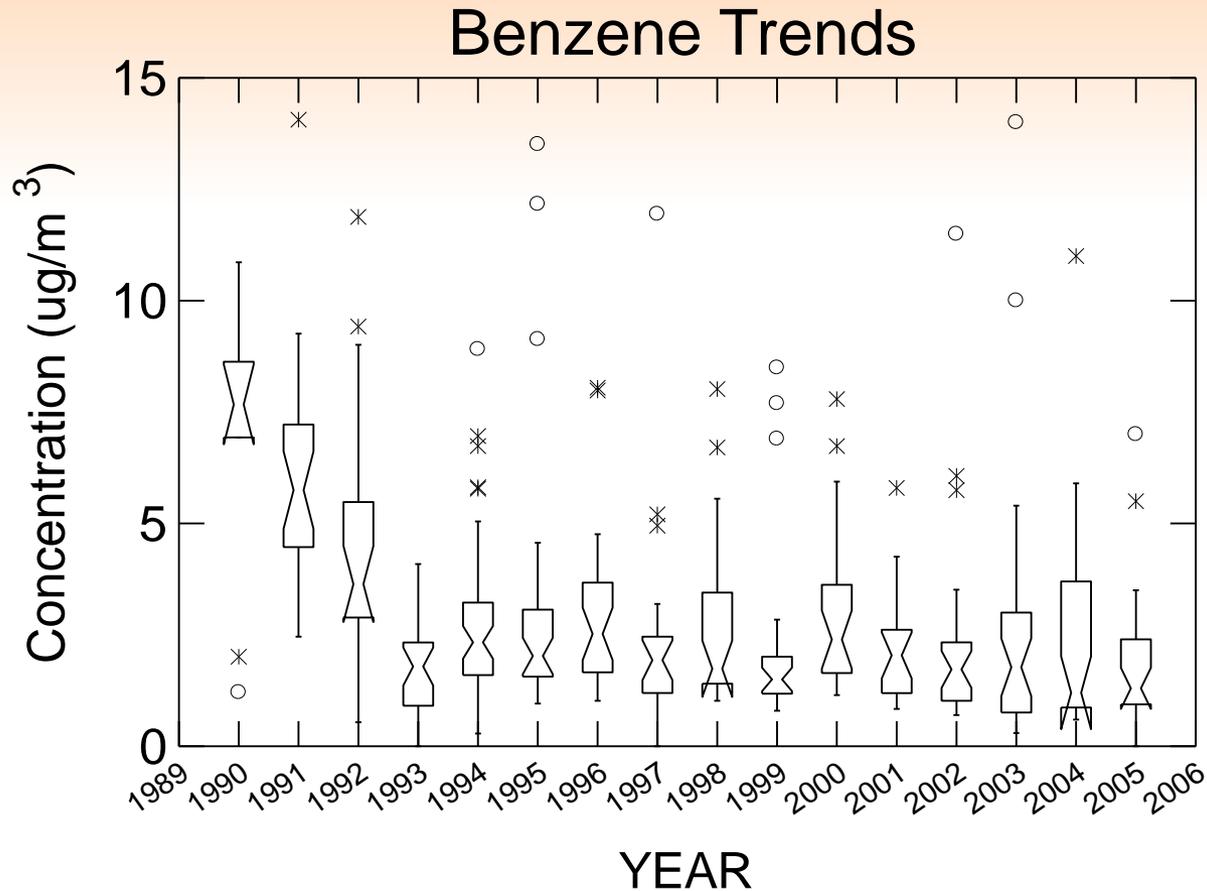
# Level I Validation Checks

- Fingerprint checks
- Scatter plots for expected relationships
- Time series analysis
- Sticking checks
- Minimum, maximum, range checks

## Fingerprint Plot



# Level II Validation



Notched box whisker plot of 24-hr average concentration of benzene by year at an urban monitoring site in the United States. Concentrations show a substantial change from 1990 to 1993.

# Level III Validation Checks

- Comparisons with remote background concentrations
- Buddy checks
- Comparisons with national concentrations

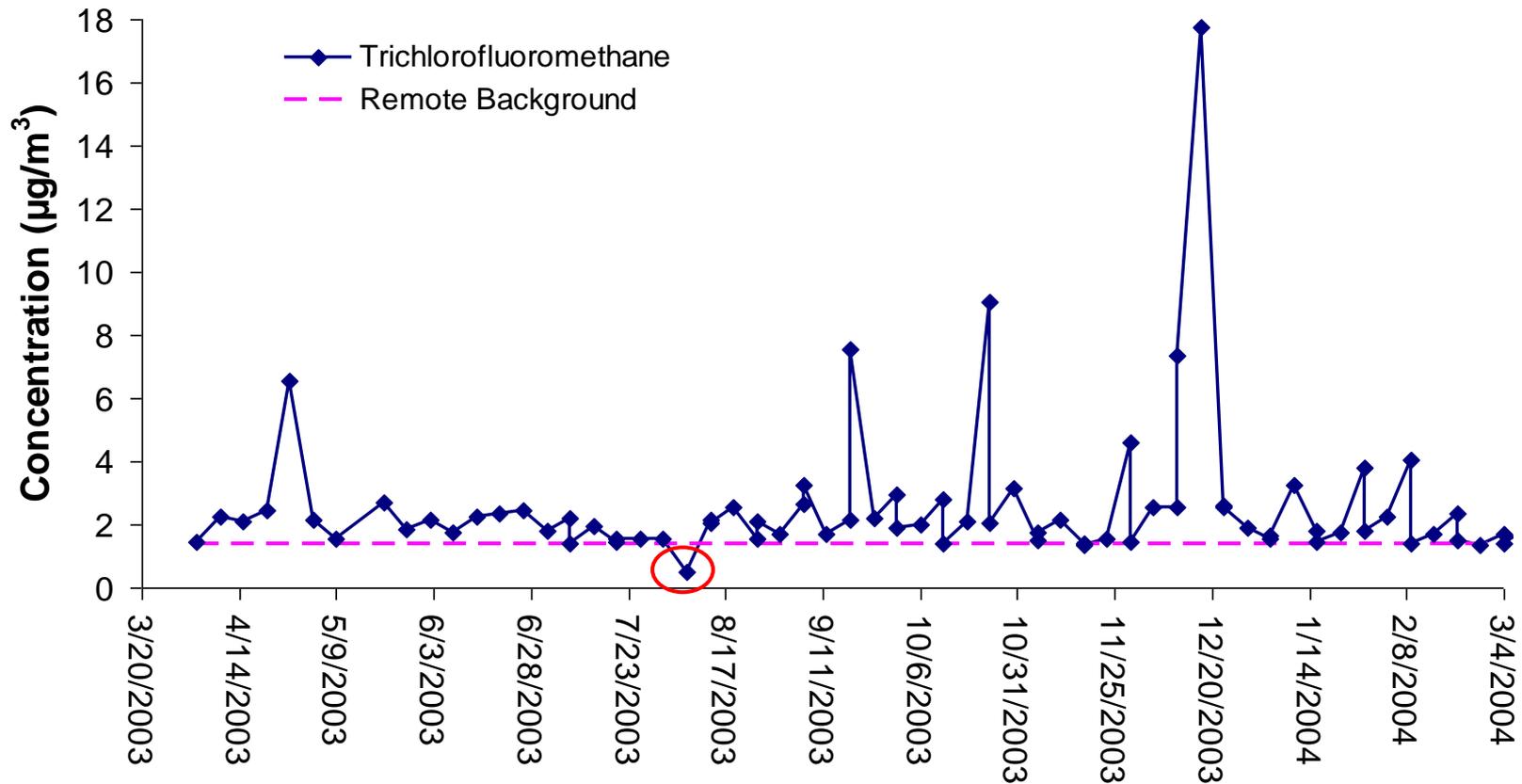
# Screening Data Using Remote Background Concentrations

- Known remote background concentrations of air toxics can be used as lower limits for data screening. A cutoff value of 30% lower than the background concentration is used as a margin of error.
- Data below this value may be identified as suspect.
- If data are identified as below the background concentration, the first things to check are
  - Units (e.g., were units reported and/or converted correctly?)
  - Sticking from substituted values such as MDL/2, MDL/10, or 0.

Pollutant	Remote Background Concentration ( $\mu\text{g}/\text{m}^3$ )	Cutoff Value ( $\mu\text{g}/\text{m}^3$ )
Acetaldehyde	0.14	0.10
Benzene	0.125	0.088
Carbon Tetrachloride	0.577	0.40
Chloroform	0.045	0.032
Formaldehyde	0.18	0.13
Methylene Chloride	0.127	0.089
Tetrachloroethylene	0.016	0.011
Methyl Chloride	1.09	0.76

Adapted from McCarthy et al. (2011) "Estimation of Background Concentrations for NATA 2008," Final Report, STI-910219-4224

# Remote Background Check Example

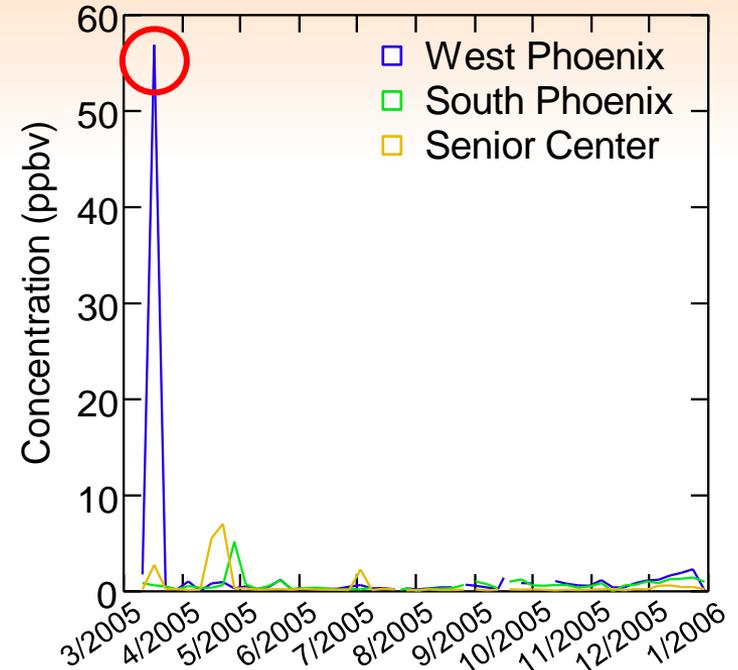


- The plot shows a time series of concentrations of trichlorofluoromethane compared to background concentrations measured at remote sites in the northern hemisphere.
- A significant dip in concentrations is circled in red. Concentrations at this monitor were typically equal to or greater than background concentrations, as expected for urban locations.
- The circled value was more than 20% below the background level and was identified as suspect for further review.

# Buddy Site Check

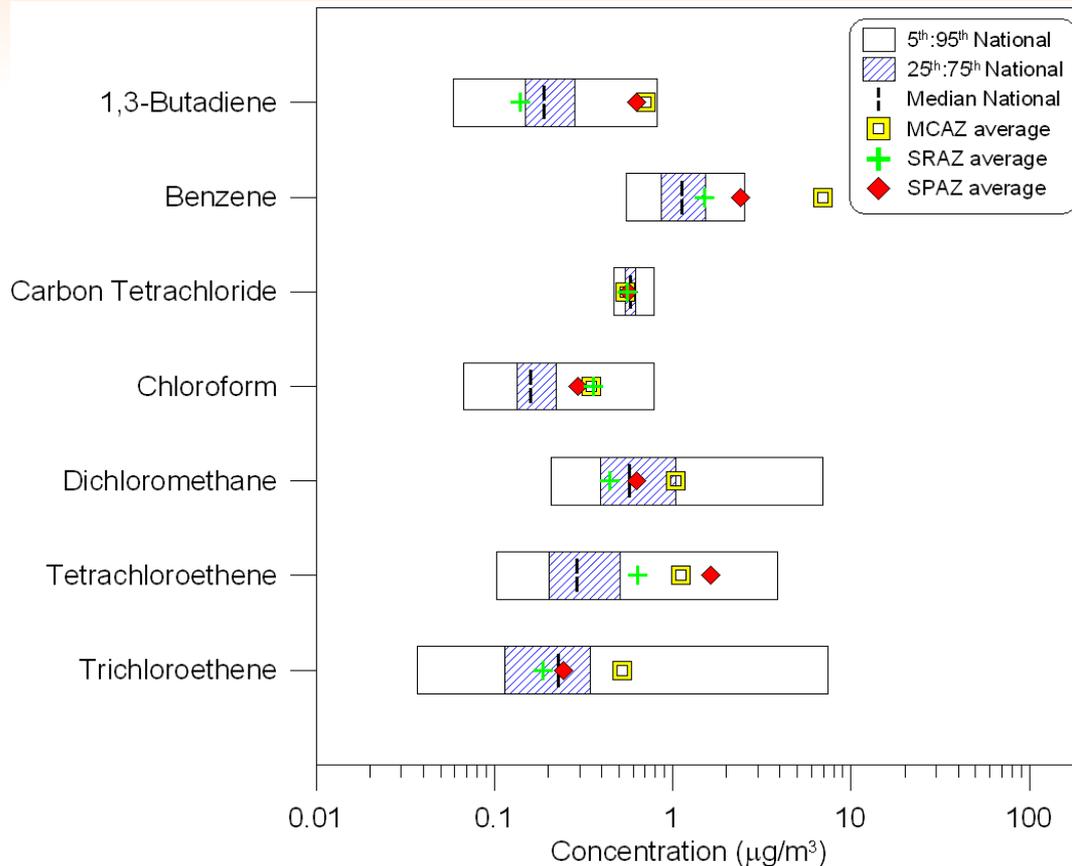
- Buddy site checks are useful in identifying suspect data.
- In the example, time series of benzene concentrations for three sites are plotted.
- There is clearly a suspect data point at the West Phoenix site in March 2005, which is not corroborated by the other sites. This indicates that the data point should be considered suspect because a concentration spike of that magnitude should register at nearby sites.
  - Investigation into these data showed that this event corresponds to a single data point significantly higher than the others.
  - Further investigation revealed that many species showed the same behavior at the West Phoenix site. The site may be impacted by a local source or sources.

## Benzene



# Comparison to National Concentrations

MCAZ = West Phoenix  
SPAZ = South Phoenix  
SRAZ = Senior Center

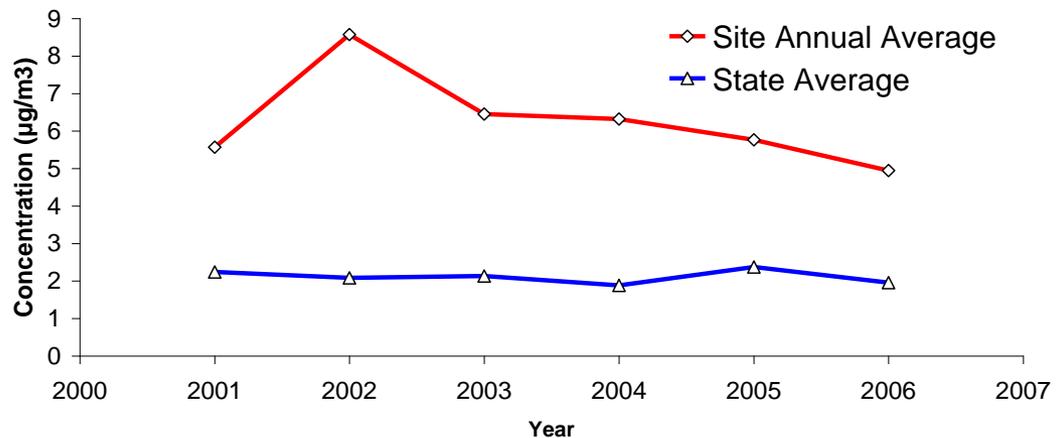
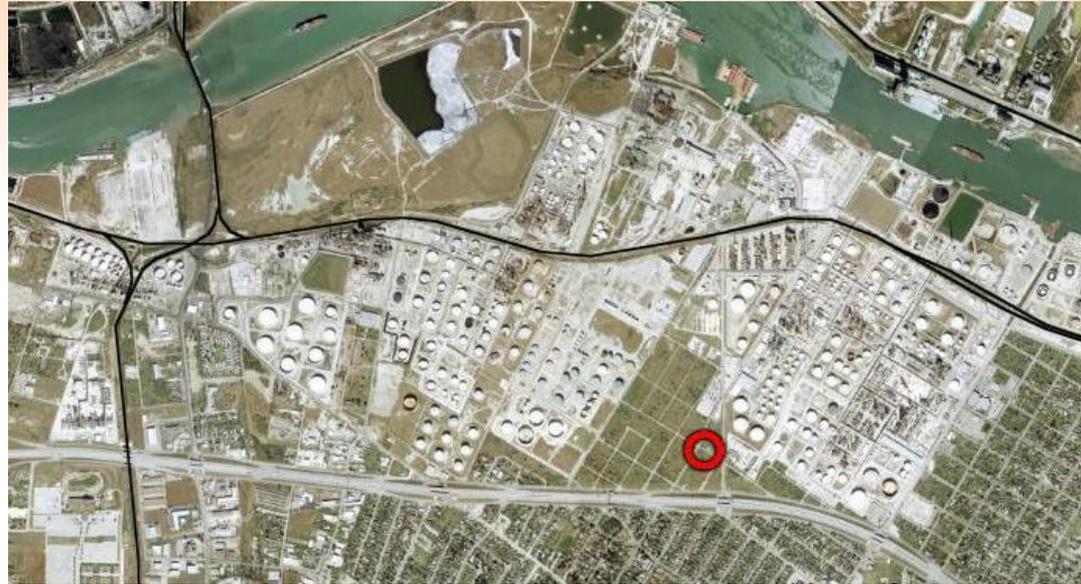


Are concentrations within standard concentration ranges?

Are the concentrations high or low? If so, why?

# Supplementing Air Toxics Data

- Metadata are not routinely available.
- Site metadata are useful in analyses for identifying local sources or roadways or physical attributes of the site.
- This satellite image shows the monitoring site (red circle) near an oil refinery that likely influences VOC concentrations at the site.
- A comparison of benzene annual averages at this site (red) to the state-wide annual average (blue) indicates benzene concentrations at this site are significantly higher.
- In this case, preliminary evidence shows the refinery may influence local benzene concentrations; however, this evidence is not conclusive.



# Data Preparation Check List

## ○ Acquire data

- Check for availability of supplementary data
  - Meteorological measurements
  - Additional species
  - Metadata
- Use supplementary data
  - Thoroughly review all metadata describing what/why/how measurements were made.
  - Find out about site characteristics, including
    - Meteorology
    - Local emissions sources
    - Geography

## ○ Know your data

- A general knowledge of air toxics behaviors is invaluable. Know and understand typical relationships and patterns that have been observed in air toxics data.

## ○ Process your data

- Investigate collocated data. Do they agree?
- Create valid data aggregates
  - Check for data completeness
  - Prepare and inspect valid aggregates and calculate the percentage of data below MDL
- Identify censored data and make MDL substitutions if necessary
  - Use knowledge of data reporting methods to identify substitution used for data below detection, if any.
  - If reporting of data below detection is unknown, separate data below detection and check for repetitive values or linear relationships detection limits
  - If data are uncensored, use “as is”
  - If data are censored, make MDL/2 substitutions or more sophisticated method as needed
  - If the data contain a mixture of censored and uncensored data,
    - Test two substitution methods for a sample analysis: (1) MDL/2 substitution for all data and (2) MDL/2 substitution for censored data, leaving uncensored data “as is.”
    - If direction and magnitude of trends results agree, keep substitution method 2.

## ○ Validate your data

- Get an overview—prepare and inspect summary statistics
- Apply visual and graphical methods to illuminate data issues and outliers
  - Buddy site check
  - Remote background comparison
  - Scatter plots
  - Time series
  - Fingerprint plots
- Flag suspect data
- Investigate suspect data using
  - Local sources/wind direction
  - Subsets of data
  - Unusual events
- Exclude invalid data
  - If you cannot prove the data are invalid, flag as suspect. These data may be removed from some analyses as an outlier even if they cannot be invalidated. Advanced analyses may provide more insight into the data.

# Data Preparation Check List (2 of 2)

## ○ Acquire data

- Check for availability of supplementary data
  - Meteorological measurements
  - Additional species
  - Metadata
- Use supplementary data
  - Thoroughly review all metadata describing what/why/how measurements were made.
  - Find out about site characteristics including
    - Meteorology
    - Local emissions sources
    - Geography

## ○ Know your data

- A general knowledge of air toxics behaviors is invaluable. Know and understand typical relationships and patterns that have been observed in air toxics data.

## ○ Process your data

- Investigate collocated data. Do they agree?
- Create valid data aggregates
  - Check for data completeness
  - Prepare and inspect valid aggregates and calculate the percentage of data below MDL
- Identify censored data and make MDL substitutions if necessary
  - Use knowledge of data reporting methods to identify substitution used for data below detection, if any.
  - If reporting of data below detection is unknown, separate data below detection and check for repetitive values or linear relationships detection limits
  - If data are uncensored, use “as is”
  - If data are censored, make MDL/2 substitutions or more sophisticated method as needed

## ○ Process your data (continued)

- Identify censored data and make MDL substitutions if necessary (continued)
  - If the data contain a mixture of censored and uncensored data,
    - Test two substitution methods for a sample analysis: (1) MDL/2 substitution for all data and (2) MDL/2 substitution for censored data, leaving uncensored data “as is.”
    - If direction and magnitude of trends results agree, keep substitution method 2.

## ○ Validate your data

- Get an overview—prepare and inspect summary statistics
- Apply visual and graphical methods to illuminate data issues and outliers
  - Buddy site check
  - Remote background comparison
  - Scatter plots
  - Time series
  - Fingerprint plots
- Flag suspect data
- Investigate suspect data using
  - Local sources/wind direction
  - Subsets of data
  - Unusual events
- Exclude invalid data
  - If you cannot prove the data are invalid, flag as suspect. These data may be removed from some analyses as an outlier even if they cannot be invalidated. Advanced analyses may provide more insight into the data.