Air Toxics Data Analysis Workbook

Prepared for:

U.S. Environmental Protection Agency
Office of Air Quality Planning and Standards
Research Triangle Park, NC

June 2009

Table of Contents (1 of 2)

Table of Contents (2 of 2)

Page numbers refer to the workbook rather than the training slides…

Disclaimer

The information and procedures set forth here are intended as a technical resource to those conducting analysis of air toxics monitoring data. This document does not constitute rulemaking by the Agency and cannot be relied on to create a substantive or procedural right enforceable by any party in litigation with the United States. As indicated by the use of non-mandatory language such as "may" and "should," it provides recommendations and does not impose any legally binding requirements. In the event of a conflict between the discussion in this document and any Federal statute or regulation, this document would not be controlling. The mention of commercial products, their source, or their use in connection with material reported herein is not to be construed as actual or implied endorsement of such products. This is a living document and may be revised periodically.

The Environmental Protection Agency welcomes public input on this document at any time. Comments should be sent to Barbara Driscoll (driscoll.barbara@epa.gov).

The training material is intended for use in webinars or other training venues by the instructor to accompany the workbook. In general, the training material is briefer, splits individual workbook pages across two or more slides, and omits some supporting material such as references.

Workbook Content Summary

- Introduction
- Definitions and acronyms
- Background
- Preparing data for analysis
- Characterizing air toxics
- Quantifying trends in air toxics
- Advanced data analysis techniques
- Suggested analyses

Workbook Purpose (1 of 2)

- This workbook was designed to
  - serve as an overview of the sizeable topic of air toxics data analysis;
  - provide suggestions on the methodology to use in analyzing air toxics data, building on the experience gained in the past several years of national-level data analysis efforts; and
  - document current methodology being used in national data analysis efforts.
- The workbook contains a different topic area in each section. Distinctions between methods used to assess the data at a national level and methods that can be applied at a site level are provided.

Workbook Purpose (2 of 2)

- Figures are used to show example analyses.
  - The figures are not intended to show the only way to perform an analysis but rather to provide the analyst with a starting point.
  - Most figure captions list the tool used to present the data, the data used in the analysis, an observation or interpretation point, and a reference.
- References are provided at the end of each section.

Introduction to Air Toxics

Introduction to Air Toxics
What's Covered in This Section?

- What are air toxics?
- Why analyze ambient air toxics data?
- Types of questions analysts want to answer
- Suggested analyses overview
- Using the workbook

What Are Air Toxics? (1 of 2)

- The Clean Air Act Amendments of 1990 define 188 hazardous air pollutants (HAPs).
  - The two terms “HAPs” and “air toxics” are used interchangeably.
- Air toxics are those pollutants known or suspected to cause cancer or other serious health effects, such as reproductive effects or birth defects.
- Examples of toxic air pollutants include
  - benzene (found in gasoline),
  - perchloroethylene (emitted from some dry cleaning facilities),
  - methylene chloride (solvent and paint stripper),
  - arsenic, mercury, chromium, and lead compounds (e.g., metal processing operations), and
  - semivolatile organic compounds (SVOCs) such as naphthalene (petroleum refining and fossil fuel and wood combustion).

What Are Air Toxics? (2 of 2)

- Most air toxics originate from anthropogenic sources, including
  - mobile sources (e.g., cars, trucks, buses),
  - stationary sources (e.g., factories, refineries, power plants), and
  - indoor sources (e.g., some building materials and cleaning solvents).
- Some air toxics are emitted by natural sources (e.g., volcanic eruptions and forest fires).
- EPA is working with state, local, and tribal governments to reduce air toxics releases to the environment.

List of 188 Hazardous Air Pollutants

1,1,2,2-Tetrachloroethane	**Cobalt (Tsp)**	**Vinyl Chloride**	Mercury (Pm10) Stp	*Acrylamide*	*Hydrochloric acid*
1,1,2-Trichloroethane	**Cobalt Pm2.5 Lc**	1,2-Dibromo-3-Chloropropane	Mercury (Vapor)	*Acrylic acid*	*Hydrogen fluoride*
1,1-Dichloroethane	**Dichloromethane**	1,3-Dichloropropene(Total)	Mercury Pm10 Lc	*Asbestos*	*Hydrogen sulfide*
1,1-Dichloroethylene	**Ethyl Acrylate**	1,4-Dioxane	Methanol	*Benzidine*	*Hydroquinone*
1,2,4-Trichlorobenzene	**Ethylbenzene**	2,4,5-Trichlorophenol	Methoxychlor	*Benzotrichloride*	*Maleic anhydride*
1,2-Dichloropropane	**Ethylene Dibromide**	2,4,6-Trichlorophenol	M-Xylene	*beta-Propiolactone*	*m-Cresol*
1,3-Butadiene	**Ethylene Dichloride**	2,4-Dinitrophenol	Nickel (Coarse Particulate)	*Bis(chloromethyl)ether*	*Methyl hydrazine*
1,4-Dichlorobenzene	**Formaldehyde**	2,4-Dinitrotoluene	Nickel Pm10 Lc	*Calcium cyanamide*	*Methyl iodide (Iodomethane)*
2,2,4-Trimethylpentane	**Hexachlorobutadiene**	3-Chloropropene	Nitrobenzene	*Captan*	*Methyl isocyanate*
Acetaldehyde	**Isopropylbenzene**	4,6-Dinitro-2-Methylphenol	O-Cresol	*Carbaryl*	*Methylene diphenyl diisocyanate*
Acetonitrile	**Lead (Pm10) Stp**	4-Nitrophenol	P-Cresol	*Carbonyl sulfide*	*N,N-Diethyl aniline*
Acrolein	**Lead (Tsp)**	Aniline	Pentachlorophenol	*Catechol*	*N-Nitrosodimethylamine*
Acrylonitrile	**Lead Pm2.5 Lc**	Antimony (Pm10) Stp	Phenol	*Chloramben*	*N-Nitrosomorpholine*
Antimony (Tsp)	**M/P-Xylene**	Antimony Pm10 Lc	Phosphorus (Tsp)	*Chlordane*	*N-Nitroso-N-methylurea*
Antimony Pm2.5 Lc	**Manganese (Pm10) Stp**	Arsenic Pm10 Lc	Phosphorus Pm10 Lc	*Chloroacetic acid*	*o-Anisidine*
Arsenic (Pm10) Stp	**Manganese (Tsp)**	Beryllium Pm10 Lc	P-Xylene	*Chlorobenzilate*	*o-Toluidine*
Arsenic (Tsp)	**Manganese Pm2.5 Lc**	Biphenyl	Selenium Pm10 Lc	*Chloromethyl methyl ether*	*Parathion*
Arsenic Pm2.5 Lc	**Mercury (Tsp)**	Bis (2-Chloroethyl)Ether	Xylene(S)	*Coke Oven Emissions*	*Pentachloronitrobenzene*
Benzene	**Mercury Pm2.5 Lc**	Bis(2-Ethylhexyl)Phthalate	*1,1-Dimethyl hydrazine*	*Cresols/Cresylic acid*	*Phosgene*
Benzyl Chloride	**Methyl Chloroform**	Cadmium Pm10 Lc	*1,2-Diphenylhydrazine*	*Cyanide Compounds*	*Phosphine*
Beryllium (Pm10) Stp	**Methyl Isobutyl Ketone**	Caprolactam	*1,2-Epoxybutane*	*DDE*	*Phthalic anhydride*
Beryllium (Tsp)	**Methyl Methacrylate**	Chlorine (Tsp)	*1,2-Propylenimine*	*Diazomethane*	*Polychlorinated biphenyls*
Bromoform	**Methyl Tert-Butyl Ether**	Chlorine Pm10 Lc	*1,3-Propane sultone*	*Dichlorvos*	*Polycylic Organic Matter*
Bromomethane	**Naphthalene**	Chromium (Coarse Particulate)	*2,3,7,8-Tetrachlorodibenzo-p-dioxin*	*Diethanolamine*	*p-Phenylenediamine*
Cadmium (Pm10) Stp	**N-Hexane**	Chromium Pm10 Lc	*2,4-D, salts and esters*	*Diethyl sulfate*	*Propoxur (Baygon)*
Cadmium (Tsp)	**Nickel (Pm10) Stp**	Cobalt Pm10 Lc	*2,4-Toluene diamine*	*Dimethyl aminoazobenzene*	*Propylene oxide*
Cadmium Pm2.5 Lc	**Nickel (Tsp)**	Dibenzofurans	*2,4-Toluene diisocyanate*	*Dimethyl carbamoyl chloride*	*Quinoline*
Carbon Disulfide	**Nickel Pm2.5 Lc**	Dimethyl Phthalate	*2-Acetylaminofluorene*	*Dimethyl formamide*	*Quinone*
Carbon Tetrachloride	**O-Xylene**	Di-N-Butyl Phthalate	*2-Chloroacetophenone*	*Dimethyl sulfate*	*Radionuclides (including radon)*
Chlorine Pm2.5 Lc	**Phosphorus Pm2.5 Lc**	Ethylene Oxide	*2-Nitropropane*	*Epichlorohydrin*	*Styrene oxide*
Chlorobenzene	**Propionaldehyde**	Heptachlor	*3,3-Dichlorobenzidene*	*Ethyl carbamate (Urethane)*	*Titanium tetrachloride*
Chloroethane	**Selenium (Pm10) Stp**	Hexachlorobenzene	*3,3-Dimethoxybenzidine*	*Ethylene glycol*	*Toxaphene*
Chloroform	**Selenium (Tsp)**	Hexachlorocyclopentadiene	*3,3'-Dimethyl benzidine*	*Ethylene imine (Aziridine)*	*Triethylamine*
Chloromethane	**Selenium Pm2.5 Lc**	Hexachloroethane	*4,4-Methylene bis(2-chloroaniline)*	*Ethylene thiourea*	*Trifluralin*
Chloroprene	**Styrene**	Isophorone	*4,4-Methylenedianiline*	*Fine mineral fibers*	*Vinyl bromide*
Chromium (Pm10) Stp	**Tetrachloroethylene**	Lead Pm10 Lc	*4-Aminobiphenyl*	*Glycol ethers*	
Chromium (Tsp)	**Toluene**	Lindane	*4-Nitrobiphenyl*	*Hexamethylene-1,6-diisocyanate*	
Chromium Pm2.5 Lc	**Trichloroethylene**	Manganese (Coarse Particulate)	*Acetamide*	*Hexamethylphosphoramide*	
Cobalt (Pm10) Stp	**Vinyl Acetate**	Manganese Pm10 Lc	*Acetophenone*	*Hydrazine*	

Abundance of data: > 20 monitoring sites with sufficient data to create a valid annual average between 2003-2005, up to 434 sites

Little data: < 20 monitoring sites with sufficient data to create a valid annual average between 2003-2005, between 1-17 sites

No Data: No valid annual averages between 2003-2005

From: http://www.epa.gov/ttn/atw/188polls.html

Why Analyze Ambient Air Toxics Data?

- Air toxics data analysis is needed to track progress in risk reduction.
- States collecting data have unique “local” perspectives on data quality, meteorology, and sources, and in articulating policy-relevant data analysis questions.
  - Data anomalies at an individual site have little influence on the overall national-scale results.
  - On a site-by-site basis, a fine level of detail is needed to understand the characteristics and trends observed.

Types of Questions Analysts May Want to Consider

- How do I ensure that the data I plan to use for analysis are of good quality?
  - *Preparing Data for Analysis*, Section 4
- How do air toxics concentrations change spatially and by time of day, day of week, and season?
  - *Characterizing Air Toxics*, Section 5 and *Background*, Section 3
- What are the most important air toxics in terms of potential risk?
  - *Advanced Analyses*, Section 7
- How do concentration levels for a given city/area compare to other cities?
  - *Characterizing Air Toxics*, Section 5
- Have air toxics concentrations declined over time in response to emission control programs?
  - *Quantifying and Interpreting Trends in Air Toxics*, Section 8
- How do the most important air toxics compare with model output (e.g., are ambient concentrations high in locations not shown by the model)?
  - *Characterizing Air Toxics*, Section 5

Suggested Analysis

Overview (1 of 2)

- A list of suggested air toxics data analyses provides direction on those analyses that may be performed by air toxics monitoring agencies and gives an overview of analyses covered in the workbook.
- EPA compiled this list based on analyses that would help regional, state, and local organizations determine which factors contribute to air toxics concentrations in their area and whether the control strategies they have implemented have been successful at reducing these pollutants.

Suggested Analysis

Overview (2 of 2)

- These suggested analyses aid the understanding of an area's air toxics concentrations.
  - Are data of sufficient quality for analysis?
  - How would air toxics be characterized in the area?
  - What are local sources of air toxics?
  - Do toxics concentrations change over time?
- For the most informative results, consider performing some of these analyses annually.
- EPA-funded reports will be placed on an air toxics website for all to share.

Suggested Analyses (1 of 4)

Questions	Example Analyses
Are data of sufficient quality for analysis?	
How have data been validated?	Run screening checks on data from AQS; identify outliers
Does suspect data quality appear in any years or species measurements?	Review collocated data; inspect summary statistics and concentration ranges; review time series plots of concentrations and detection limits
Have data been censored?	Assess concentration distributions; compare concentrations to detection limits
Are sufficient samples available for detailed analyses?	Determine number of samples/species with concentrations above detection

Suggested Analyses (2 of 4)

Questions	Example Analyses
What is the nature and extent of air toxics problems in your area?	
What are the most abundant air toxics at each site on a risk-weighted basis?	Determine median concentrations and concentration ranges and compare to appropriate risk levels
How do these species vary by measurement season, month, and time of day? Are findings consistent with national level results?	Prepare box plots of concentrations by season, month, and time of day; compare to national results and expectations based on local conditions
Do species show any day-of-week patterns?	Prepare box plots of concentrations by day of week; compare results to expected patterns of local emissions
How do concentrations compare to other locations, risk levels, remote background, or reference concentrations?	Compare monitor-level data to national-perspective plots

Suggested Analyses (3 of 4)

Questions	Example Analyses
What are local sources of air toxics?	
What are the potential toxics sources in the area?	Investigate Google map of area; overlay VOC, $PM_{2.5}$, and air toxics emission inventory information
Do the air toxics corroborate the source mixture?	• Examine key species noted as tracers for the expected sources in the area using scatter plots and correlation matrices • Compare concentrations of air toxics and nontoxic tracer species to further assess sources (e.g., $PM_{2.5}$ components, hydrocarbons)

Suggested Analyses (4 of 4)

Questions	Example Analyses
Do air toxics concentrations change over time?	
What are the annual trends in air toxics concentrations?	Prepare annual box plots of key species to evaluate trends
How might changes in air toxics concentrations be related to emissions controls?	• Compare trends in co-emitted pollutants • Assess timing of controls and expected reductions relevant to local monitoring of pollutants.

Using the Workbook

- This workbook documents methodology used in national-scale analyses, extends these methodologies to possible use in local-scale analyses, and suggests methodology for further exploration.
- Examples are provided from the national-scale analyses and some analyses were custom-designed for the workbook.
- Space available in the workbook is limited; therefore, many details are, of necessity, provided in the literature. A reference section is provided at the end of each chapter.

References

- Agency for Toxic Substances and Disease Registry (ASTDR) (2007) Frequently asked questions about contaminants found at hazardous waste sites. Available on the Internet at http://www.atsdr.cdc.gov/toxfaq.html.
- U.S. Environmental Protection Agency, FERA (Fate, Exposure and Risk Analysis) Risk Assessment and Modeling web site. Available on the Internet at http://www.epa.gov/ttn/fera/risk_atoxic.html
- U.S. Environmental Protection Agency (2007a) EPA air toxics web site. Available on the Internet at http://www.epa.gov/ttn/atw/allabout.html
- U.S. Environmental Protection Agency (2007b) About air toxics, health and ecological effects. Available on the Internet at http://www.epa.gov/air/toxicair/newtoxics.html.

Definitions and Acronyms

This section lists definitions of terms and acronyms used in this workbook.

Definitions and Acronyms (1 of 10)

Aerosol A particle of solid and/or liquid matter that can remain suspended in the air because of its small size (generally under one micron).

AIRNow The U.S. EPA, NOAA, tribal, state, and local agencies developed the AIRNow web site to provide the public with easy access to national air quality information. The web site offers daily air quality index (AQI) forecasts as well as real-time AQI conditions for over 300 cities across the United States, and provides links to more detailed state and local air quality web sites <http://airnow.gov/>.

Airshed A geographic area that, because of topography, meteorology, and/or climate, is frequently affected by the same air mass.

AQS Air Quality System; the EPA's repository of ambient air quality data http://www.epa.gov/ttn/airs/airsaqs/.

Anthropogenic Caused or produced by human activities.

Anthropogenic emissions Emissions from man-made sources as opposed to natural (biogenic) sources**.**

Back trajectory A trace backwards in time showing where an air mass has been.

Black Carbon (BC) Black carbon measured using light absorption, typically with an Aethalometer™. Used in the air toxics monitoring network as a potential surrogate measure (although not unique or quantitative) of diesel particulate matter.

Cancer benchmark A potential regulatory threshold concentration of concern related to long term exposure to a chemical associated with increased cancer risk.

Cd Cadmium.

Censored Data The measured value is replaced with a proxy: Typical examples are MDL, MDL/2, MDL/10, or zero.

Definitions and Acronyms (2 of 10)

Census tract Census tracts are small, relatively permanent statistical subdivisions of a county. Census tracts are delineated for most metropolitan areas (MAs) and other densely populated counties by local census statistical areas committees following Census Bureau guidelines (more than 3,000 census tracts have been established in 221 counties outside MA's). Six states (California, Connecticut, Delaware, Hawaii, New Jersey, and Rhode Island) and the District of Columbia are covered entirely by census tracts. Census tracts usually represent between 2,500 and 8,000 people and, when first delineated, are designed to be homogeneous with respect to population characteristics, economic status, and living conditions. Census tracts do not cross county boundaries. The spatial size of census tracts varies widely depending on the density of settlement <http://www.census.gov/geo/www/cen_tract.html>.

Cluster analysis A multivariate procedure for grouping data by similarity among samples (i.e., samples with similar chemical compound concentrations are grouped).

CMAQ Community Multiscale Air Quality system. An air quality simulation model of tropospheric ozone, acid deposition, visibility, and fine particulate matter from urban to regional scales.

CMB Chemical mass balance model. A receptor model.

Coefficient of Correlation, r A statistic representing how closely two variables co-vary; they can vary from -1 (perfect negative correlation) through 0 (no correlation) to +1 (perfect positive correlation).

Collinearity A situation in which a near-perfect linear relationship exists among some or all of the independent variables in a regression model; in practical terms, there is some degree of redundancy or overlap among the variables.

Definitions and Acronyms (3 of 10)

Conditional probability function (CPF) A method that analyzes local source impacts from varying wind directions using the source contribution estimates from PMF coupled with the corresponding wind directions.

Confidence Interval (CI) CI for a population parameter is an interval with an associated probability p that is generated from a random sample of an underlying population such that if the sampling was repeated numerous times and the confidence interval recalculated from each sample according to the same method, a proportion p of the confidence intervals would contain the population parameter in question.

Covariance A statistical measure of correlation of the fluctuations of two different quantities.

Cr Chromium.

CSN Chemical Speciation Network

Dispersion model A source-oriented approach in which a pollutant emission rate and meteorological information are input into a mathematical model that disperses (and may also chemically transform) the emitted pollutant, generating a prediction of the resulting pollutant concentration at a point in space and time.

DL – Detection limit (see method detection limit).

DPM Diesel particulate matter.

Edge A line that defines the boundary of the relationship between two parameters on a scatter plot.

Elemental carbon (EC) Black carbon material with little or no hydrogen; non-volatile carbon material; often called black carbon or soot.

Emission Inventory (EI) A list of air pollutants emitted into a community's atmosphere in amounts (commonly tons) per day or year, by type of source.

EPA U.S. Environmental Protection Agency.

EPA PMF A standalone version of PMF created by the EPA in 2005.

Environmental justice The fair treatment and meaningful involvement of all people regardless of race, color, national origin, or income with respect to the development, implementation, and enforcement of environmental laws, regulations, and policies.

Definitions and Acronyms (4 of 10)

F-test The F-test provides a statistical measure of the confidence that a relationship exists between the two variables (i.e., the regression line does not have a slope of zero, which would indicate the dependent variable is not related to the independent variable).

F-value Output of the F-test. Large F-values indicate a stronger correlation between the two variables (i.e., the slope of the regression line is NOT zero).

Factor analysis A procedure for grouping data by similarity among variables (i.e., variables that are highly correlated are grouped).

Factor strength (source strength) See Source contribution.

Federal Reference Method (FRM) Provides for the measurement of the mass concentration of fine particulate matter having an aerodynamic diameter less than or equal to a nominal 2.5 microns ($PM_{2.5}$) in ambient air over a 24-hr period for purposes of determining whether the primary and secondary National Ambient Air Quality Standards for fine particulate matter are met. Designation of a particle sampler as a Federal Reference Method (FRM) is based on a demonstration that a vendor's instrument meets the design specifications, performance requirements, and quality control standards specified in the regulation.

Fine particles Particulate matter with diameter less than 2.5 microns; $PM_{2.5}$.

HAPs (hazardous air pollutants) Hazardous air pollutants, also known as air toxics, have been associated with a number of adverse human health effects, including cancers, asthma and other respiratory ailments, and neurological problems such as learning disabilities and hyperactivity.

HYSPLIT HYbrid Single-Particle Lagrangian Integrated Trajectory model; a system for computing simple air parcel trajectories <http://www.arl.noaa.gov/ready/hysplit4.html>.

IMPROVE Interagency Monitoring of Protected Visual Environments. A collaborative monitoring program to establish present visibility levels and trends, and to identify sources of man-made impairment <http://vista.cira.colostate.edu/improve/Default.htm>.

Interquartile range The difference between the 75th and 25th percentiles of a data set.

Definitions and Acronyms (5 of 10)

Level 0 validation Routine checks made during the initial data processing and generation of data, including proper data file identification, review of unusual events, review of field data sheets and result reports, instrument performance checks, and deterministic relationships.

Level I validation Tests for internal consistency to identify values in the data that appear atypical when compared to values of the entire data set.

Level II validation Comparison of the current data set with historical data to verify consistency over time. This level can be considered a part of the data interpretation or analysis process.

Level III validation Tests for parallel consistency with data sets from the same population (i.e., region, period of time, air mass, etc.) to identify systematic bias. This level can also be considered a part of the data interpretation or analysis process.

LC Local conditions; refers to ambient PM measurements.

MACT Maximum achievable control technology. MACTs are technology-based air emission standards established under Title III of the 1990 Clean Air Act Amendments <http://www.epa.gov/region08/compliance/mact/mact.html>.

Mean The sum of all values divided by the number of samples.

Median The middle value in a sorted list of samples if there is an odd number of samples, or the average of the two middle values if there is an even number of samples.

Method Detection Limit (MDL) 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 the analysis of a sample in a given matrix containing the analyte

Mobile sources Motor vehicles and other moving objects that release pollution; mobile sources include cars, trucks, buses, planes, trains, motorcycles, and gasoline-powered lawn mowers. Mobile sources are divided into two groups: road vehicles, which include cars, trucks, and buses, and non-road vehicles, which include trains, planes, and lawn mowers.

Definitions and Acronyms (6 of 10)

National Ambient Air Quality Standards (NAAQS) Health-based pollutant concentration limits established by the EPA that apply to outside air.

NATA National air toxics assessment <http://www.epa.gov/ttn/atw/nata1999/nsata99.html>. EPA's national-scale assessment of 1999 air toxics emissions. The purpose of the national-scale assessment is to identify and prioritize air toxics, emission source types and locations that are of greatest potential concern in terms of contributing to population risk.

NATTS National air toxics trends stations <http://www.epa.gov/ttn/amtic/natts.html>.

NEI National emissions inventory <http://www.epa.gov/ttn/chief/net/>.

NOAA National Oceanic and Atmospheric Administration.

NWS National Weather Service.

OH Hydroxyl radical; the driving force behind the daytime reactions of hydrocarbons in the troposphere.

O_3 Ozone; a major component of smog. Ozone is not emitted directly into the air but is formed by the reaction of VOCs and NO_x in the presence of heat and sunlight.

Organic carbon (OC) Consists of hundreds of separate semi-volatile and particulate compounds.

Outliers Data physically, spatially, or temporally inconsistent.

P-value Provides a measure of the percentage confidence that the slope is not zero: % confidence slope is not zero = 100%(1 – P). Generally, 95% confidence is used as a cutoff value, corresponding to a P-value of 0.05.

PAMS Photochemical Assessment Monitoring Stations <http://epa.gov/air/oaqps/pams/freqfile.html>.

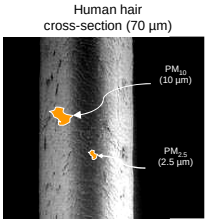
Particulate matter (PM) A generic term referring to liquid and/or solid particles suspended in the air.

Definitions and Acronyms (7 of 10)

Percentile The *p*th percentile of a data set is the number such that *p*% of the data is less than that number.

$PM_{2.5}$ Particulate matter less than 2.5 microns. Tiny solid and/or liquid particles, generally soot and aerosols. The size of the particles (2.5 microns or smaller, about 0.0001 inches or less) allows them to easily enter the air sacs deep in the lungs where they may cause adverse health effects; $PM_{2.5}$ also causes visibility reduction.

PM_{10} Particulate matter less than 10 microns. Tiny solid and/or liquid particles of soot, dust, smoke, fumes, and aerosols. The size of the particles (10 microns or smaller, about 0.0004 inches or less) allows them to easily enter the air sacs in the lungs where they may be deposited, resulting in adverse health effects. PM_{10} also causes visibility reduction and is a criteria air pollutant.



PMF Positive matrix factorization; a receptor model. PMF can be used to determine source profiles and source contributions based on the ambient data.

POC Pollutant occurrence code used in the AQS.

Definitions and Acronyms (8 of 10)

Point source Point sources include industrial and nonindustrial stationary equipment or processes considered significant sources of air pollution emissions. A facility is considered to have significant emissions if it emits about one ton or more in a calendar year. Examples of point sources include industrial and commercial boilers, electric utility boilers, turbine engines, industrial surface coating facilities, refinery and chemical processing operations, and petroleum storage tanks.

Potential Source Contribution Function (PSCF) A method that combines the source contribution estimates from PMF with the air parcel backward trajectories to identify possible source areas and pathways that give rise to the observed high particulate mass concentrations from the potential sources.

Precursor Compounds that change chemically or physically after being emitted into the air and eventually produce air pollutants. For example, sulfur and nitrogen oxides are precursors for particulate matter.

Primary particles The fraction of PM_{10} and $PM_{2.5}$ that is directly emitted from combustion and fugitive dust sources.

QA Quality assurance; a set of external tasks to provide certainty that the quality control system is satisfactory. These tasks include independent performance audits, on-site system audits, interlaboratory comparisons, and periodic evaluations of internal quality control data.

QC Quality control; a set of internal tasks performed to provide accurate and precise measured ambient air quality data. These tasks address sample collection, handling, analysis, and reporting (e.g., periodic calibrations, routine service checks, instrument-specific monthly quality control maintenance checks, and duplicate analyses on split and spiked samples).

R-squared, r^2 Statistical measure of how well a regression line approximates real data points; an r^2 of 1.0 (100%) indicates a perfect fit.

Definitions and Acronyms (9 of 10)

Receptor model A receptor-oriented approach for identifying and quantifying the sources of ambient air contaminants at a receptor primarily on the basis of concentration measurements at that receptor.

Reference Concentration (RfC) An estimate (with uncertainty of perhaps an order of magnitude) of a continuous inhalation exposure to the human population (including sensitive subgroups) that is likely to be without an appreciable risk of deleterious effects during a lifetime.

Reid Vapor Pressure (RVP) A measure of gasoline volatility.

RFG Reformulated gasoline.

Residuals Measured concentrations minus modeled concentrations.

SEARCH SouthEastern Aerosol Research and Characterization Study.

Secondary formation The fraction of a pollutant that is formed in the atmosphere (e.g., formaldehyde is both emitted directly and formed in the atmosphere through secondary photochemical processes).

Selected ion monitoring (SIM) A mass spectral mode in which the mass spectrometer is set to scan over a very small mass range, typically one mass unit, providing higher sensitivity results than a full mass scan.

Slope Statistical measure of the average ratio of the predicted to measured concentrations of a species; a slope closer to 1.0 demonstrates a closer fit.

Source apportionment The process of apportioning ambient pollutants to an emissions source. Also known as source attribution.

Source contribution Total mass of material from a source measured in a sample.

Source-dispersion model See Dispersion model.

Source profile Listing of individual chemical species emitted by a specific source category.

Speciation Trends Network (STN) A network of sampling locations established by the EPA in 2001 to characterize $PM_{2.5}$ composition in urban areas. Roughly 300 sites nationwide are part of this network.

Definitions and Acronyms (10 of 10)

Standard Deviation A measure of how much the average varies. The square root of the average squared deviation of the observations from their mean.

Standard operating procedure (SOP) A set of instructions used to ensure data quality.

Standardized residual Ratio of the residual to the uncertainty of a species in a specific sample determined by the user.

State implementation plan (SIP) A detailed description of the programs a state will use to carry out its responsibilities under the Clean Air Act. State implementation plans are collections of the regulations used by a state to reduce air pollution. The Clean Air Act requires that the EPA approve each state implementation plan.

SVOC Semi-volatile organic compound.

TRI Toxic Release Inventory. Publicly available EPA database that contains information about toxic chemical releases and other waste management activities reported annually by certain covered industry groups as well as federal facilities <http://www.epa.gov/tri/index.htm>.

TSP Total suspended particulate.

Uncensored data Data reported "as is" with no substitution for values below detection.

Variance The square of the standard deviation.

VOC Volatile organic compound.

WD Wind direction.

WS Wind speed.

XRF Energy dispersive X-ray fluorescence. Method used to quantify particulate metals.

References

Bay Area Air Quality Management District (2005) Air quality glossary. Available on the Internet at <http://www.baaqmd.gov/dst/glossary.htm>.

California Air Resources Board (2003) Glossary of air pollution terms. Available on the Internet at <http://arbis.arb.ca.gov/html/gloss.htm>.

Minnesota Pollution Control Agency (2005) General glossary. Available on the Internet at <http://www.pca.state.mn.us/gloss/>.

National Park Service (2005) Glossary of terms used by the NPS Inventory and Monitoring Program. Available on the Internet at <http://science.nature.nps.gov/im//monitor/glossary.htm>.

Sam Houston State University (2005) Atmospheric chemistry glossary. Web site prepared by Sam Houston State University, Department of Chemistry, Huntsville, TX, by the Department of Chemistry. Available on the Internet at <http://www.shsu.edu/~chemistry/Glossary/glos.html>.

U.S. Environmental Protection Agency (2002) The plain English guide to the Clean Air Act: Glossary. Available on the Internet at <http://www.epa.gov/oar/oaqps/peg_caa/pegcaa10.html#topic10>.

U.S. Environmental Protection Agency (2005) AIRTrends 1997 report: list of acronyms. Available on the Internet at <http://www.epa.gov/air/airtrends/aqtrnd97/acron.html>.

Background

What are air toxics and why are they important?

Background

What's Covered in This Section?

- Air toxics overview
- Health risks from air toxics; terminology
- Air toxics emissions
- Physical properties
- Formation, destruction, and transport of air toxics
- History of sampling; objectives of air toxics and other monitoring programs
- Air toxics sampling and analysis
- Critical issues for data interpretation

Air Toxics

Overview (1 of 3)

- ***What are air toxics?***
  - Air toxics are gaseous, aerosol, or particle pollutants present in the air in varying concentrations with characteristics such as toxicity or persistence that can be hazardous to human, plant, or animal life.
  - The terms “air toxics” and “hazardous air pollutants” (HAPs) are used interchangeably in this document.
  - Air toxics include the following general categories of compounds: volatile and semi-volatile organic compounds (VOCs, SVOCs), polycyclic aromatic hydrocarbons (PAHs), heavy metals, and carbonyl compounds.

U.S. Environmental Protection Agency (2007c, g)

Air Toxics

Overview (2 of 3)

- ***What are the health and environmental effects of toxic air pollutants?***
  - People exposed to toxic air pollutants at sufficient concentrations and durations may have an increased chance of getting cancer or experiencing other serious health effects.
  - Both high values and annual means of air toxics concentrations are of interest because some air toxics have both episodic, short-term health effects and chronic, long-term health effects.
  - Other health effects can include damage to the immune system, as well as neurological, reproductive (e.g., reduced fertility), developmental, respiratory, and other health problems.
  - Some toxic air pollutants, such as mercury, can deposit onto soils or surface waters where they are taken up by plants and ingested by animals and are eventually magnified up through the food chain.
  - Animals may experience health problems if exposed to sufficient quantities of air toxics over time.

U.S. Environmental Protection Agency (2007c, g)

Air Toxics

Overview (3 of 3)

- ***How are people exposed to air toxics?***
  - Breathing contaminated air.
  - Eating contaminated food products, such as fish from contaminated waters; meat, milk, or eggs from animals that feed on contaminated plants; and fruits and vegetables grown in contaminated soil on which air toxics have been deposited.
  - Drinking water contaminated by toxic air pollutants.
  - Ingesting contaminated soil.
  - Touching contaminated soil, dust, or water.
  - Accumulating some persistent toxic air pollutants in body tissues after toxic air pollutants have entered the body. Predators typically accumulate even greater pollutant concentrations than their contaminated prey. As a result, people and other animals at the top of the food chain who eat contaminated fish or meat are exposed to concentrations that are much higher than the concentrations in the water, air, or soil.

U.S. Environmental Protection Agency (2007c, g)

Health Risks from Air Toxics

- Simply put, health risks are a measure of the chance that you will experience health problems.

  Health risk = Hazard x exposure

- Health risk is the probability that exposure to a hazardous substance will make you sick.
- Exposure to toxic air pollutants can increase your health risks.
- Ambient concentrations of air toxics are compared to chronic exposure risk levels derived from scientific assessments conducted by the EPA and other environmental agencies.

U.S. Environmental Protection Agency (2007a, b)

Air Toxics Emissions

What Are the Sources of Air Toxics?

- Air toxics are both directly emitted by sources and formed in the atmosphere.
- **Major sources** include chemical plants, steel mills, oil refineries, and hazardous waste incinerators for which there is a specific location provided in the inventory.
- **Area sources** are made up of many smaller sources releasing pollutants to the outdoor air in a defined area.
- **Mobile sources** include highway vehicles, trains, marine vessels, and non-road equipment (such as construction equipment).
- **Natural sources** – Some air toxics are also released from natural sources such as volcanoes or fires; typically in the inventory these would be included in area source emissions.

Air Toxics Emissions

Source Type Characteristics

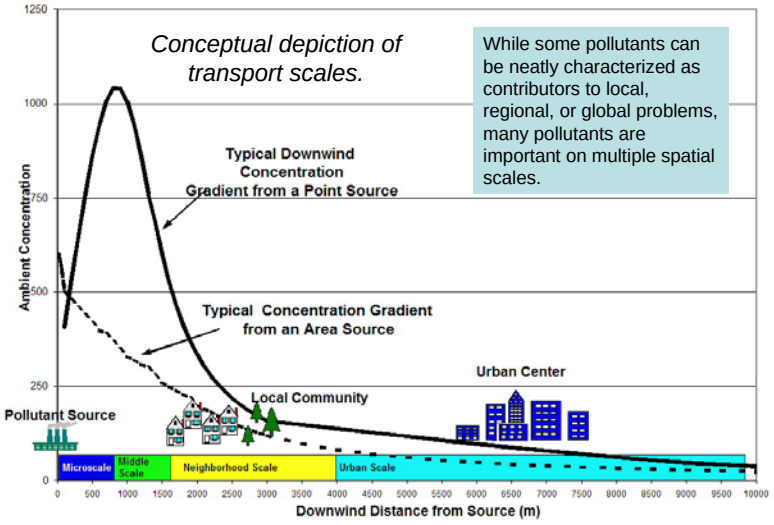
Understanding the emission source type of a particular air toxic can help the analyst begin to develop a conceptual model of concentration patterns and gradients that might be expected.

- Major source emissions, for example, are a localized source of toxics and may show steep concentration gradients.
- Area source emissions are typically well-distributed emissions sources because there are multiple sources in an area.
- Mobile source air toxics exhibit both point source and area source characteristics.

Physical Properties

- Physical properties of air toxics span the entire range of pollutants present in the atmosphere
  - As particles and gases and in semi-volatile form.
  - As both primary (directly emitted) and secondary (formed in the atmosphere).
  - From mostly anthropogenic sources, but include some biogenic sources.
  - Have a wide range of atmospheric lifetimes.
- Some air toxics such as VOCs (e.g., benzene and toluene) are precursors to ozone and particulate matter (PM); and other toxics such as heavy metals are components of PM.

Formation, Destruction, Transport (1 of 2)



While some pollutants can be neatly characterized as contributors to local, regional, or global problems, many pollutants are important on multiple spatial scales.

Formation, Destruction, Transport (2 of 2)

- Concentrations of pollutants that are secondarily formed in the atmosphere
  - are often highest downwind of the source of precursor compounds
  - generally do not have steep concentration gradients near the original precursor emissions source
- Transport distance is determined by
  - atmospheric chemistry (pollutant lifetimes and formation and removal processes)
  - meteorology (air mass movement and precipitation)
  - topography (mountains and valleys that affect air movement)
- Short-lived pollutants can only travel short distances from where they are emitted (10s to 100s of miles). Longer-lived pollutants can travel large distances from where they are formed or emitted (e.g., toxic metals in $PM_{2.5}$) and may be more regionally homogenous.
- Some unreactive pollutants can remain in the atmosphere for months, years, or decades and spread across the Earth (e.g., carbon tetrachloride).

Residence Time

Overview

- Residence time is a pollutant-specific measure of the average lifetime of a molecule in the atmosphere.
- It is dependent on chemical and physical removal pathways that include
  - *Chemical:* reaction with hydroxyl radical (OH), photolysis
  - *Physical:* wet or dry deposition
- Why is it important to understand residence times?
  - Residence times can provide insight into the spatial and temporal variability of air toxics.
  - Longer residence times result in less spatial variability (e.g., carbon tetrachloride).
  - Conversely, short residence times should result in steep gradients in concentrations near sources and temporal patterns that are dependent on emissions schedules.
- Residence times are not characterized well for all air toxics. Some air toxics and their residence times are listed in the appendix to this section.

History of Sampling

- Air toxics measurements have been collected across the country since the 1960s as part of various programs and measurement studies.
- National monitoring efforts have included programs specific to air toxics:
  - National Air Toxics Trends Stations (NATTS)
  - Urban Air Toxics Monitoring Program (UATMP)
- Some ambient monitoring networks are designed for other purposes but also provide air toxics data:
  - Photochemical Assessment Monitoring Station (PAMS) program
  - Chemical Speciation Network (CSN) which includes the Speciation Trends Network (STN)
  - Interagency Monitoring of Protected Visual Environments (IMPROVE)
- State and local agencies have also operated long-running monitoring operations and special studies to understand air toxics in their communities.

NATTS Sampling

Overview

NATTS sampling began in 2003 with 23 sites; the first complete year of data was 2004. The principal objective of the NATTS network is to provide long-term monitoring data across representative areas of the country for certain priority HAPs (e.g., benzene, formaldehyde, 1,3-butadiene, acrolein, and hexavalent chromium) in order to establish national trends for these and other HAPs.

There are currently 27 national air toxics trends sites: 21 urban and 6 rural.

NATTS Sampling

Objectives

Primary objectives of NATTS monitoring

- Providing air toxics data of sufficient quality to identify trends, characterize ambient concentrations in representative areas, and evaluate air quality models.
- Providing tools and guidance that enable consistent, high certainty measurements.
- Using these consistent measurements to facilitate measuring progress towards national emission and risk reduction goals.
- Considering all NATTS sites to be NCORE level 2 sites, thereby providing rich data sets to address multi-pollutant issues. NCORE level 2 sites are “backbone” sites providing consistent, long-term data for multiple pollutant types.

Urban Air Toxics Monitoring Program (UATMP)

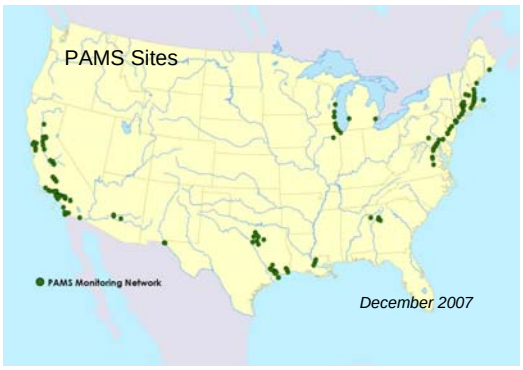
- The UATMP has provided sample collection and analysis support since 1987 to encourage state, local, and tribal agencies to understand and appreciate the nature and extent of potentially toxic air pollution in urban areas.
- Participation in the UATMP is voluntary; aside from the NATTS, target pollutants and monitor siting are at the discretion of each participant agency.
- UATMP assures analytical consistency among participants.

2007 UATMP Sites

U.S. Environmental Protection Agency (2006f)

PAMS Sampling

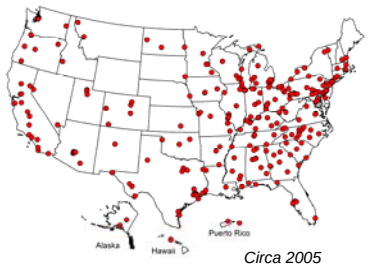
- The goal of the PAMS network is to help assess ozone control programs by
  - identifying key constituents and parameters;
  - tracking trends;
  - characterizing transport;
  - assisting in forecasting episodes; and
  - assisting in improving emission inventories.
- Toxic VOCs sampled by the PAMS network include benzene, formaldehyde, xylenes, toluene, ethylbenzene, styrene, and acetaldehyde.
- PAMS sites make subdaily measurements at the same sites that are useful in assessing diurnal trends.



U.S. Environmental Protection Agency (2006c)

CSN Sampling

- The Chemical Speciation Network is a companion network of the mass-based Federal Reference Method (FRM) network implemented to support the $PM_{2.5}$ National Ambient Air Quality Standards (NAAQS).
- The purpose of the CSN is to provide nationally consistent speciated $PM_{2.5}$ data to assess trends at representative sites in urban areas across the country.
- As part of a routine monitoring program, the CSN quantifies mass concentrations and $PM_{2.5}$ constituents, including numerous trace elements, ions (sulfate, nitrate, sodium, potassium, ammonium), elemental carbon, and organic carbon.
- CSN data are available via AQS.



Circa 2005

U.S. Environmental Protection Agency (2007f)

IMPROVE Sampling

- Interagency Monitoring of Protected Visual Environments (IMPROVE) program provides $PM_{2.5}$ speciated and mass measurements in 156 Class I areas (national parks and wildness areas). Speciated $PM_{2.5}$ metals are the only toxics measured in this network.
- Data are available in AQS.
- IMPROVE data can also be accessed from the VIEWS* web site.
- IMPROVE also provides site photos and local topographical maps which are very useful for data analyses.

IMPROVE Site Locations

*VIEWS: Visibility Exchange Web System

Local Scale Monitoring Projects

- EPA began programs to fund local-scale monitoring projects starting in the 2004 fiscal year.
- The goal of local monitoring is to provide more flexibility to address middle- and neighborhood-scale (0.5 km to 4 km) issues that are not handled well by national networks, given the diversity of toxics issues across the nation.
- Specific objectives include identifying and profiling air toxics sources, developing and assessing emerging measurement methods, characterizing the degree and extent of local air toxics problems, and tracking progress of air toxics reduction activities.
- Projects are selected through an open competition process. Grant topics, funding levels, and number of awards are set for each grant cycle.
- Local scale monitoring is typically only conducted from 1-2 years.

U.S. Environmental Protection Agency (2006c).

Air Toxics Sampling and Analysis (1 of 2)

- Because air toxics are present in the atmosphere in gaseous, particulate, and semi-volatile form, no single measurement technique is adequate.
- EPA offers 17 approved sampling and analysis methods for toxic gases; among the most commonly used methods are the following:
  - Compendium method TO-11A. Used to measure formaldehyde and other carbonyl compounds.
  - Compendium method TO-13A. Used to measure Polycyclic Aromatic Hydrocarbon (PAH) compounds.
  - Compendium method TO-15. Created to target 97 compounds on the list of 187 hazardous air pollutants.

Air Toxics Sampling and Analysis (2 of 2)

- EPA-approved methods for collection and analysis of suspended particulate matter are documented in the "Compendium of Methods for the Determination of Inorganic Compounds in Ambient Air."
  - Chapter IO-3, Chemical Species Analysis of Filter-Collected Suspended Particulate Matter (SPM), is of considerable importance to the air toxics ambient monitoring program.
  - Several different methods for speciated particulate analyses are available.
  - Each have advantages and disadvantages depending on the target analytes and desired minimum detection limits.
  - For Hazardous Air Pollutant (HAP) metals, IO-3.5 (Inductively Coupled Plasma / Mass Spectrometry (ICP/MS)) offers the lowest detection limits.

Differences Among Sampling Networks

- When using data from different sampling networks, it is important to consider
  - The multiple sampling networks from which data were drawn for these analyses vary in their objectives and sampling and analytical methods. *Data may not always be comparable.*
  - Sampling, analysis, method detection limits, objectives, site characteristics, etc. have changed over time. *Care is needed in interpreting temporal and spatial trends.*
- Analysts need to gather, and understand, all metadata prior to conducting analyses.

Critical Issues for Interpretation

- Issues to consider when planning and performing data analysis:
  - Data quality
  - Data availability
  - Sampling duration
  - Sampling frequency
  - Complementary data

Sampling Design

- To develop a sampling design or monitoring plan, the following should be considered:
  - Monitoring objectives, including consideration of geophysical setting, meteorology, types and characteristics of sources, and existing monitoring programs.
  - Data quality objectives needed to answer questions to be asked of the data (i.e., how precisely or accurately do the questions need to be answered?).
  - Options for what, when, where, how frequently, and for how long to monitor; these are related to the selection of appropriate monitoring equipment and laboratory analyses.
  - Data quality assurance and validation approach, including collocated data requirements, QA programs for analytical laboratories, and data validation guidelines for ambient data.
  - Options for data analysis and exploration, including available tools, data analyses, data needs, and training needs.

Resources

Monitoring Networks

- NATTS: http://www.epa.gov/ttn/amtic/natts.html
- UATMP: http://www.epa.gov/ttn/amtic/uatm.html
- PAMS: http://www.epa.gov/ttn/amtic/pamsmain.html
- CSN: http://www.epa.gov/ttn/amtic/speciepg.htm
- IMPROVE: A source of speciated $PM_{2.5}$ data http://vista.cira.colostate.edu/views/
- Local scale monitoring programs: http://www.epa.gov/ttn/amtic/local.html

Appendix
Residence Times

- Approximate atmospheric residence times for some air toxics are listed here.
- These values were found at http://www.scorecard.org/chemical-profiles/. To find the atmospheric persistence of other air toxics, enter the pollutant's name in the chemical profile. Once the pollutant page is available, select "links" and the entry for "CalEPA Air Resources Board Toxic Air Contaminant Summary". A summary of physical properties is provided including atmospheric persistence.

Species	Lifetime by reaction with OH
Carbon Tetrachloride	decades
Chloroform	months
Tetrachloroethylene	months
Methylene Chloride	months
Benzene	84 hrs
1,2-Dichloropropane	weeks*
Trichloroethylene	84 hrs
Acrylonitrile	2.4 days
Ethylbenzene	2 days
Vinyl Chloride	27 hrs
Formaldehyde	26 hrs
Acrolein	17 hrs
Naphthalene	16 hrs
Acetaldehyde	12 hrs
1,3-Butadiene	2.8 hrs
Arsenic and other toxic metal compounds	N/A**

* Wet deposition is also a sink
** Lifetime is dependant on particle deposition and is typically days to weeks. Deposition time is primarily determined by the size of the particles.

Preparing Data for Analysis

How do I get my data ready for analysis?
How do I treat data below detection?

Overview

- This section provides suggestions for acquiring and preparing data sets for analysis, laying the foundation for subsequent sections of the workbook.
- Data preparation is sometimes more difficult and time-consuming than the data analyses.
- It is vital to carefully construct a data set so that data quality and integrity are assured.
- Performing data validation is a start on data analysis.

Data Quality Objectives

- Preparation of data for subsequent analyses is tied to the data quality objectives (DQOs) to be achieved. A DQO is measurement performance or acceptance criteria established as part of the study design. DQOs relate the quality of data needed to the established limits on the chance of making a decision error or of incorrectly answering a study question.
- In setting DQOs, consider
  - who will use the data;
  - what the project's goals/objectives/questions or issues are;
  - what decision(s) will be made from the information obtained;
  - what type, quantity, and quality of data are specified;
  - how "good" the data have to be to support the decision to be made.
- EPA provides guidance on setting DQOs.

Preparing Data for Analysis

What's Covered in This Section?

- Data availability
  - What data are available?
  - Sources for ambient air toxics data
  - Accessing data systems and acquiring data
    - AQS
    - IMPROVE
    - SEARCH
    - Other archives
  - Supplementing air toxics data
  - Know your data
- Data processing
  - Investigating collocated data
  - Preparing daily, seasonal, and annual averages
  - Determining data completeness
  - Treating data below detection
- Data validation
  - Procedures and tools
  - Handling suspect data

What Data Are Available?

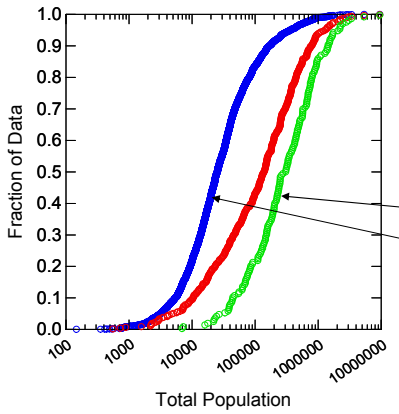
Air Toxics Overview (1 of 2)

- Air toxics ambient monitoring data is typically collected in three major durations: 1-hr, 3-hr, 24-hr.
- Sampling frequencies vary: subdaily, daily, 1-in-3-day, 1-in-6-day, 1-in-12-day.
- Some sites have operated as long-term (multiple-year) sites while others may report data for a short study only (e.g., a week or two).
- Data can be reported in a range of units, consistency is essential.
- For data to be useful, a minimum of monitor locations, concentration units, method codes, and parameter names is required.

What Data Are Available?

Air Toxics Overview (2 of 2)

Air toxics measurements are primarily captured in urban areas. VOC* measurements, for example, are typically made in higher population areas relative to all counties in the United States.



Blue = All U.S. Counties
Red = Counties with toxic metal measurements
Green = Counties with toxic VOC measurements

Note that counties with toxic VOC Measurements have higher population than "all counties"

* VOC: Volatile Organic Compound

What Data Are Available?

Sources for Ambient Air Toxics Data

- EPA's Air Quality System (AQS)
- IMPROVE speciated $PM_{2.5}$ data (VIEWS website)
- SEARCH speciated $PM_{2.5}$ data (Atmospheric Research Analysis website)
- Air Quality Archive (AQA) (1990-2005) developed during Phase V national air toxics analysis project; includes legacy air toxics archive data (data posted here http://www.epa.gov/ttn/amtic/toxdat.html)
- Local, state, and tribal air quality agency databases (i.e., some data are not yet submitted to AQS)

IMPROVE = Interagency Monitoring of Protected Visual Environments
VIEWS = Visibility Information Exchange Web System
SEARCH = SouthEastern Aerosol Research and Characterization Study

AQS Data

Overview

- AQS is the EPA's principal data repository, containing the most complete set of toxics (and other) data available.
  - AMP501 request provides raw data in R-2 format.
  - Data are available from 1995 to the present in AQS.
  - Annual air toxics data are required to be submitted to AQS within 180 days of end of Q4, i.e., 2007 data would be entered by July 2008.
  - Data from AQS are provided in a pipe-delimited format that needs to be transformed and processed.
- Some data, such as criteria pollutant summaries, are available for download without a user ID; most air toxics are not yet available this way.

AQS Data
Codes

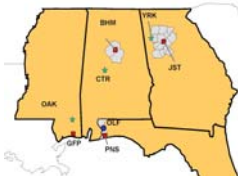
- Key Codes
  - AQS site code; identifies a particular monitoring site.
  - AQS parameter code; identifies the pollutant measured.
  - AQS parameter occurrence code (POC); distinguishes among monitors for the same pollutant at the same site.
  - AQS method code; unique for each combination of sample collection and analysis.
- Each code contains additional metadata which would be unnecessarily repetitive if included in the R-2 file.

  For example, default method detection limits (MDLs) are not provided in the R-2 file—this information must be looked up on the AQS website using the method query tool. Alternate MDLs are included in the R-2 file because they are unique to each record.

Other Data Archives (1 of 2)

- IMPROVE data – $PM_{2.5}$ speciated and mass measurements in 156 Class I areas (national parks and wildness areas). Speciated $PM_{2.5}$ metals are the only toxics measured in this network. Further described in Section 3, "Background".
- SEARCH data – $PM_{2.5}$ species and mass measurements at 8 sites in the Southeast from 1998 to the present. Speciated $PM_{2.5}$ metals are the only toxics measured in this network. At the time of the national analysis, these data were not available in AQS.
  - SEARCH data are publicly available via the Internet and can be downloaded on a site-by-site basis in a Microsoft Excel output format.
  - Site photographs and other useful metadata are available at the web site, http://www.atmospheric-research.com/newindex.html.

SEARCH Site Locations



Other Data Archives (2 of 2)

- As part of several projects, an air quality archive (AQA) was developed as an analysis-ready database that includes data from AQS (1990-2005), IMPROVE and SEARCH data, and data from the legacy air toxics archive.
- This database contains nearly 1 billion raw data records, 27 million raw toxics records, and complete validated and temporally aggregated data sets.
- Key data summaries have been posted http://www.epa.gov/ttn/amtic/toxdat.html:
  - 24-hour CSV Files (very large file)
  - Monthly CSV Files
  - Quarterly CSV Files
  - Annual Average CSV Files
  - SAS Files (all data, very large file)

Supplementing Air Toxics Data (1 of 2)

Supporting Information that Will Help in Data Interpretation

- Criteria pollutant species (AQS) – *multipollutant relationships, transport, diurnal/seasonal evaluation, source identification*
- Meteorological data (AQS, NWS) – *transport, mixing, source direction, meteorological adjustment of trends*
- All $PM_{2.5}$ speciation data (OC, EC, sulfate, nitrate, etc.) – *source identification*
- Aethalometer™ data (black carbon) – *diurnal characterization, source identification*
- All speciated hydrocarbon data (e.g., full PAMS target list) – *air parcel age (transport), source identification*
- Special studies data (e.g., continuous speciated PM data, ammonia) – *diurnal characteristics, source identification*

Supplementing Air Toxics Data (2 of 2)

Supporting Information that Will Help in Data Interpretation

- Monitoring objectives – *time-frame of data, reasoning for site locations*
- Site characteristics (e.g., photos) – *may explain data anomalies, source identification*
- Monitoring scale (likely varies by pollutant) – *air parcel age (transport), source identification*
- Emission inventory, especially point sources – *source identification*
- Population density – *relative concentration level*
- Vehicle traffic counts: *diurnal patterns, source identification*

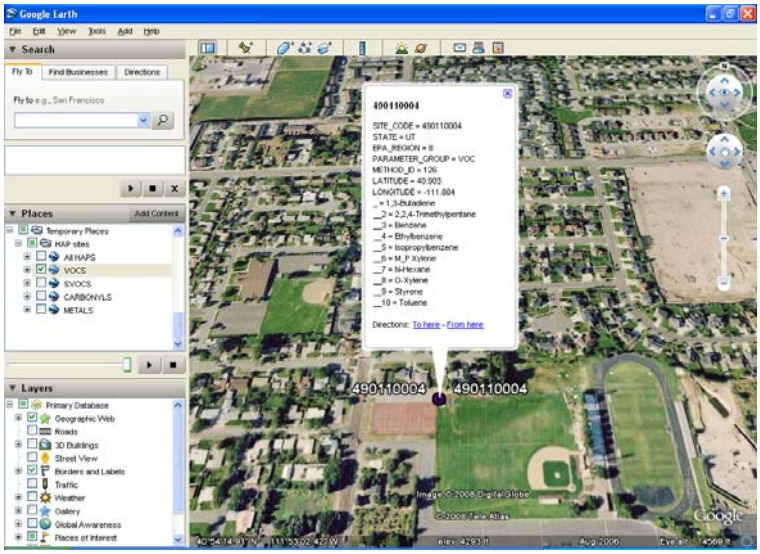
Supplementing Air Toxics Data (1 of 3)

Using Metadata

- Although some metadata are available through AQS, metadata are not routinely populated or updated.
- Site metadata can assist in analyses by illuminating sources (such as local sources or roadways) or physical attributes of the site.
- Satellite images can be obtained from Google Earth, a publicly available program that contains satellite coverage of the entire planet and is very useful to investigate monitor siting.
- Use caution when interpreting maps—reported precisions of monitor locations vary and not all significant sources will be easy to identify visually.

Google Earth File

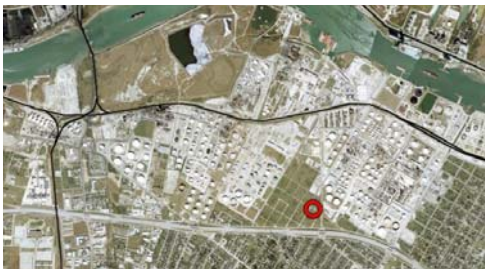
http://www.epa.gov/airexplorer/monitor_kml.htm

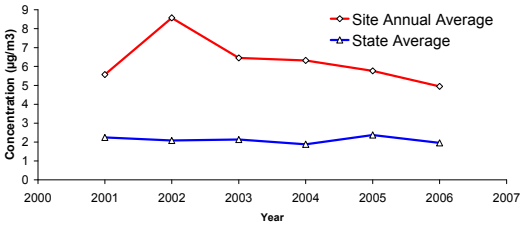


Supplementing Air Toxics Data (2 of 3)

Using Metadata

- The site (red circle) is near an oil refinery that will likely have a significant influence on VOC concentrations.
- The graph compares benzene annual averages at this site (red) to the state-wide annual average (blue). Benzene concentrations at this site are significantly increased.
- Preliminary evidence shows the refinery may be influencing local benzene concentrations.

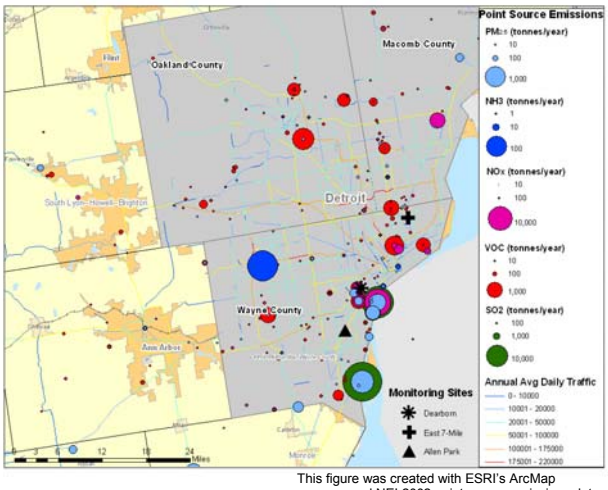




Supplementing Air Toxics Data (3 of 3)

Using Metadata

- Point source emissions of criteria pollutants and annual average daily traffic counts in the Detroit area near three monitoring sites.
- The Dearborn site is closest to major industry – higher concentrations of VOCs and $PM_{2.5}$ at this site could be explained by these sources.



This figure was created with ESRI's ArcMap program and NEI 2002 point source emissions data.

Converting Units (1 of 2)

- Frequently used units for gaseous air toxics include $\mu g/m^3$, parts per billion (ppb), and parts per billion carbon (ppbC).
- The preferred units for risk assessment are $\mu g/m^3$. The data are not always delivered or reported in these units.
- Useful equations for converting data units:

$$[\text{conc. in } \mu g/m^3] = ([\text{conc. in ppb}] * MW * 298 * P)/(24.45 * T * 760)$$
$$[\text{conc. in ppb}] = ([\text{conc. in } \mu g/m^3] * 24.45 * T * 760)/(MW * 298 * P)$$
$$ppbC = ppb \times (\#\text{ of carbons in the molecule})$$

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

Converting Units (2 of 2)

Examples

Benzene (C_6H_6)– convert 1 ppb to µg/m^3 at standard T and P
[conc. in µg/m^3] = ([1 ppb] * 78.11)/(24.45) = 3.195 µg/m^3
where T = 298 K (25 C) and P = 760 mm Hg

Carbon tetrachloride (CCl_4)– convert 1 µg/m^3 to ppb at 0 C, 1 atm.
[conc. in µg/m^3] = ([1 ppb] * 153.82*298)/(24.45*273) = 6.867 µg/m^3
where P = 760 mm Hg

The EPA provides a thorough walk-through of the unit conversion process:
http://www.epa.gov/athens/learn2model/part-two/onsite/ia_unit_conversion_detail.htm

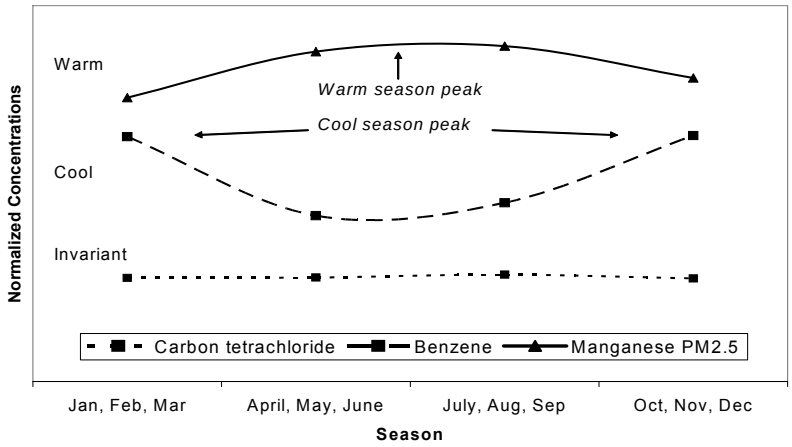
Know Your Data

Overview

- Before beginning data validation, it helps to know the typical patterns in an air toxics data set to set expectations and identify data anomalies.
  - Diurnal and seasonal patterns help analysts understand possible impacts on data aggregations when some data are missing.
- By using the power of the central tendencies in a large national data set, typical air toxics relationships are provided.
  - Patterns at individual sites may differ from the typical examples shown—understanding why there are differences becomes part of the data validation and data analysis steps.

Know Your Data

Typical Air Toxics Relationships: Seasonal Trends (1 of 2)



The plot shows an example seasonal pattern for carbon tetrachloride, benzene, and manganese $PM_{2.5}$ at a national level.

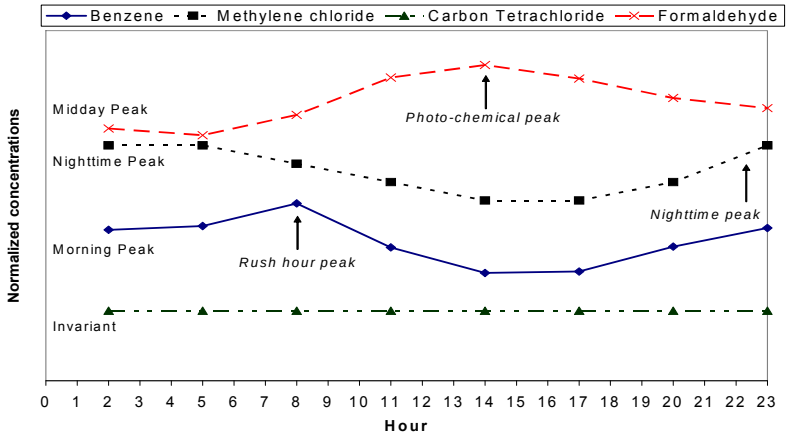
Know Your Data

Typical Air Toxics Relationships: Seasonal Trends (2 of 2)

- Pollutants that typically correlate well
  - Acetaldehyde and formaldehyde, similar sources and reactivity
  - Benzene and 1,3-butadiene, especially at locations influenced by mobile source emissions
  - Toluene concentrations, typically higher than benzene concentrations
  - Toluene and ethylbenzene, especially at locations influenced by mobile source emissions
- National seasonal patterns
  - Warm season peak: formaldehyde, acetaldehyde, chloroform, manganese $PM_{2.5}$
  - Cool season peak: benzene, 1,3-butadiene, hexane, chlorine $PM_{2.5}$ (especially at locations where roads are salted in winter)
  - Invariant: carbon tetrachloride

Know Your Data

Typical Air Toxics Relationships: Diurnal Trends (1 of 2)



The plot shows example diurnal patterns of benzene, methylene chloride, carbon tetrachloride, and formaldehyde at a national level. It was created with Microsoft Excel.

Know Your Data

Typical Air Toxics Relationships: Diurnal Trends (2 of 2)

- Midday peak, photochemical production:
  - acetaldehyde, formaldehyde
- Morning peak, mobile sources:
  - benzene, 1,3-butadiene, xylenes, hexane, ethylbenzene, toluene, 2,2,4-trimethylpentane
- Nighttime peak, affected by dilution:
  - methylene chloride, mercury vapor
- Invariant, global background:
  - carbon tetrachloride

Collocated Data

Overview

- 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.

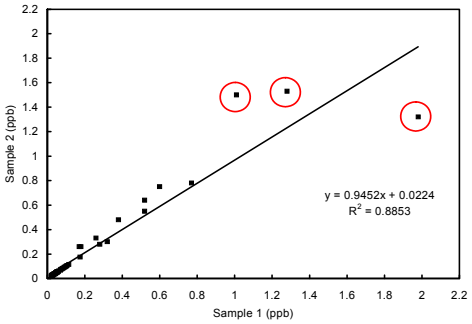
  These samples provide a measure of the precision of the chemical analysis, but do not provide any error estimates for the sample collection method.
  - In contrast, collocated samples are two samples collected at the same location and time by equivalent samplers and chemically analyzed by the same method.

  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

Collocated Data

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
- In the graph, three species 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.

Collocated Data

Aggregating Collocated Data (1 of 2)

- Double-counting collocated data should be avoided when creating aggregates such as annual averages.
  - If scatter plots of the collocated measurements correlate well, the values can be averaged together for a given site, method, date, and time.
  - If the collocated measurements do not agree, there can be no certainty which (if any) measurement is correct and the data should be excluded from analyses.

If disagreement is a regular occurrence, confidence in other data collected with the same instruments at that site is reduced.

Collocated Data

Aggregating Collocated Data (2 of 2)

- After determining that collocated measurements agree, average the two data sets together as follows:
  - If one measurement is missing, use the collocated value as the average value. Investigate the value to make sure it is consistent with the rest of the data.
  - If both values are below detection, treat them as any other data (i.e., average them together).
  - If one measurement is below detection and one is not, use the value above detection as a conservative approach.
- In some monitoring programs, only data from the primary sample are used in data analysis and the collocated sample is used only for QA purposes.

Data Completeness

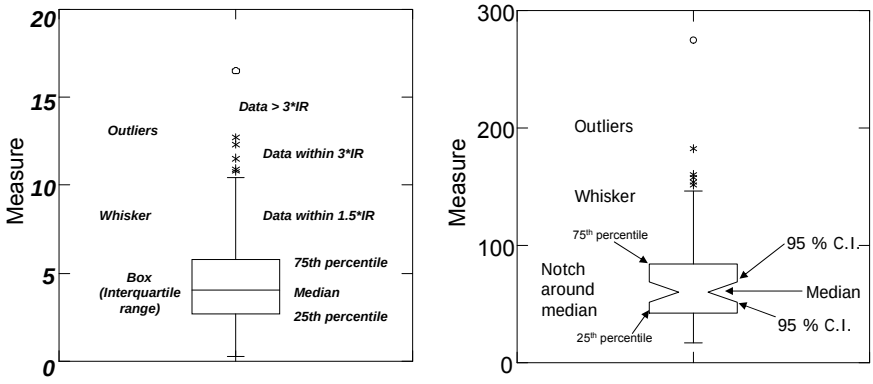
Overview

- Ensure that data are comparable across sites, years, or other subsets of the data for analysis.
- Completeness criteria are necessary in creating valid aggregated values (such as annual averages) to verify that the distribution of measured values within the aggregation window is representative of that entire period.
- Data completeness is computed using the reported sampling frequency (when available) as a measure of how many samples should be collected in a given period versus the number of samples that were collected.
  - 75% completeness is the suggested minimum value for data.
  - Using higher or lower completeness criteria may be appropriate for certain analyses.

Data Completeness

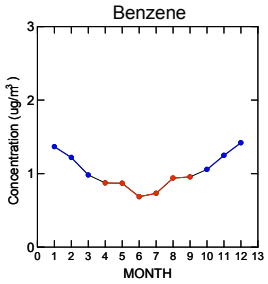
Interpreting Notched Box Plots

Notched box whisker plots are useful for showing the central trends of the data (i.e., the median) while also showing variability (i.e., the box and whiskers).



Data Completeness

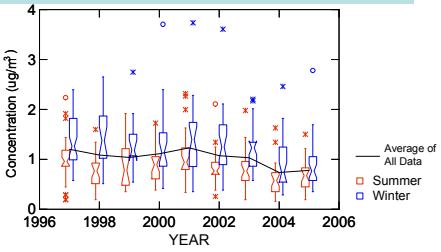
Example Effect of Aggregating Incomplete Data



Seasonal pattern of 24-hr benzene samples from an urban site. Lower concentrations in summer are typical of national concentrations and are driven by dilution from higher mixing heights in summer.

This is NOT how aggregations should be constructed.

The annual averages here were constructed using only summer (red) or winter (blue) data to illustrate aggregation results from an incomplete data set. Incomplete data cause the summer "annual averages" to be biased low and the winter "annual averages" to be biased high; the black line shows the true average of all data.



Data Aggregation

Creating Valid 24-hr Averages

- In the calculation process, it is important to verify that 24-hr averages are representative of a significant portion of the day because diurnal fluctuations in pollutant concentration throughout the day may bias the average if incomplete data are used.
- We suggest a 75% daily completeness criteria be used to ensure that a large portion of the day is represented. These criteria by sample frequency are shown in the table below.

Sample Duration	75% Daily Completeness Cutoff (# of samples)
1-hr	18
2-hr	9
3-hr	6
4-hr	5
6-hr	3
8-hr	3
12-hr	2

Data Aggregation

Creating Valid Monthly Averages

- It is suggested 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 monthly values with certainty. Therefore, at least four samples should be required in a month.

Frequency	75% Monthly Completeness Cutoff
Daily	23
Every 3rd day	8
Every 6th day	4
Other	4

- Unassigned frequencies mean that no frequency was reported with the data and a frequency could not be easily determined. The completeness criteria then defaults to the minimum to preserve data, but should be identified for later QC if possible.

Data Aggregation

Creating Valid Quarterly and Annual Averages

- Annual averages are calculated by first computing valid quarterly averages
- Quarterly Averages
  - Quarterly averages are calculated from valid 24-hr averages.
  - 75% of data at the expected daily sampling frequency is suggested

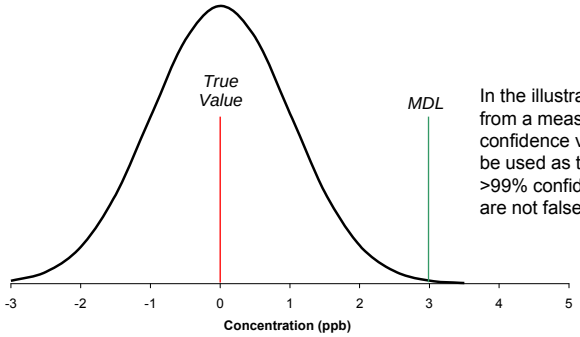
Frequency	75% Quarterly Completeness Cutoff
Daily	68
Every 3rd Day	24
Every 6th Day	12
Every 12th Day	6
Unassigned	6

  - At least 58 days are suggested between the first and last sample in a quarter to ensure sampling represented the entire quarter.
  - Unassigned frequencies mean that no frequency was reported with the data and a frequency could not be easily determined. The completeness criteria then defaults to the minimum to preserve data, but should be identified for later QC if possible.
- Annual Averages – three of four valid quarterly averages are required.

Method Detection Limits

Overview

- The EPA 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

MDLs Are Not Low Enough For Most Air Toxics Measurements

- 52% of all air toxics measurements reported in AQS from 1990-2005 are at or below the MDL.
- This percentage varies widely across pollutants; some are close to 100% below MDL.
- Data below MDL can be reported in two ways.
  - Uncensored: The measured value is reported.
  - Censored: The measured value is replaced with a proxy. Typical examples are MDL, MDL/2, MDL/10, or zero.
- We suggest that data below detection not be removed from analyses. A measurement below detection does not necessarily indicate a value of zero because ambient concentrations can be lower than currently available MDLs.
  - Data below detection are representative of the lower ambient concentration range, and removing them from analyses will bias results toward higher concentrations and may cause incorrect conclusions.

Identifying Censored Data (1 of 2)

- Data are typically reported as concentration values with accompanying MDLs.
  - In AQS, the MDL is either a default value associated with the analytical method (MDL) or a value assigned by the reporting entity for that specific record (alternate MDL).
- Identify censored values by treating data below detection.
  - Reporting of censored data will most likely differ between sites and may even be different by method, parameter, or time period for a given site.
- Identify and separate data at or below the detection limit along with the associated MDL and date/time. If alternate MDLs are available, use these alternates over the default MDLs.
  - Alternate MDLs may be different for each sample run causing a distribution of values if MDL/x substitutions were used. That values below MDL are not all the same does not mean they are not censored.

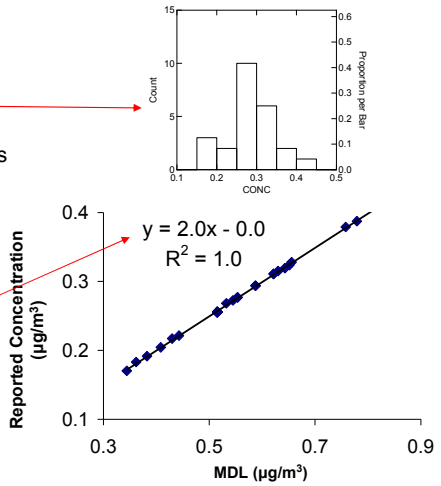
Identifying Censored Data (2 of 2)

- Examine the data for obvious substitution. Count the number of times each value at or below detection is reported for a given site, parameter, and method. Are the majority of data reported as the same value (e.g., zero or MDL/2)?
  - If data are largely reported as two or more values, investigate the temporal variation of the data. Are there large step changes where reporting methods or MDLs have changed?
  - Do the duplicate values indicate a typical censoring method (e.g., MDL/2, MDL/10)?
- Check for MDL/X substitution.
  - Make a scatter plot of the value vs. MDL to see if the data fall on a straight line.
  - If the data form a straight line, the slope of the regression line will indicate the value by which the MDL has been divided.
    - Is the value a reasonable number that would be used for MDL substitution (e.g., 1,2,5 or 10)?
    - The distribution of the ratios should be highly variable if the data are not censored.

Identifying Censored Data

Example Alternate MDL/2 Substitution

- The data shown in the table are values for a given air toxic below detection in a selected year.
- The reported data, at first glance, appear to be "real" concentrations (e.g., the histogram shows a distribution of concentrations).
- The ratio of MDL to reported concentration equals 2. In this example, the reported concentrations have been substituted with MDL/2.



Reported Concentration (µg/m³)	MDL (µg/m³)
0.19161	0.38237
0.20438	0.40834
0.22141	0.44283
0.38748	0.77921
0.40451	0.81327
0.37896	0.75792
0.17032	0.34404
0.18309	0.36193
0.27251	0.54502
0.31935	0.64295
0.31083	0.62166
0.29380	0.58760
0.32361	0.65147
0.26825	0.53225
0.27677	0.55354
0.31509	0.63018
0.25548	0.51521
0.32786	0.65573
0.27677	0.55354
0.25548	0.51521
0.25548	0.51521
0.25548	0.51521
0.29380	0.58760
0.31083	0.62166

Method Detection Limits

Treating Data Below Detection

- In a site-level analysis, in which the analyst knows how the data have been reported, more sophisticated methods may be employed.
  - If uncensored values are reported below MDL, use the data "as is" with no substitution.
  - If uncensored values are not available, use MDL/2 substitution for data at or below MDL if trying to calculate an annual mean value:
    - Substitution will lead to a 10-40% bias when fewer than 85% of the data are below MDL.
    - At >85% of data below MDL, uncertainties are large and one may only reliably state that the concentration is below MDL.
- Alternatives to MDL/2 substitution are more statistically intensive; however, in some cases they may yield better results. Note at a high degree of censoring (>70% censored data), no technique will produce good estimates of summary statistics.

Method Detection Limits

Treating Data Below Detection

- EPA recommends some approaches other than MDL/2 substitution:
  - Regression order statistics (ROS) and probability plotting (MR) methods. ROS and MR methods are superior when distribution shape population is unknown or nonparametric.
    - ROS produces more accurate results when >30% of the data is below detection.
  - Maximum likelihood estimation (MLE). MLE methods have been shown to have the smallest mean-squared error (i.e., higher accuracy) of available techniques when the data distribution is exactly normal or lognormal.
    - MLE does not work well for data sets with <50 detected values.
  - Kaplan-Meier is effective for data sets when less than 70% of the data is censored and the distribution is nonparametric.

Method Detection Limits

Treating Data Below Detection

- Mixed Data Sets
  - For data sets that have a mix of censored and uncensored data, compare two substitution methods: (1) substitute MDL/2 for censored values and leave uncensored values “as is” and (2) substitute MDL/2 for all data below detection.
  - Results that are comparable using both substitution methods increase confidence in the results, and substitution method 1 should be retained. If the results do not agree, a more sophisticated method for estimating the data below MDL may be employed.
- In all cases, flag data below detection and calculate the percentage of data below MDL for all aggregated values.

Data Treatment Methods

The selection of a data treatment method for below MDL data depends on the amount of data below MDL and the data quality objectives which are to be met. Methods explored in previous air toxics work are discussed next.

- Ignore data below MDL.
  - *Not recommended.* Reduces number of samples. Results in a bias of higher values in summary statistics.
- Replace data below MDL with zero.
  - *Not recommended.* May bias summary statistics low.
- Replace data below MDL with the actual MDL.
  - *Not recommended.* May bias summary statistics high.
- Replace data below MDL with % non-detects*MDL
  - *Not recommended.* Found to be similar to MDL/2 substitution.
- Replace data below MDL with MDL/2.
  - *Recommended as a simple method for calculating mean values with relatively small bias.*
- Replace data below MDL with more statistically intensive approaches (such as Kaplan-Meier, Maximum Likelihood Estimation, and Robust Regression on Order Statistics [KM, MLE, and ROS])
  - *Recommend for sophisticated analyses* such as quantifying percentiles in the data rather than simply the mean.

Maximum Likelihood Estimation (MLE)

- Maximum likelihood estimation (MLE) (also called Cohen's method) is a popular statistical method used for fitting a mathematical model to data.
- This method relies on knowing (or assuming) the underlying statistical distribution (e.g., lognormal) from which the data are derived.
- Uncensored data are used to calculate fitting parameters that represent the best fit to the distribution.
- MLE is sensitive to outliers and does not perform well if the data do not follow the assumed distribution.
- MLE requires at least 50 uncensored values to work well, so 1-in-6-day sampling will usually not be sufficient for calculating annual statistics using this technique.

MLE Calculations

Using Statistical Software

- The MLE model is a parametric analysis because the distribution is assumed -- usually assumed to be lognormal for atmospheric data.
- Each data value is assigned a range of possible concentrations:
  - Censored data: Lower value = 0, Higher value = MDL
  - Uncensored data: Lower value = Higher value = Reported value
- The statistical software procedure may require a distribution for the input, or require you to log-transform your data if a normal distribution is assumed.
- Summary statistics will be produced that provide estimates of mean, standard deviation, and some percentiles for the data set of interest.

Nonparametric Kaplan-Meier (KM)

- Nonparametric methods rely only on ranks of data and make no assumptions about the statistical distribution of the data.
- Nonparametric methods are insensitive to outliers.

KM Using Statistical Software

- Kaplan-Meier can be accessed under Survival Analysis in most statistical packages.
  - This analysis usually expects data to be right-censored (i.e., values greater than X, rather than less than X).
  - Data may need to be "flipped". Take your highest value and set it as the upper-bound. Subtract all values from it to get your input data set. Censored data are considered less than the MDL.
    - Original data set = 10, 7, 3, *2, 1.5*, 0.7, *0.3 (red = MDL-censored)*
    - Flipped data set = 0, 3, 7, *8, 8.5*, 9.3, *9.7*
  - Input your flipped data set along with a second column indicating the censored data values.
- The output will include a survival plot (cumulative distribution function) and estimated summary statistics for the flipped data set.
  - Re-flip the summary statistics for mean, median, and percentiles.
  - Measures of variances (standard deviation, confidence intervals) are independent of flipping and do not need to be changed from the output values.

Robust Regression on Order Statistics (ROS)

- These techniques calculate summary statistics with a regression equation on a probability plot.
- ROS assumes a distribution only for censored data.
- This technique is better for data sets with <30 observations and is therefore suited to typical air toxics data sets.

ROS using Statistical Software

- Data are input as reported values and MDL-censored values. MDL-censored values will need a column indicating they are censored.
- ROS statistics calculate the probability that observed data are below each MDL value. If there is only one MDL value, this is just the fraction of data below MDL.
  - Original data set = 10, 7, 3, *2*, *1.5*, 0.7, *0.3, 0.3 (red = below MDL)*
    - Probability > 2 = 0.375
    - Probability > 1.5 = 0.375
    - Probability > 0.3 = 0.583
  - Using these probabilities, probability plotting positions are calculated for all detected and censored observations using the detected data to determine a best-fit distribution.
  - Summary statistics are output from this dataset.

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.

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

Warren and Nussbaum, 2009

Treating Data <MDL

Example

- This example walks through the Maximum Likelihood Estimation (MLE) and Kaplan-Meier (KM) replacement methods.
- The MLE method requires that data without the nondetects be normally distributed and that there be only one detection limit in the data set. Neither requirement is routinely met with air toxics data.
- The KM method does not require knowing the distribution of the data and can accommodate multiple detection limits. KM is a "flipped" version of censored survival data analysis.

1.752	1.045
1.563	<1.000 (0.977)
1.498	<1.000 (0.944)
1.477	<1.000 (0.919)
1.418	<1.000 (0.897)
1.358	<1.000 (0.818)
1.327	<1.000 (0.806)
1.289	<0.800 (0.777)
1.148	<0.800 (0.622)
1.060	<0.800 (0.455)

Pollutant Concentrations (µg/m^3)
Assumes MDL of 1.000 or 0.800
(Actual values also shown)

From material supplied by Warren and Nussbaum (2009); in Appendix to Section 4

Maximum Likelihood

Example

- Let $X_1, X_2, ..., X_m, ..., X_n$ represent all the n data values ranked from largest to smallest. The first "m" values represent the data values above the detection limit (DL), and the remaining "n-m" data points are those below DL.
- Compute the sample mean and the sample variance from only the "m" above detection data values. The mean will be too large because the small undetected values have been ignored, and the variance too small.
- The mean will be lowered and the variance enlarged through the use of factors:

$$h = \frac{n - m}{n} \qquad \gamma = \frac{s_d^2}{\left(\overline{X}_d - DL\right)^2}$$

$\overline{X}_d$ is the sample mean
s_d is the sample standard deviation
m is the number of detected values
n is the total number of values

- Use the table on the next page to obtain

$$\hat{\lambda}(\gamma, h)$$

From material supplied by Warren and Nussbaum (2009)

EPA/QA/G-9S, Table A-11

	h											
γ	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.80	.90
.00	.31862	.4021	.4941	.5961	.7096	.8388	.9808	1.145	1.336	1.561	2.176	3.283
.05	.32793	.4130	.5066	.6101	.7252	.8540	.9994	1.166	1.358	1.585	2.203	3.314
.10	.33662	.4233	.5184	.6234	.7400	.8703	1.017	1.185	1.379	1.608	2.229	3.345
.15	.34480	.4330	.5296	.6361	.7542	.8860	1.035	1.204	1.400	1.630	2.255	3.376
.20	.35255	.4422	.5403	.6483	.7673	.9012	1.051	1.222	1.419	1.651	2.280	3.405
.25	.35993	.4510	.5506	.6600	.7810	.9158	1.067	1.240	1.439	1.672	2.305	3.435
.30	.36700	.4595	.5604	.6713	.7937	.9300	1.083	1.257	1.457	1.693	2.329	3.464
.35	.37379	.4676	.5699	.6821	.8060	.9437	1.098	1.274	1.475	1.713	2.353	3.492
.40	.38033	.4735	.5791	.6927	.8179	.9570	1.113	1.290	1.494	1.732	2.376	3.520
.45	.38665	.4831	.5880	.7029	.8295	.9700	1.127	1.306	1.511	1.751	2.399	3.547
.50	.39276	.4904	.5967	.7129	.8408	.9826	1.141	1.321	1.528	1.770	2.421	3.575
.55	.39679	.4976	.6061	.7225	.8517	.9950	1.155	1.337	1.545	1.788	2.443	3.601
.60	.40447	.5045	.6133	.7320	.8625	1.007	1.169	1.351	1.561	1.806	2.465	3.628
.65	.41008	.5114	.6213	.7412	.8729	1.019	1.182	1.368	1.577	1.824	2.486	3.654
.70	.41555	.5180	.6291	.7502	.8832	1.030	1.195	1.380	1.593	1.841	2.507	3.679
.75	.42090	.5245	.6367	.7590	.8932	1.042	1.207	1.394	1.608	1.851	2.528	3.705
.80	.42612	.5308	.6441	.7676	.9031	1.053	1.220	1.408	1.624	1.875	2.548	3.730
.85	.43122	.5370	.6515	.7781	.9127	1.064	1.232	1.422	1.639	1.892	2.568	3.754
.90	.43622	.5430	.6586	.7844	.9222	1.074	1.244	1.435	1.653	1.908	2.588	3.779
.95	.44112	.5490	.6656	.7925	.9314	1.085	1.255	1.448	1.668	1.924	2.607	3.803
1.00	.44592	.5548	.6724	.8005	.9406	1.095	1.287	1.461	1.882	1.940	2.626	3.827

Maximum Likelihood

Example Continued

- Estimate the corrected sample mean and corrected sample variance to account for the data below the DL:

$$\overline{X} = \overline{X}_d - \hat{\lambda}\left(\overline{X}_d - DL\right) \qquad s^2 = s_d^2 + \hat{\lambda}\left(\overline{X}_d - DL\right)^2$$

- Let $X_1, X_2, ..., X_m, ..., X_n$ represent all the n data values ranked from largest to smallest: *1.752, 1.563, 1.498, 1.477, 1.418, 1.358, 1.327, 1.289, 1.148, 1.060, 1.045, <1.000, <1.000, <1.000, <1.000, <1.000, <1.000, <1.000, <1.000, <1.000*
- The first "m" values represent the data values above the DL, and the remaining "n-m" data points are those below the detection limit: *n = 20, m = 11, n-m = 9*
- Compute the sample mean and the sample variance from only the "m" above detection data values: *Mean = 1.358 Variance = 0.0524*
- The first factor (h): *11/20 = 0.55*
- The second factor (γ): *0.0524/(1.358 – 1.000)2 = 0.409*
- The third factor (h, γ, Table A-11): *1.113*
- Estimate the corrected sample mean and corrected sample variance to account for the data below the DL: *Mean = 1.358 – 1.113(1.358 – 1) = 0.960* and *variance = 0.0524 + 1.113(1.358 – 1)2 = 0.195*

From material supplied by Warren and Nussbaum (2009)

Kaplan-Meier

Example

- For this example, the maximum was 1.752, using 2 as the flip point: *1.752 when flipped is 0.248, 1.563 becomes 0.437, etc.*
- This method will find a specific probability (denoted as g_i) for each X_i (the flipped value) using an "Incremental Survival Probability"
- The "g_i" and "X_i" are combined to estimate the mean and variance:
  Mean = $\sum g_i X_i$ Variance = $\sum g_i X_i^2$ - (Mean)^2
- The Mean is then flipped back to the original scale; variance is left as is.
- The computation is summarized on the next slide.
  - Col 1: The actual data values (non-detects indicated by a dashed line)
  - Col 2: The "flipped data" = 2 minus the actual value
  - Col 3: Rank order (the missing ranks belong to non-detects)
  - Col 4: b = n – r + 1 where n = total (20), r = rank
  - Col 5: d = number of observations for this value (1 in this case)
  - Col 6: p = (b – d)/b
  - Col 7: S = The S from the previous row multiplied by the p for the current row (starts at 1.0000). E.g., 10th data value: *S = 0.5500 x 10/11 = 0.500*
  - Col 8: g = The S from the previous row minus the S for the current row (starts at 1.000). E.g., 10th data value: *g = 0.5000 – 0.4500 = 0.0500*.
- The X_is are the flipped values and the g_is come from the table.
  - Mean = 0.05 x 0.248 + …+ 0.16875 x 1.200 = 0.8620
  - Variance = 0.05 x 0.2482 +…+0.16875 x 1.2002 – 0.86202 = 0.085
- The true Mean is then 2 – 0.8620 = 1.138 and the variance 0.085

From material supplied by Warren and Nussbaum (2009)

Kaplan-Meier

Example

Data	Flip on 2	rank	b = n-r+1	d	p=(b-d)/b	S	g
1.752	0.248	1	20	1	19/20	0.9500	0.0500
1.563	0.437	2	19	1	18/19	0.9000	0.0500
1.498	0.502	3	18	1	17/18	0.8500	0.0500
1.477	0.523	4	17	1	16/17	0.8000	0.0500
1.418	0.582	5	16	1	15/16	0.7500	0.0500
1.358	0.642	6	15	1	14/15	0.7000	0.0500
1.327	0.673	7	14	1	13/14	0.6500	0.0500
1.289	0.711	8	13	1	12/13	0.6000	0.0500
1.148	0.852	9	12	1	11/12	0.5500	0.0500
1.060	0.940	10	11	1	10/11	0.5000	0.0500
1.045	0.955	11	10	1	9/11	0.4500	0.0500
0.977	1.023	13	8	1	8/9	0.3938	0.05625
0.944	1.056	14	7	1	7/8	0.3375	0.05625
0.919	1.081	15	6	1	6/7	0.2813	0.05625
0.897	1.103	16	5	1	5/6	0.2250	0.05625
0.818	1.182	17	4	1	4/5	0.1688	0.05625
<0.800	>1.200	18	3	3	0	0	0.16875

Comparison of Methods

Example

	True	Zero	DL	½ DL	MLE	ROS	K-M
Mean	1.108	0.747	1.422	0.972	0.960	1.197	1.138
Var	0.117	0.505	0.099	0.302	0.195	0.048	0.085

- In this example, the easiest methods—substitution with zero, DL, or ½ DL—give poor results.
- MLE and ROS (not shown in the example) provide fairly good mean and variance values considering the high non-detect rate (45%) in this example. However, these methods require significant work to calculate the estimates.
- Kaplan-Meier provides reasonable estimates for this example, and works when there are multiple detection limits. However, this method also requires significant work to calculate the estimates.

From material supplied by Warren and Nussbaum (2009)

Data Validation

Introduction (1 of 2)

- 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.
- Validation guidelines are built on knowledge of typical air toxics emissions sources; formation, loss, and transport processes; chemical relationships; and site-specific knowledge.
- 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

Introduction (2 of 2)

- 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.
- The data analyst should perform Level 1 steps and perform additional validation when other data sets are available.
- There is no substitute for the local knowledge of monitoring sites; operators or those who have extensive knowledge of the area are a unique resource for data analysts.

* U.S. Environmental Protection Agency, 1999.

Data Validation

Initial Approach

- 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 flag.
  - Comparison to remote background concentrations: urban air toxics concentrations should not be lower than remote background concentrations.

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 Years Day at many sites.

Data Validation

Tips and Tricks

- Overall
  - Proceed from the big picture to the details. For example, proceed from inspecting species groups to individual species.
  - Inspect every specie, even to confirm that a specie normally absent met that expectation.
  - Know the site topography, prevalent meteorology, and major emissions sources nearby.
- Inspect time series for the following
  - Large "jumps" or "dips" in concentrations which may indicate a change in analysis method or MDL.
  - Periodicity of peaks. (Is there a pattern? Can the pattern be related to emissions or meteorology?)
  - Expected seasonal behavior (e.g., photochemically formed species concentrations usually peak during summer).
  - Expected relationships among species (e.g., benzene and toluene typically correlate).

Data Validation

To Further Investigate 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.

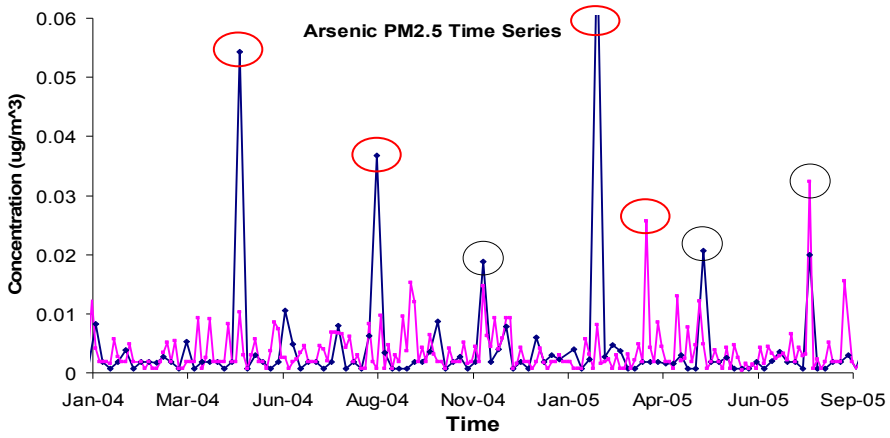
Data Validation

Using Summary Statistics

- Investigate summary statistics to begin to understand your data.
- Compare data ranges to “typical” ranges as a reality check.
- National summary statistics based on 2003 to 2005 annual averages for selected species can be found in the appendix to this section.
- These data can be used as benchmarks for site-specific comparison; for example, if your data are significantly higher than the national 95th percentile, there may be errors in the data.
  - Note that calculation of summary statistics smoothes extreme events so comparison of daily data to these numbers, for example, may not be adequate; individual high concentration days may legitimately be higher than the summary statistics.
  - We suggest a comparison between similar summary statistics rather than a comparison of summary statistics to raw data.

Data Validation

Buddy Check Example

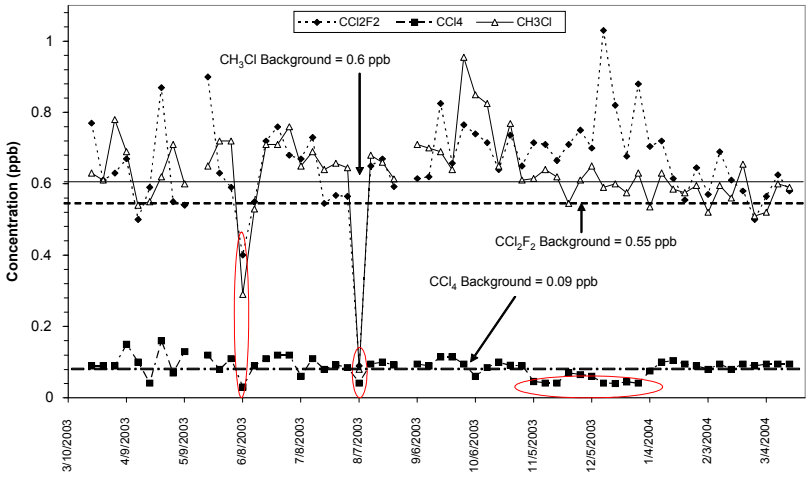


Sample time series of 24-hr arsenic $PM_{2.5}$ measurements at two sites about five miles apart. Both sites show above average arsenic concentrations and are located near a major emissions source. The figure was created in Microsoft Excel.

Screening Data Using Remote Background Concentrations

- Knowledge of remote background concentrations of air toxics can be used as lower limits for data screening.
  - A cutoff value of 20% 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.
- This screen was applied to the national data set. Data failing this check were not used in subsequent analyses.

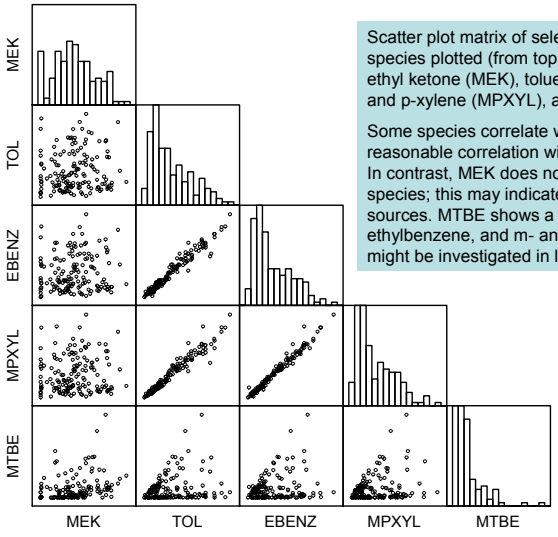
Screening Data Using Remote Background Concentrations



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 Examples

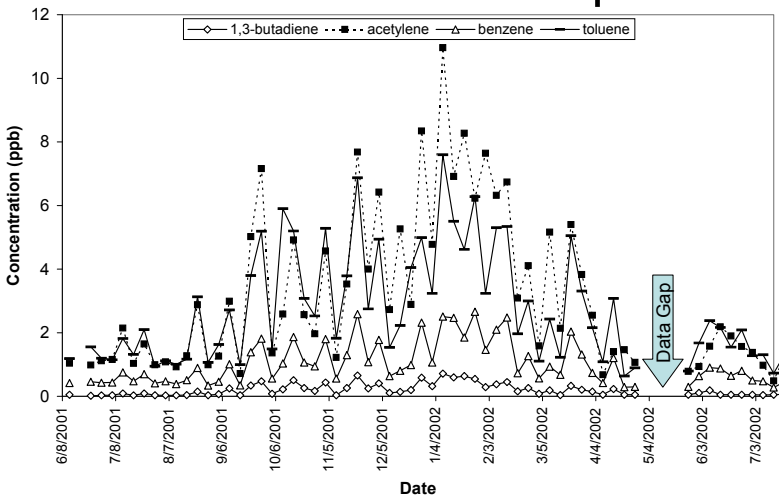
Scatter Plots



Scatter plot matrix of selected species from an urban site. The species plotted (from top to bottom and left to right) are methyl ethyl ketone (MEK), toluene (TOL), ethylbenzene (EBENZ), m- and p-xylene (MPXYL), and methyl tert-butyl ether (MTBE).

Some species correlate well. For example, toluene has a reasonable correlation with ethylbenzene and m- and p-xylene. In contrast, MEK does not correlate with any of the other species; this may indicate that MEK is emitted from different sources. MTBE shows a bifurcated relationship with toluene, ethylbenzene, and m- and p-xylene. This interesting relationship might be investigated in later validation steps and analysis.

Data Validation Examples

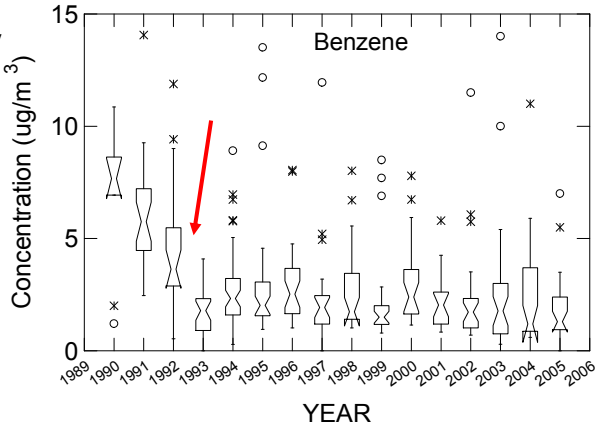


Twenty-four-hour average concentrations (ppb) of acetylene, 1,3-butadiene, benzene, and toluene collected at an urban site every sixth day from July 2001 through July 2002. Expected seasonal and inter-species relationships were observed.

Data Validation Examples

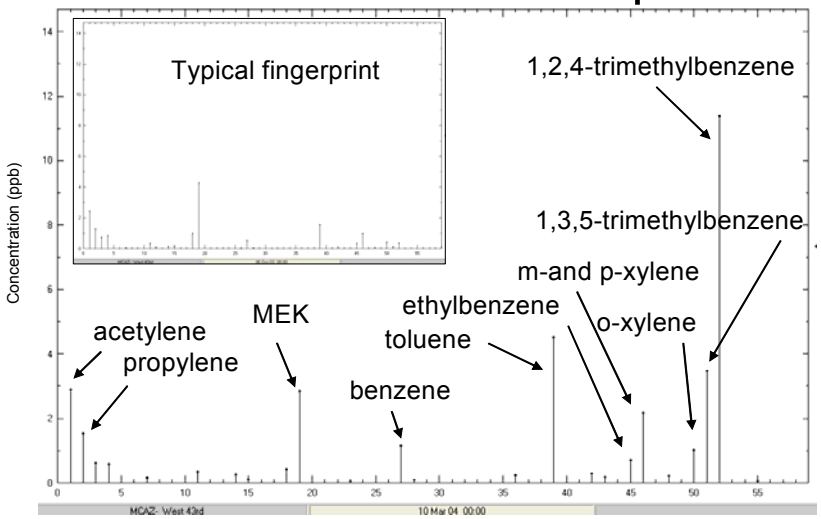
It is immediately clear by the large concentration change from 1990-1993 that something affected the data and should be investigated.

- Were there significant method or MDL changes during this time?
- Is this change due to emissions regulations or is there another explanation?



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.

Data Validation Examples

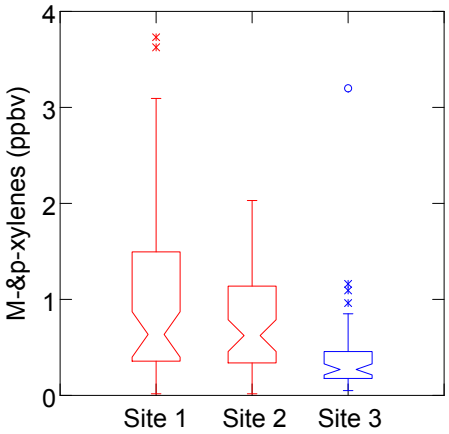


Example fingerprint plot of 24-hr concentrations (ppb). The inset figure shows a more typical fingerprint at the same site on another date.

Data Validation Examples

Using Metadata – Urban vs. Rural Sites

- Concentrations at each site do not need to be the same but do need to be consistent with our expectations of concentrations at urban and rural sites.
- Sites 1 and 2 show the highest concentrations because these sites are relatively close to an Interstate highway and are located in urban areas.
- In contrast, Site 3 shows relatively low m-&p-xylenes concentrations, as expected for a site outside the urban area.

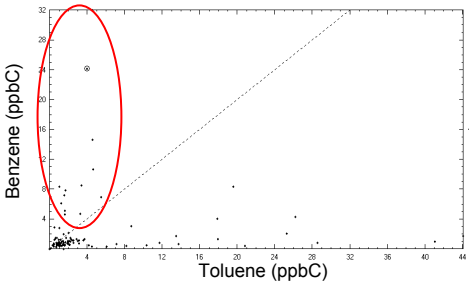


Notched box whisker plot of 24-hr m-&p-xylenes concentrations at three monitoring stations in 2005. Sites 1 and 2 are urban and Site 3 represents a rural site.

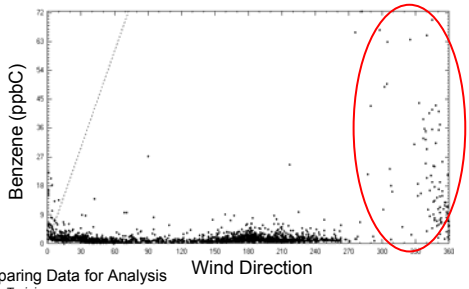
Data Validation Examples

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).



Data Validation

Handling Suspect Data

- During the process of data validation, the analyst may identify data as suspect but not be able to prove that the data are invalid.
- Analysts may decide to exclude these suspect data from central tendency computations (e.g., annual average) or other analyses.
- These data may warrant additional investigation using case studies (i.e., inspection of individual dates).

Summary

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.
- **Data processing**
  - ❑ 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
    - ○ 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.
- **Data validation**
  - ❑ 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 can not be invalidated. Advanced analyses may provide more insight into the data.

Appendix – National Summary Statistics

- The appendix contains a table of national summary statistics based upon annual averages from 2003 to 2005.
- These data are useful for comparison of data ranges to "typical" national ranges.
- These data can be used as benchmarks for site-specific comparison; for example, if data are significantly higher than the national 95th percentile, there may be errors in the data.

Pollutant	AQS Code	% Below Detection	# of Monitoring Sites	5th Percentile Concentration (µg/m3)	25th Percentile Concentration (µg/m3)	Median Concentration (µg/m3)	75th Percentile Concentration (µg/m3)	95th Percentile Concentration (µg/m3)
1,1,2,2-Tetrachloroethane	43818	97	228	6.9E-02	1.6E-01	1.7E-01	3.1E-01	1.1E+00
1,1,2-Trichloroethane	43820	98	211	5.5E-02	1.3E-01	1.4E-01	1.9E-01	9.0E-01
1,1-Dichloroethane	43813	97	224	1.0E-02	6.1E-02	1.0E-01	1.0E-01	6.8E-01
1,1-Dichloroethylene	43826	98	225	2.0E-02	9.5E-02	9.9E-02	1.1E-01	6.5E-01
1,2,4-Trichlorobenzene	45810	90	164	1.2E-02	6.2E-02	1.5E-01	6.4E-01	1.2E+00
1,2-Dichloropropane	43829	96	229	1.5E-02	7.7E-02	7.9E-02	1.5E-01	7.6E-01
1,3-Butadiene	43218	26	278	3.5E-02	9.5E-02	1.6E-01	2.4E-01	8.4E-01
1,4-Dichlorobenzene	45807	64	202	1.9E-02	1.1E-01	2.4E-01	5.2E-01	9.9E-01
1,4-Dioxane	46201	94	14	4.5E-02	4.9E-02	6.9E-02	9.2E-02	1.2E-01
2,2,4-Trimethylpentane	43250	13	125	1.1E-01	2.9E-01	4.8E-01	7.8E-01	2.4E+00
3-Chloropropene	43335	100	13	1.1E-01	1.2E-01	1.6E-01	1.6E-01	1.9E-01
Acenaphthene	17147	44	33	5.6E-04	5.7E-03	1.4E-02	3.9E-02	7.2E-02
Acenaphthylene	17148	68	33	2.4E-04	6.8E-04	3.4E-03	3.9E-02	4.4E-02

Excerpt

Resources

Data Acquisition

- Data acquisition
- Quality assurance
- Metadata
- Advanced methods for estimating data structure below detection
- Information and methods
- Data validation
- Data analysis

Characterizing Air Toxics

What are the diurnal, seasonal, and spatial characteristics of air toxics?

What do these characteristics tell us about emission sources, transport, and chemistry?

Characterizing Air Toxics
What's Covered in This Section

- Temporal Patterns
  - Diurnal
  - Day-of-week
  - Seasonal
- Spatial Patterns
  - Spatial characterization
    - National concentration plots for perspective
    - Maps
  - Variability within and between cities
  - Hot and cold spot analysis
  - Comparing urban and rural sites
- Risk screening

Characterizing Air Toxics
Overview *(1 of 2)*

- Spatial and temporal characterizations of air toxics data are the basis for improving our understanding of emissions and the atmospheric processes that influence pollutant formation, distribution, and removal.
- Characterization analyses help us develop a conceptual model of processes affecting air toxics concentrations and also provide an opportunity to compare data to existing conceptual models to identify interesting or problematic data.

Characterizing Air Toxics
Overview *(2 of 2)*

- Typical questions which may be addressed using these types of analyses
  - Where are air toxics concentrations highest or lowest?
  - How do pollutant concentrations vary relative to each other – and what does this tell us about their sources?
  - What and where are the air toxics of concern?
  - How do urban and rural sites compare?
  - How do air toxics concentrations compare to criteria pollutants (e.g., ozone and $PM_{2.5}$)?
  - What local or regional sources influence a particular measurement site?

Quantifying Patterns
Methods *(1 of 2)*

- When investigating temporal patterns, analysts use statistical measures to understand if concentrations are statistically different.
- Testing statistical significance using t-test
  - The t-test is a very common method for assessing the difference in mean values of two groups of data (e.g., the difference in means of two years of data).
  - This test assumes that both data sets are normally distributed, a fact that is not true for many air toxics measurements. However, this is not a problem as long as there are sufficient data in each group (>~100). Each data set is also required to contain the same number of samples.
  - If there are fewer than 100 data points per group, a more advanced, non-parametric, test must be used. Some examples are Kruskal-Wallis, Kolmogorov-Smirnov, Anderson-Darling (sample sizes of 10 to 40 only).

StatSoft, Inc. (2005)

Quantifying Patterns
Methods *(2 of 2)*

- Testing statistical significance using notched box plots
  - For the national analyses, SYSTAT notched box plots were used as a quick check of statistical significance between two groups. The notches on a box plot represent the range of the upper to lower 95th percentile confidence intervals surrounding the median (a full description of notched box plots can be found in *Preparing Data For Analysis*, Section 4). If the notches of two box plots do not overlap, the median concentrations are statistically significantly different.
  - Testing with notched box plots provides a qualitative view of significance on the median concentration value, not the mean.
- Most of these statistical methods can be performed with Microsoft Excel or SYSTAT, as well as many other statistical programs.

StatSoft, Inc. (2005)

Characterizing Temporal Patterns
Motivation

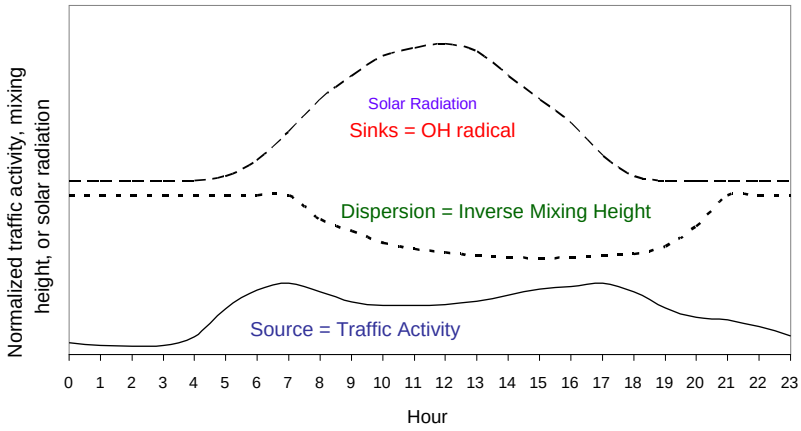
- To more fully understand potential contributing air toxics sources, analysts may also wish to consider:
  - Diurnal patterns. How does the daily cycle of air toxics concentrations relate to emissions and meteorology? Are diurnal patterns properly reflected in exposure models?
  - Day-of-week patterns. Does the weekly cycle of air toxics concentrations tell us anything about emissions sources?
  - Seasonal patterns. Do air toxics concentrations show seasonal patterns and do these patterns make sense with respect to what we know about formation, transport, and removal processes?

Diurnal Patterns
Overview

- Air toxics data are not routinely collected on a subdaily basis; most data are reported as 24-hr averages. The diurnal variation of some air toxics is unknown because of data limitations.
- Subdaily data allow us to
  - Evaluate diurnal variation,
  - Understand general atmospheric processes (the physics, chemistry, and sources of air toxics),
  - Assess the performance of models that are attempting to capture diurnal cycles, and
  - Provide input to receptor-based models.
- Reasons to understand diurnal patterns include
  - Assessment of human exposure and health effects,
  - Identification of local sources vs. regional transport, and
  - Contribution to an understanding of the physics and chemistry of air toxics.

Diurnal Patterns
Conceptual Model

Concentrations = (Sources – Sinks + Transport)/Dispersion



Diurnal Patterns
Approach (1 of 4)

- Suggested data requirements
  - 75% sampling completeness for each site, pollutant, and day to ensure that (1) data are representative of a full day and (2) data are consistent with completeness requirements used to construct other aggregates.
  - Percent below detection tracked for each pollutant and year.
  - In initial national level analyses, a minimum of 10 measurements for each air toxic and hour was set to include as many air toxics as possible in the analysis; more measurements are recommended if they are available.
  - Data should be inspected on both a concentration and normalized basis for each available duration. Normalization enables a comparison of diurnal patterns among sites and pollutants even if pollutant concentrations vary widely.

Diurnal Patterns

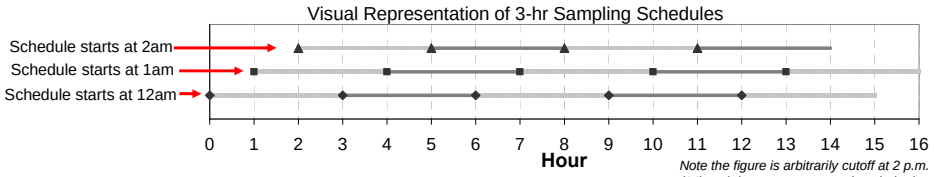
Approach *(2 of 4)*

- Data are normalized using the average concentration for each individual day, site, duration, and pollutant. To normalize data,
  - Calculate the average concentration by date, site, pollutant, and duration.
  - Divide the corresponding subdaily data by this average.
- The resulting normalized values provide an indication of the magnitude of difference of the hourly concentration from the average concentration for that day. A value of 1 indicates that the hourly concentration value is the same as the daily average concentration. Values > 1 are greater than the average value (e.g., a value of 2 is 2 x average value) while values < 1 are lower than the average value (e.g., a value of 0.5 is ½ the average value).

Diurnal Patterns

Approach *(3 of 4)*

- Subdaily measurements made on different sampling schedules must be taken into account when aggregating multi-site data.
  - Diurnal analyses can be obscured by the different sample schedules when aggregating multi-site data if the number of samples for each hour is different across hours. This issue needs to be considered when data from different jurisdictions are used (such as at the national scale).
  - Hypothetical case: some sites used a 2 a.m. sample schedule and other sites used a 1 a.m. sample schedule. Consider the first three hours of the day—the sample that begins at 2 a.m. includes all three sampling schedules (i.e., all three samples overlap). For aggregating data with multiple sampling schedules, calculate a weighted average of the hour representing the middle of staggered sampling schedules (i.e., 2 a.m. sampling schedule for 3-hr duration) from the raw data before completing the next steps.



Note the figure is arbitrarily cutoff at 2 p.m. (14) and does not represent the whole day.

Diurnal Patterns

Approach *(4 of 4)*

- Summary statistics may be generated by pollutant and hour for the concentration and normalized data sets.
  - Inspect various parameterizations of the data (e.g., 10th, 50th, and 90th percentiles), especially when more than 50% of data is below detection.
  - Include the standard deviation or confidence interval as a measure of uncertainty in the data.
- Subdaily patterns can be visualized by using line graphs of summary statistics with confidence intervals or notched box plots.

Diurnal Patterns

Effect of Sampling Schedule *(1 of 2)*

- Table 1 shows the raw measurements by begin-hour (i.e., the time that would be reported with the measurement).
- Table 2 provides the aggregated weighted averages.

Table 1. Raw Measurements

Begin Hour Of Measurement	Number of Measurements	Median Concentration (µg/m^3)
0	66	0.777
1	66	0.708
2	64	0.729
3	66	0.665
4	66	0.697
5	65	0.857
6	70	0.947
7	71	0.995
8	68	0.836
9	66	0.692
10	64	0.554
11	64	0.490
12	78	0.500
13	70	0.463
14	67	0.479
15	67	0.479
16	66	0.495
17	64	0.511
18	66	0.585
19	66	0.692
20	64	0.793
21	64	0.852
22	64	0.852
23	64	0.814

Table 2. Aggregated Measurements

Aggregated Hour	Weighted Average Median Concentration (µg/m^3)
2	0.738
5	0.739
8	0.927
11	0.580
14	0.482
23	0.839

Weighted Average (WA) Formula:

$$WA = (1/\sum N_i)*\sum N_i C_i$$

N = Number of Measurements
C = Concentration

Example calculation, aggregated to 2 a.m. sample schedule:

[1/(66+66+64)]*[66*0.777+66*0.708+64*0.729] = 0.738

Diurnal Patterns

Effect of Sampling Schedule *(1 of 2)*

- Table 1 shows the raw measurements by begin-hour (i.e., the time that would be reported with the measurement).
- Table 2 provides the aggregated weighted averages.

Table 1. Raw Measurements

Begin Hour Of Measurement	Number of Measurements	Median Concentration (µg/m^3)
0	66	0.777
1	66	0.708
2	64	0.729
3	66	0.665
4	66	0.697
5	65	0.857
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22	64	0.852
23	64	0.814

Table 2. Aggregated Measurements

Aggregated Hour	Weighted Average Median Concentration (µg/m^3)
2	0.738
5	0.739
8	0.927
11	0.580
14	0.482

Weighted Average (WA) Formula:

$$WA = (1/\sum N_i)*\sum N_i C_i$$

N = Number of Measurements

C = Concentration

Example calculation, aggregated to 2 a.m. sample schedule:

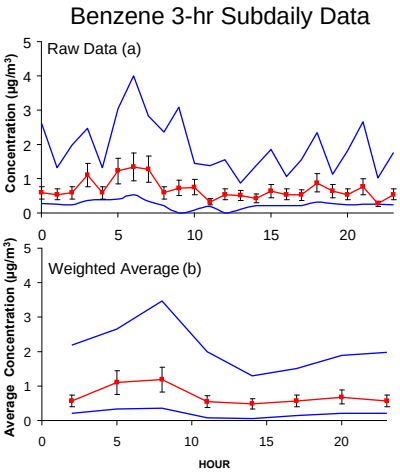
[1/(66+66+64)]*[66*0.777+66*0.708+64*0.729]

= 0.738

Diurnal Patterns

Effect of Sampling Schedule *(2 of 2)*

- Figure (a) shows the 10th, 50th, and 90th percentile of national 3-hr benzene data. The noise in this pattern is due to varying amounts of data available from three sampling schedules which begin at 12, 1, or 2 a.m.
  - Sampling-schedule differences are typical when aggregating 3-hr or 4-hr measurements and can obscure diurnal patterns.
- Figure (b) shows the same data as a weighted average by the most representative hour.
- Averaging clarifies the diurnal pattern showing a morning peak trend as would be expected for benzene concentrations at most sites.

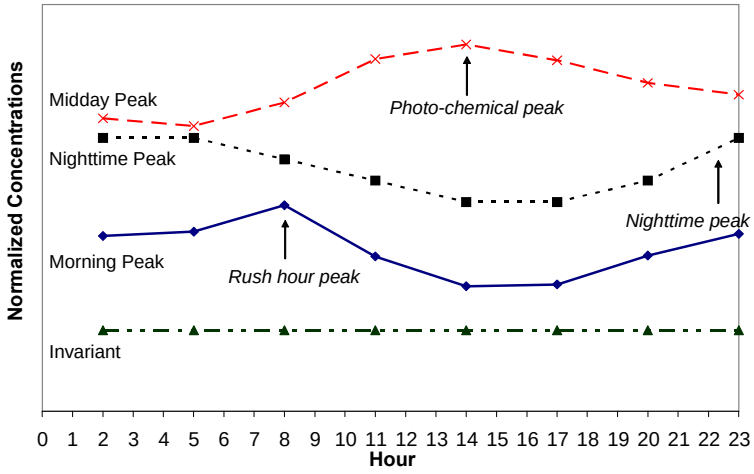


The 10th, 50th, and 90th percentile of national 3-hr benzene data.

Diurnal Patterns

Commonly Observed Patterns

Sample of four commonly observed diurnal patterns using national 3-hr duration data. The sources, sinks, transport, and dispersion leading to each pattern are discussed in this section. *Data were normalized as described in the approach to diurnal patterns.*



Diurnal Patterns

Morning Peak

- Morning peak patterns are observed from the combination of traffic emissions and mixing height dilution.
- The morning rush hour occurs while mixing heights are relatively low, causing a peak in concentration while emissions outweigh dilution.
- By mid-morning, mixing height dilution has outweighed traffic emissions, reducing concentrations below their nighttime value and obscuring the remaining traffic emission patterns.
- Evening concentration increases are a consequence of mixing height lowering.

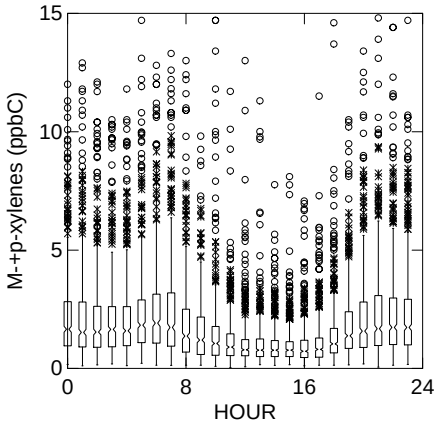
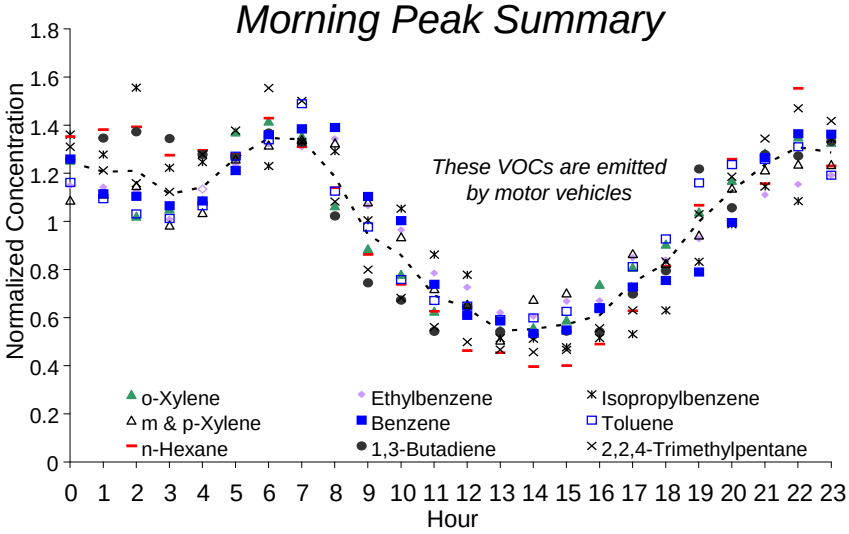


Figure shows notched box plot of m-&p-xylenes concentrations by hour at an urban site. Several years of data are included.

Diurnal Patterns

Morning Peak Summary

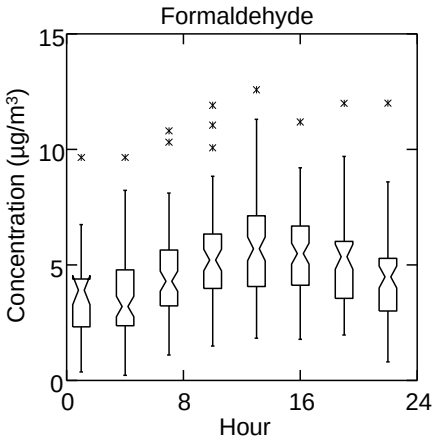


1990-2005 national hourly data normalized by site, pollutant, and day for all pollutants that exhibited a morning peak pattern on the national scale. Data were normalized as described in the approach to diurnal patterns.

Diurnal Patterns

Daytime Peak

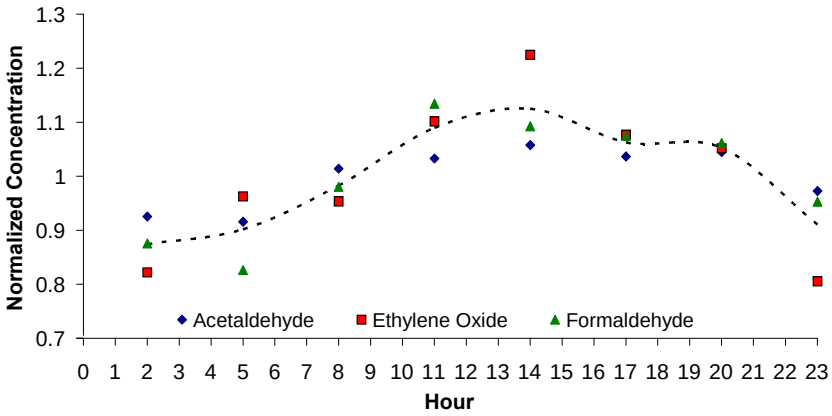
- The daytime pattern is driven by *in situ* secondary photochemical production mechanisms and mirrors the pattern of solar radiation.
  - Precursors of afternoon peak pollutants are typically emitted by motor vehicle sources and OH sinks. Afternoon peak pollutants experience daily dilution patterns in a manner similar to morning peak pollutants.
  - Secondary production of a pollutant (such as formaldehyde) must outweigh all these factors in order to create the observed pattern.



The figure shows notched box plots of national 3-hr formaldehyde concentrations by the middle sampling schedule.

Diurnal Patterns

Daytime Peak Summary

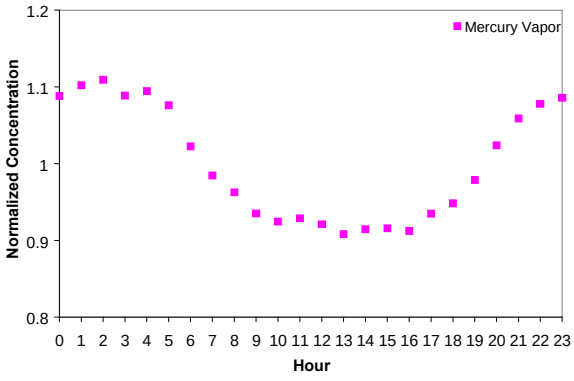


1990-2005 national-scale 3-hr duration data normalized by site, pollutant, and day for all pollutants that exhibit an afternoon peak pattern.

Diurnal Patterns

Evening Peak

- Mercury vapor is the only air toxic to exhibit a clear evening peak pattern in the air toxics investigated at the national level. However, data from only a few sites were available so this analysis may not be representative of a national pattern.
- Dilution appears to be the key factor affecting evening peak pollutants; emissions and sinks are likely invariant at the subdaily level.

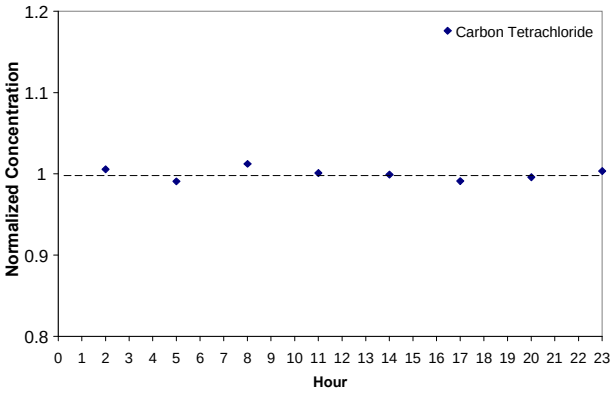


1990-2005 national hourly mercury vapor data normalized by site, pollutant, and day.

Diurnal Patterns

Invariant

- Invariant patterns are observed for global background pollutants (i.e., pollutant is no longer emitted).
- These pollutants show no sources or sinks and are evenly distributed worldwide so that transport and dilution have no effect on concentration.

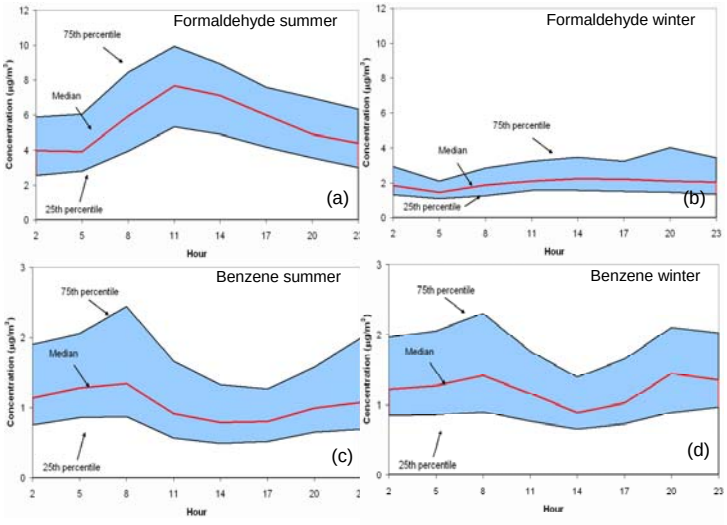


1990-2005 national 3-hr carbon tetrachloride data normalized by site, pollutant, and day. Carbon tetrachloride is the only pollutant to exhibit an invariant diurnal pattern on the national scale.

Diurnal Patterns

Seasonal Differences

- The diurnal pattern of formaldehyde is highly affected by season because the main production of formaldehyde depends on sunlight which is less abundant in winter months; thus, midday production decreases significantly during these months.
- The diurnal pattern of benzene shows less seasonal dependence because it is driven by diurnal meteorology that is consistent throughout the year and benzene is less photochemically reactive.



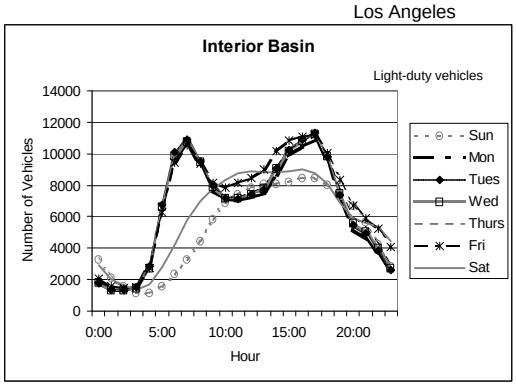
Diurnal Patterns
Summary

- Diurnal patterns of air toxics are influenced by sources, sinks, and dispersion processes that vary on a subdaily basis.
- Diurnal patterns are useful in classifying source type, transport, and reactivity of air toxics. These patterns can be used to improve exposure modeling, air quality modeling, and emissions inventories.
- Most air toxics data typically follow four diurnal patterns although many air toxics have not been characterized because of sampling and detection limitations.
  - Morning peak. Driven by mobile source emissions and mixing height dilution
  - Afternoon peak. Driven by secondary photochemical production
  - Nighttime peak. Driven by mixing height dilution
  - Invariant. Typical of global background pollutants that are not dependent on sources, sinks, transport, or dilution.
- If the diurnal pattern of a pollutant differs from the typical patterns shown at a national level, the analyst should explore possible reasons for the variation such as the presence of a nearby source.

Day-of-Week Patterns
Overview and Conceptual Model (1 of 2)

- Expectations
  - Emission sources that operate every day, 24 hours per day (e.g., refineries) will not show a day-of-week pattern.
  - Emission sources with lower emissions on weekends should lead to lower ambient weekend concentrations of the emitted air toxics. Traffic studies show that in many cities, light-duty vehicle activity is lower on Sunday compared to other days of the week.

Los Angeles

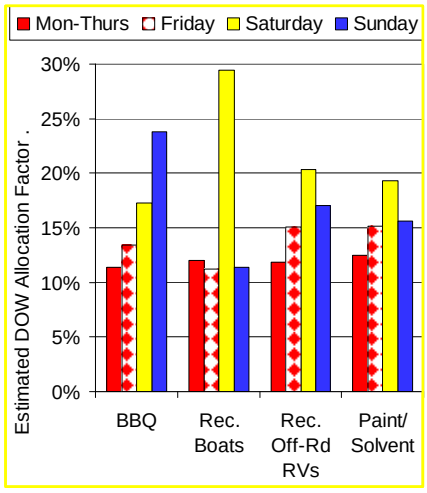


Chinkin et al., 2003

Day-of-Week Patterns

Overview and Conceptual Model (2 of 2)

- Emission sources with higher emissions on weekends should lead to high ambient weekend concentrations of the emitted air toxics. For example, studies in the Los Angeles area showed that recreational vehicle emissions may be higher on Saturdays.



Estimated allocation of residential emissions activity by day of week in Los Angeles (Coe et al., 2003)

Day-of-Week Patterns

Approach (1 of 2)

- Day-of-week patterns are typically constructed from 24-hr averages.
  - If subdaily data are available, look at data subsets (e.g., morning, afternoon).
  - When creating day-of-week trends of an air toxic that exhibits morning peak diurnal patterns, the rush hour peak data subset (i.e., 6 to 9 a.m.) will provide more information about the mobile source signature than the 24-hr average.
  - Mobile source signatures typically show day-of-week patterns, while mixing height dilution will occur on any day of the week.
  - 24-hr averages will be more heavily weighted by mixing height dilution and may obscure mobile source day-of-week trends.
- Investigate the day-of-week pattern of multiple statistics (e.g., 10th, 50th, and, 90th percentile) with the standard deviation or confidence intervals as a measure of uncertainty.
- If data are insufficient for each day to determine a pattern, weekday vs. weekend patterns may be investigated.

Day-of-Week Patterns

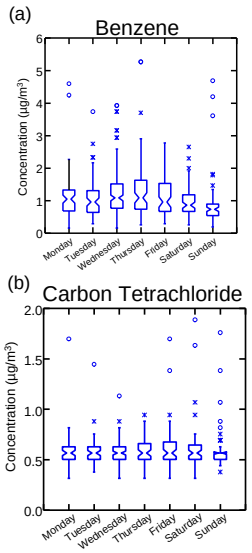
Approach *(2 of 2)*

- A sufficient number of records for each day of the week is needed to create a representative day-of-week pattern. The actual data requirements will vary depending on the analysis types and variability of the data, among other factors.
  - Statistically, decreasing the sample size increases the confidence interval (CI). In general, if the 95% CIs of two data subsets (e.g., weekend vs. weekday concentrations) do not overlap, there is good evidence that the subset population means are different; therefore, it will be more difficult to discern statistically significant patterns with smaller sample sizes.
  - Quantify patterns using the statistical treatments described earlier in this section.

Day-of-Week Patterns

Example *(1 of 2)*

- Benzene concentrations at an urban site are statistically significantly lower on Sunday. The concentrations on Saturday seem slightly lower, but differences are not statistically significant.
  - These results are consistent with our conceptual model of light-duty vehicle traffic.
- For carbon tetrachloride, we expect concentrations to be the same every day.
  - The central tendencies of the concentrations are consistent from day to day.

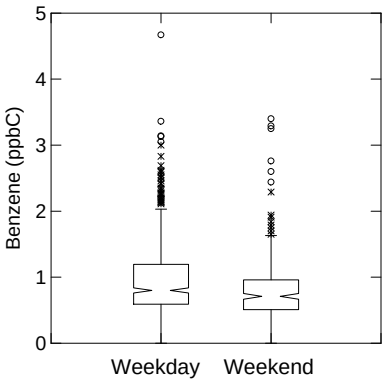


The figures show notched box plots of 24-hr concentrations by day of week at selected sites.

Day-of-Week Patterns

Example (2 of 2)

- Sometimes, not enough data are available to determine patterns by day of week—in some cases, the data can be combined into weekday vs. weekend groups.
- In the example, benzene concentrations at an urban site are lower on weekends than on weekdays (the difference in medians is statistically significant). These findings make sense because of the urban location of the monitor and lower motor vehicle emissions on the weekend compared to weekdays.



The figure shows a notched box plot of 1-hr benzene concentrations on weekdays vs. weekends at an urban site. All time periods were included—and weekend concentrations are statistically significantly lower than weekday concentrations.

Day-of-Week Patterns

Summary

- Typically, mobile source air toxics show the most obvious day-of-week pattern consistent with traffic patterns.
- In general, day-of-week patterns can be difficult to discern due to interference from other sources, sinks, or meteorology.
- A low number of samples can obscure underlying patterns.
- In exploratory investigations of national-level data, few non-mobile source air toxics showed a clear day-of-week pattern.
- Note that day-of-week patterns are highly dependent on the proximity of the monitor's site to sources, the emission sources' schedule, and meteorology (e.g., wind direction); site-level examinations may provide a better explanation.

Seasonal Patterns
Overview

Understanding seasonal differences in air toxics concentrations helps analysts

- Formulate or evaluate a conceptual model of emissions, formation, removal, and transport of an air toxic.
- Better understand source types.
- Continue to validate data, i.e., do data meet expectations for seasonal variation?
- Construct and interpret annual averages when a season's data are missing from the average (e.g., if the data for a winter quarter are missing, what biases in the annual average can be expected?).

Seasonal Patterns
Conceptual Model (1 of 2)

- Cool season expectations
  - Mixing heights are lower in the cold months. Low mixing heights create less air available for pollutant dispersion which causes higher ambient concentrations.
  - Temperatures are lower and sunlight is reduced in cold months. This combination can lead to a reduction in evaporative emissions (e.g., gasoline) and reduced photochemistry. Reductions in temperature and sunlight also limit formation of hydroxyl radicals which efficiently oxidize many air toxics.
  - Typically more precipitation occurs during winter months and reduces dust emissions.

Seasonal Patterns
Conceptual Model *(2 of 2)*

- Warm season expectations
  - Mixing heights are higher in warm months, allowing more dilution and transport of air toxics which, in turn, reduces ambient concentrations.
  - Higher temperatures and increased sunlight in warm months lead to an increase in evaporative emissions and photochemistry.
  - Conditions are typically drier, producing more dust.
  - Depending on the site, recreational activities such as boating may increase in summer and result in higher gasoline-related emissions.
  - Wildfire activity can also cause an increase in concentrations of pollutants emitted in smoke.

Seasonal Patterns
National Trends

- Seasonal patterns observed at a national level are shown in the table.
- These air toxics were selected because there were sufficient data for analyses.
  - Minimum of three valid seasonal averages by site and year
  - At least 20 monitoring sites meeting the above criteria
  - Additionally, limited to pollutants investigated in diurnal variability and annual analyses to focus on similar pollutants.
- Most of the VOCs, with the exceptions of styrene and isopropylbenzene, are cool-season pollutants as expected.
- We are not sure why carbon tetrachloride shows a warm season peak—we expected it to be invariant. No obvious data issues suggested this pattern.

McCarthy et. al, 2007

Seasonal Patterns

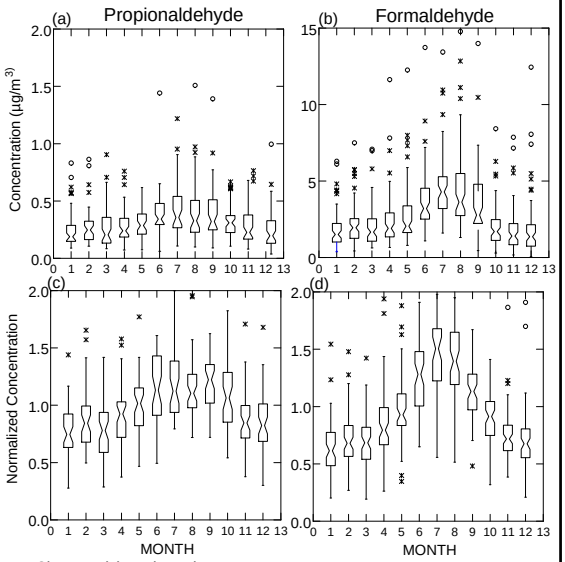
Approach

- Investigate seasonal variability patterns using normalized monthly and/or quarterly averages.
- Keep track of the percentage of data below detection; pollutants and years for which >85% of data were below detection may result in too much bias to draw conclusions.
- Preferably, inspect monthly data for seasonal patterns if sufficient data are available.
- Normalize the data using the average value for each year, site, and pollutant.
  - Calculate an annual average for each year, site, and pollutant.
  - Divide the corresponding monthly or quarterly average by the annual average.
- Investigate seasonal patterns of normalized data using notched box plots or summary statistics with a measure of confidence (e.g., standard deviation or confidence intervals).

Seasonal Patterns

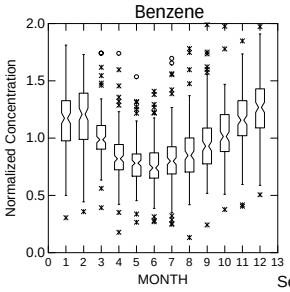
Using Normalized National-Scale Data

- Propionaldehyde and formaldehyde show higher concentrations in summer (Figures a and b).
- However, normalized concentration patterns (Figures c and d) show that the monthly pattern of formaldehyde is more significant than that of propionaldehyde.
- On a relative basis, Figures c and d show that concentrations of formaldehyde are nearly three times higher in the summer than in winter.

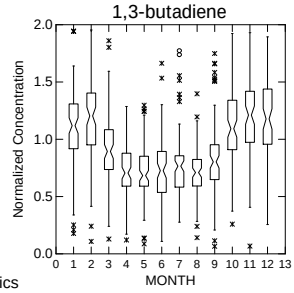


Seasonal Patterns
Cool Season Peak

- Cool seasonal patterns are generally observed because mixing heights are lower in winter and the enhanced removal by photooxidation observed during summer is absent.
- Heating-related emissions, such as wood burning, will typically be higher during winter months, contributing to increased concentrations of some air toxics.
- Benzene and 1,3-butadiene, two mobile source air toxics, show cool season peaks on the national scale.

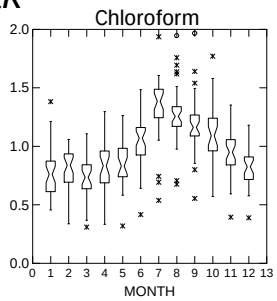


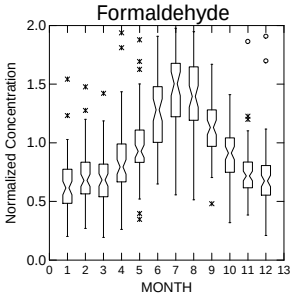
Figures show normalized monthly national concentration distributions for 2003-2005.



Seasonal Patterns
Warm Season Peak

- To display a warm peak pattern, summertime sources (emissions or secondary production) must significantly outweigh the higher mixing heights that occur during warm months.
- Chloroform emissions from water treatment processes and swimming pools may be enhanced during summer months, explaining the observed pattern.
- It has been estimated that 85-95% of formaldehyde concentrations originate from secondary photochemical production, which supports the observed warm season peak (Grosjean et al., 1983).



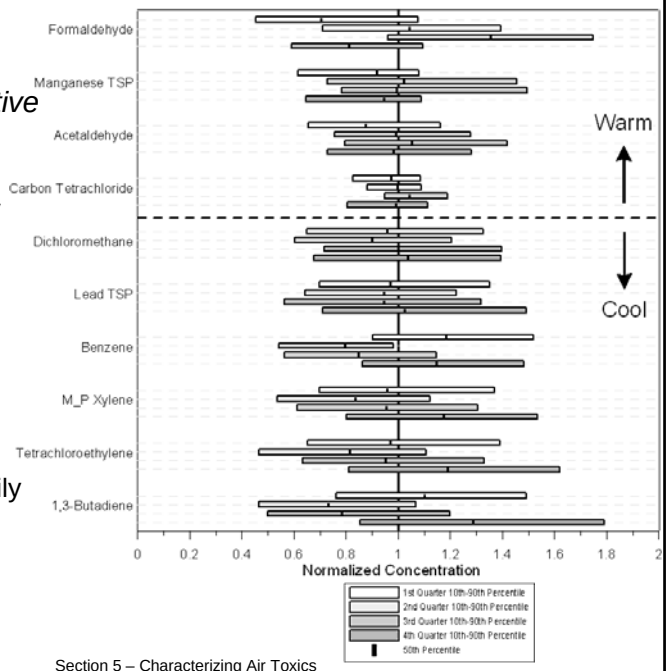


Figures show normalized monthly national concentration distributions for 2003-2005.

Seasonal Patterns

A National Perspective

- Warm season peaks are likely due to secondary photochemical production and dust; it is unclear why carbon tetrachloride shows a warm season peak.
- Cool season peaks are primarily due to lower mixing heights in the winter.



Seasonal Patterns

Summary

- Three seasonal patterns were observed at a national level.
  - *Warm season peak.* Photochemical production of secondary air toxics (e.g., formaldehyde and acetaldehyde) can be important at some sites. Concentrations (e.g., manganese) may also be high because of dust events and seasonally increased emissions (e.g., chloroform).
  - *Cool season peak.* Concentrations can be high because of lower inversions, changes in emissions through the use of wood-burning or fuel oil for home heating, and reduced photochemical reactivity.
  - *Invariant*. Invariant seasonal patterns are not commonly observed, but are typical of global background pollutants that are not affected by emissions changes or dilution which cause seasonal patterns of other air toxics.
- Understanding seasonal patterns assists in air toxics data analysis by providing insight into the chemistry, sources, and transport of air toxics. Deviation from expected seasonal patterns at a site may indicate additional sources of interest or transport.

Spatial Patterns
Overview

- Air toxics data are typically collected in urban locations. Given the large number of air toxics, their often disparate sources, and the wide range of chemical and physical properties, understanding spatial patterns and gradients is important.
- Understanding these gradients may help us
  - Improve monitoring networks. (Are we measuring in the right places to meet network objectives? Do we have the right number of monitors?)
  - Improve emission inventories. (How finely do emissions need to be spatially allocated?)
  - Improve models, including exposure models. (Are gradients in pollutants being properly represented in the model?)
  - Identify contributing sources. (Are concentrations higher when winds are predominantly from the direction of a source?)

Spatial Patterns
Conceptual Model

The concentration of a given species at any location is determined by local production, local sinks, and transport.

- *Production.* Local emissions—higher emissions lead to higher concentrations.
- *Loss*. Local removal (chemical or deposition)—reactive compounds and large particles are removed faster resulting in lower concentrations.
- *Transport*. Movement of species in the atmosphere—pollutants from sources are dispersed or diluted; local concentrations can either increase or decrease.

$$\frac{d(\text{Concentration})}{dt} = \text{Production} - \text{Loss} + \text{Transport}$$

Spatial Patterns
Methods

- Calculate one site average value for each air toxic for the time period of interest. This method removes temporal variability and focuses on spatial patterns.
  - The method is only valid if sites are temporally comparable. If not, results may be driven by a mixture of temporal and spatial patterns and will be difficult to interpret.
  - Construct averages from valid aggregates. For example, if data are available for 2003-2005, you might first calculate the three valid annual averages then aggregate these averages to one site average. If data are insufficient to create valid annual averages, use valid seasonal or monthly averages. Note that site average values may be biased by temporal patterns if data are not representative of the full year. Relative spatial comparisons are still valid as long as data are available for all sites during the same time period.
  - It is best to use multiple years of data to mitigate meteorological effects.
  - Keep track of the percent of data below detection for each site average.

Spatial Patterns
Methods *(1 of 2)*

- Visualize concentration ranges by plotting summary statistics for each pollutant.

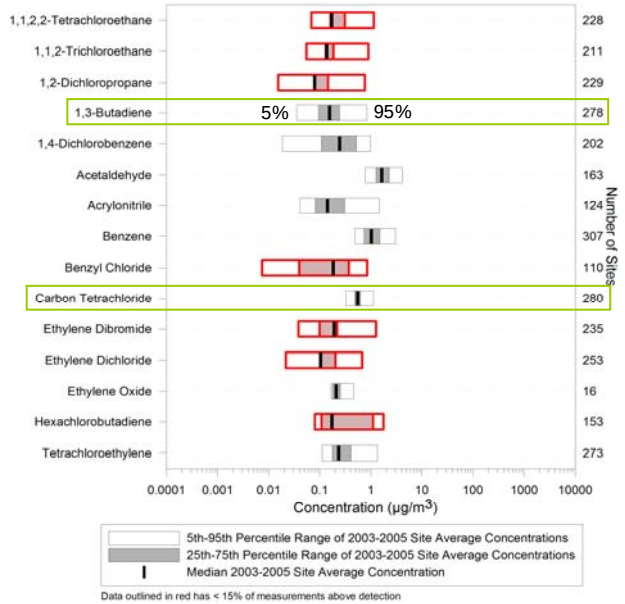
  Supplementary data, such as risk levels, remote background concentrations, and method detection limits (MDLs), are useful to put concentration data into perspective.
- Visualize site level concentrations using a mapping program to overlay supplementary data, such as the percent of data below detection, to enrich conclusions.
- The visualization methods may illuminate site-level data anomalies which become apparent upon comparison to other sites.

National Concentration Plots

Overview *(2 of 2)*

- The following national site average concentrations for 2003-2005 air toxics concentrations show one way to visualize summary statistics and supplementary data.
- The following figures show the 5th, 25th, 50th (median), 75th, and 95th concentration ranges by pollutant; supplementary data are then overlaid as a progression. Wide ranges in concentration across sites indicate greater spatial variability of that pollutant.
- The number of sites included are shown on the right axis for each pollutant.
- Pollutants outlined in red represent <15% of samples nationally above their respective MDLs. The distribution of concentrations for these pollutants are mostly based on MDL/2 and should not be considered quantitative.

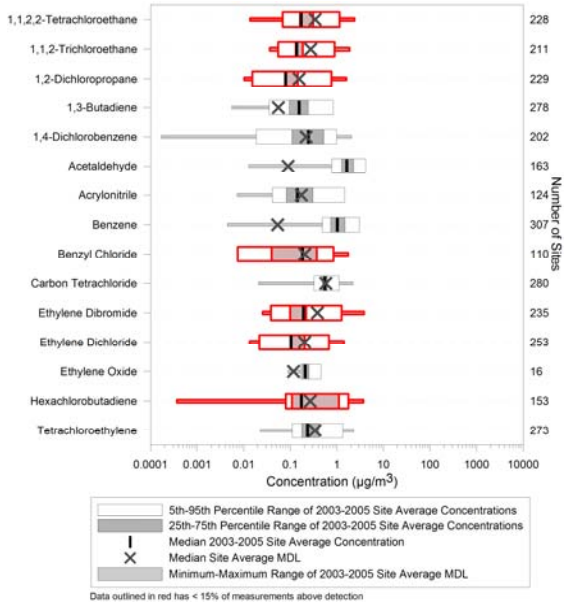
National Concentration Plots



Interpretation

- Spatial variability is represented by the width of the bar—nationally, air toxics concentrations typically varied by a factor of 3 to 10.
- High spatial variability of 1,3-butadiene is due to its relatively high reactivity.
- Conversely, carbon tetrachloride shows less spatial variability due to its low removal rate from the atmosphere and the absence of domestic emissions.

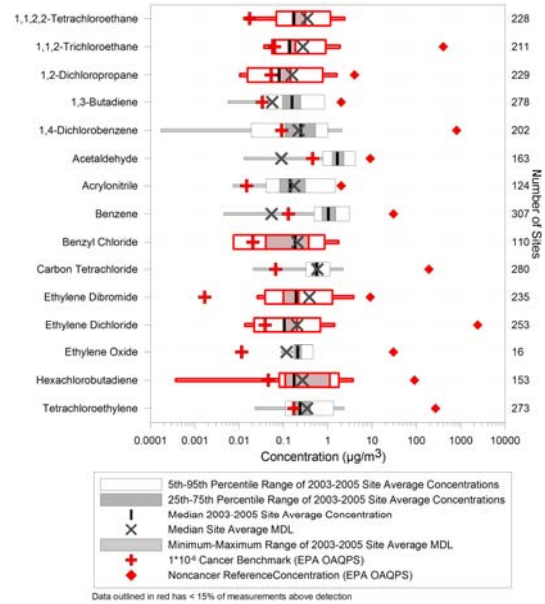
National Concentration Plots



Adding MDLs

- MDL ranges (thin lines) and median MDLs (X's) are added to the plot to illustrate how well pollutants are monitored.
- The minimum-maximum range of MDL concentrations and the median MDL concentration for a 2003-2005 site average are shown.
- The median concentration of the pollutants outlined in red are always below the median MDL.

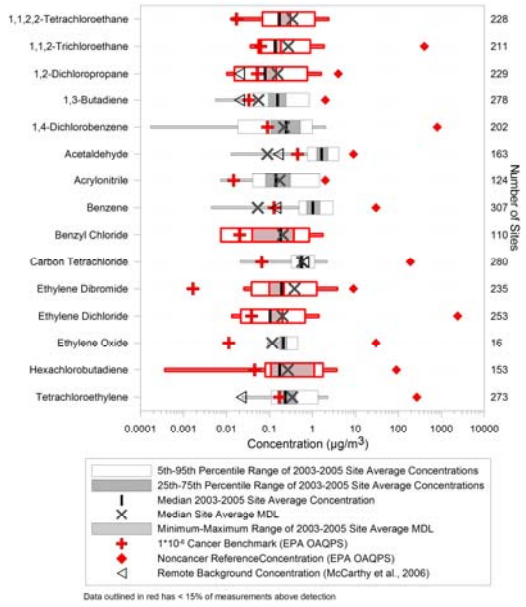
National Concentration Plots



Risk Levels

- Chronic exposure concentration associated with a 1-in-a-million cancer risk (red crosses) and noncancer reference concentrations (red diamonds) are added to the plot to show a relationship to human health.
- National measured annual average air toxics concentrations are usually above the chronic exposure concentration associated with a 1-in-a-million cancer risk and below noncancer reference concentrations.
- Note that the pollutant concentration ranges outlined in red may actually be below levels of concern, but the data are not resolved well enough to characterize risk.

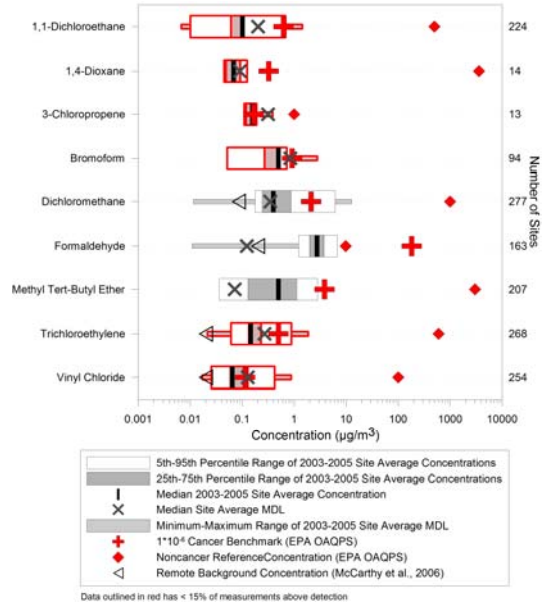
National Concentration Plots



Remote Background

- Remote background concentrations (triangles) are added to the plot to show the lowest levels expected to be seen in the remote atmosphere; urban concentrations of most air toxics should not typically fall below this value.
- As expected, most air toxics are a factor of 5-10 above their remote background concentrations, with the exception of carbon tetrachloride – the only air toxic dominated by background concentrations.

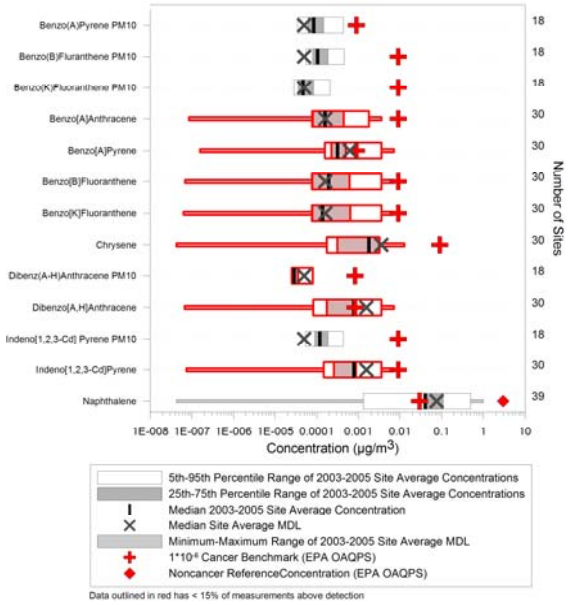
National Concentration Plots



Additional VOCs

- These VOCs are usually below their 1-in-a-million cancer risk level and noncancer reference concentrations.
- Note that the 1-in-a-million cancer risk level for formaldehyde was changed in 2004 from 0.08 to 182 µg/m^3. 1-in-a-million cancer risk levels plotted are provided by EPA OAQPS.
- See the NATA website for more information regarding risk characterization, http://www.epa.gov/ttn/atw/nata1999/nsata99.html. For example, analysts can investigate the potential for health effects from air toxics by target organ/system.

National Concentration Plots

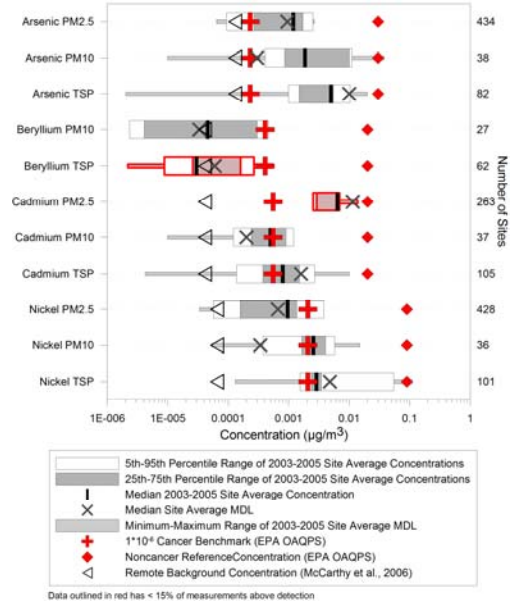


*SVOCs**

- The figure indicates that most SVOCs are below their 1-in-a-million cancer risk levels. However, the data quality for many SVOCs is poor—less than 15% of measurements are above the detection limit.
- Only naphthalene is above its 1-in-a-million cancer risk level at most sites.
- Routine measurements of SVOCs are relatively rare across the United States.

* semi-volatile organic compounds

National Concentration Plots



Metals

- All metals are well below their noncancer reference concentrations.
- With respect to 1-in-a-million cancer risk levels, arsenic is the most important of these metals, with more than 75% of sites measuring concentrations above the 1-in-a-million cancer risk level for $PM_{2.5}$.
- $PM_{2.5}$ metals are more commonly measured in rural and remote locations via the IMPROVE network; therefore, the lower range of $PM_{2.5}$ concentrations commonly overlaps remote background concentrations.

National Concentration Plots
Summary

- The national concentration plots provide perspective for local, state, regional, and tribal analysts to see how their data compare.
- Air toxics concentrations typically vary spatially by a factor of 3 to 10, depending on the pollutant.
- Almost all air toxics are below noncancer reference concentrations (except acrolein, not shown).
- At a national level, some air toxics are above their respective associated with a 1-in-a-million cancer risk (http://www.epa.gov/ttn/atw/toxsource/table1.pdf).
- Most air toxics are well above their remote background concentrations.

Spatial Patterns – Maps
Overview

- National concentration plots placing air toxics in a national context provide useful information for quantifying air toxics spatial variability. To view spatial patterns, though, it is also useful to plot site-level data on a map.
- Example maps of site average and risk-weighted concentrations (i.e., risk estimates based on ambient measurements) from 2003 through 2005 are shown in the following slides.
  - These maps help analysts characterize the national picture of air toxics and are most useful in a qualitative sense to compare among sites, look for spatial patterns, and note data anomalies.
  - The maps also illustrate a method of displaying data that can be applied to sites within a city, state, or region.
- In the examples, concentrations are displayed as proportional symbols which are color-coded to impart additional information.
- Maps are useful for communicating a range of information—similar depictions can be made using risk estimates, percent change per year, or ratios—over a range of spatial dimensions (e.g., city, state, or region).
- The volume of concentrations is indicated on the maps by the diameter of the circle (the three sizes in the map legends) while the underlying percent of data below detection is signified by color.

Spatial Patterns – Maps

Benzene Concentrations 2003-2005

The largest circle on the map corresponds to 17 µg/m^3.

- Benzene concentrations have ambient measurements above detection across the country with only a few exceptions (i.e., 0-50% of the measurements at most sites are below detection).
- Concentrations are consistent for areas dominated by mobile sources (e.g., the Northeast and California) while isolated high concentrations generally coincide with significant point source emissions of benzene such as refineries and coking operations.

Spatial Patterns – Maps

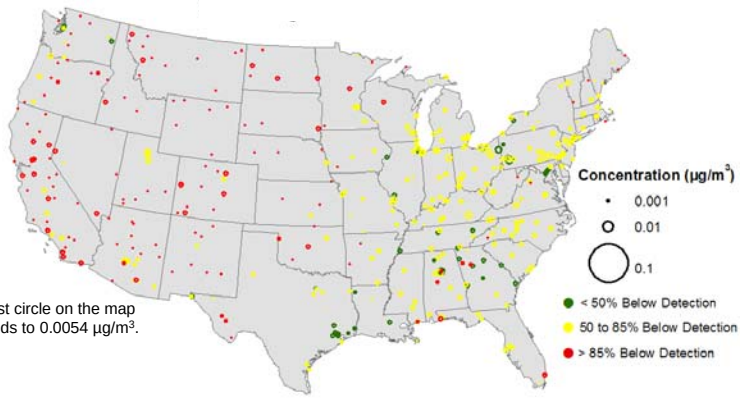
1,3-Butadiene Concentrations 2003-2005

The largest circle on the map corresponds to 6.6 µg/m^3.

- The ability to obtain 1,3-butadiene concentration measurements above the MDL across the United States varies (note all the red circles and their varying sizes).
- Higher concentrations generally coincide with locations of known point source emissions.
- Differences in monitoring methods and methods application have resulted in large differences in reported MDLs across the United States.

Spatial Patterns – Maps

Arsenic $PM_{2.5}$ Concentrations 2003-2005

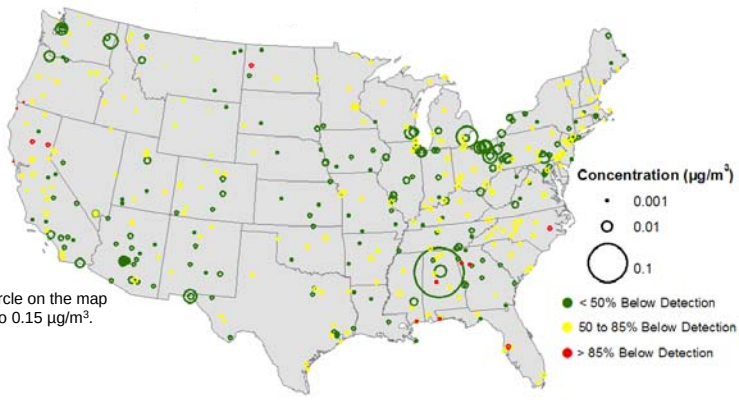


The largest circle on the map corresponds to 0.0054 µg/m³.

- Arsenic concentrations are widely measured across the United States, and the entire range of data availability is observed from more than 50% of data above detection to less than 15% above detection.
- Significant MDL differences between networks make determining spatial patterns difficult.
- In general, concentrations are higher and more often above detection in the eastern half of the country.

Spatial Patterns – Maps

Manganese $PM_{2.5}$ Concentrations 2003-2005



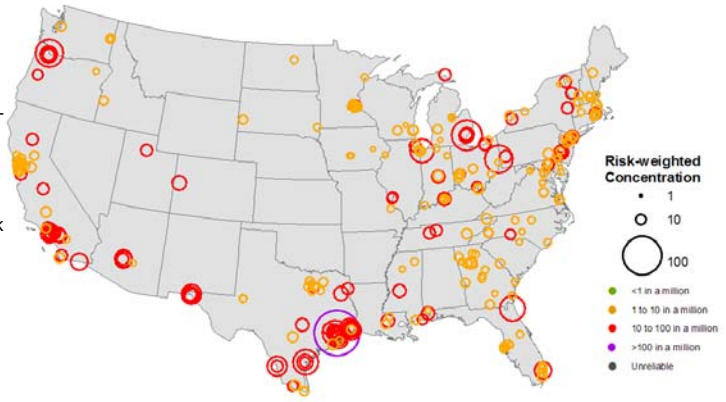
The largest circle on the map corresponds to 0.15 µg/m³.

- In contrast to arsenic, manganese concentrations are widely measured across the country with most data recorded above the detection limit.
- Concentrations vary spatially and several "hot spots" can be identified that may lend themselves to additional investigation at a site level.

Spatial Patterns – Maps

Benzene Risk Estimates 2003-2005

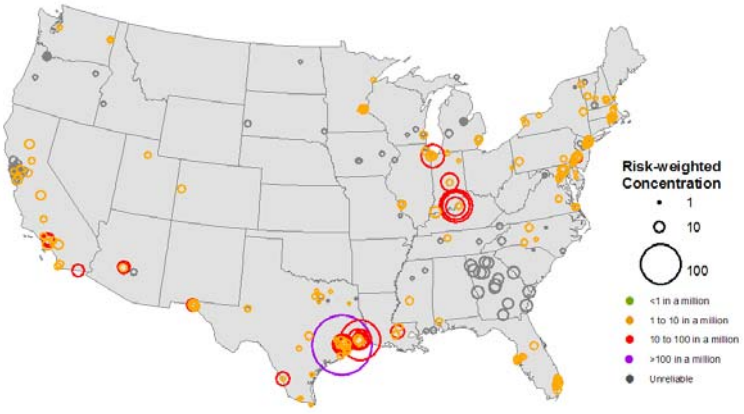
Note:
2003-2005 average concentrations are divided by the 1-in-a-million cancer risk concentration. Circle diameter represents this ratio while the chronic risk assessment is indicated by color. Sites at which >85% of data are below detection are considered unreliable (grey).



Benzene risk associated with measured ambient concentrations is almost always above the 1-in-a-million cancer risk level across the United States. Many areas are also above the 10-in-a-million cancer risk.

Spatial Patterns – Maps

1,3-Butadiene Risk Estimates 2003-2005



Where measured reliably, 1,3-butadiene concentrations are almost always above the 1-in-a-million cancer risk level. Some areas do not measure concentrations well enough to evaluate risk (grey symbols). Highest concentrations are located in areas with known point source emissions (e.g., Houston and Louisville).

Variability Within and Between Cities
Overview

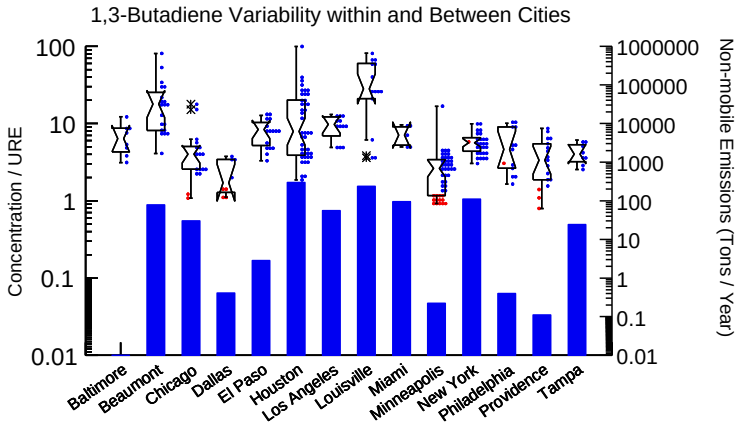
- The aim of such analysis is to understand how representative a given site is with respect to air toxics concentrations in a city.
  - What is the variability of air toxics concentrations within cities and what are the implications for aggregating data at the city level?
  - Where do sites need to be located to accurately characterize variability within a city?
  - How many sites are needed to characterize spatial variability within a city?
  - How does within-city variability differ across cities?
- There may also be interest in assessing variability in air toxics from city to city.
  - What are the concentration distributions across all monitoring sites?
  - Do specific cities, states, or regions have demonstrably higher or lower concentrations?
  - Do demonstrably lower concentrations occur at rural and remote sites?
  - Are concentration differences associated with monitoring agency differences?

Variability Within and Between Cities
Approach

- Valid annual averages are calculated for each monitor in a city. To reduce noise from year-to-year changes (e.g., the effect of meteorology), it is best to use multiple years of data when available. The national study used 2003-2005 data.
- Data can be visualized using notched box plots by air toxic, city, and year. If variation between years at a given city is minor, notched box plots by air toxic and city only can be constructed to increase the amount of data.
- Advanced plotting techniques
  - Include a color-coded measure of the percent of data below detection to understand the reliability of the data.
  - Divide annual averages by the cancer risk (such as the URE) to show variation in risk estimates within and between cities.
  - Include a measure of relevant emissions by city to explain possible reasons for high or low concentrations

Variability Within and Between Cities

Example



Benzene risk-weighted annual average variation for 2003-2005 for selected U.S. cities along with non-mobile emissions. Notched boxes include annual averages for each monitor within a city, providing within-city variation. Dots over the notched boxes show the individual data points and whether they are above (blue) or below (red) the average MDL. Bars show county-level non-mobile emissions of 1,3-butadiene from EPA's AirData.

Variability Within and Between Cities

National Perspective

- At a national level, spatial variability within cities was found to be pollutant- (or pollutant group-) specific.
- Most toxic measurements are highly variable within cities; risk values span an order of magnitude within some cities.
- The spatial variability between cities is a good metric to estimate the variability within cities *a priori*. Spatial variability analysis helps set expectations for sampling in a new city.
- Cities with point source emissions (e.g., Houston) showed higher within-city variability than those dominated by area/mobile sources (e.g., Los Angeles).

Hot and Cold Spot Analysis
Overview

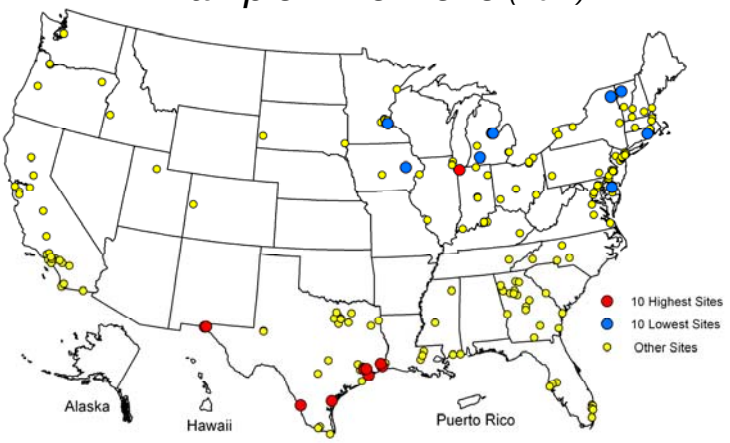
- Hot and cold spot analysis is an investigation of sites where the highest and lowest concentrations occur.
- The objective of this analysis includes
  - Data validation. The highest and lowest values may be due to some type of error, possibly reporting.
  - Comparison to the spatial conceptual model. Are the highest concentrations consistent with known sources, transport, and dispersion?
  - Risk screening. Where are the toxic concentrations highest?

Hot and Cold Spot Analysis
Approach

- Create valid annual averages for each site and pollutant and rank each site by its concentration (highest to lowest). The number of high- and low-ranked concentration sites investigated depends on the number of available sites.
- Map all sites, marking the highest and lowest ranked sites to investigate spatial variation.
- Identify why high or low concentrations occur at those sites and whether the occurrence of those concentrations meets expectations.
  - Review metadata about the sites (e.g., Google Earth images, local emissions, and meteorology). Do concentrations meet spatial conceptual models with respect to scale, sources, transport, and dispersion?
  - Inspect time series of concentration and MDL (e.g., is the value stuck, are data outliers driving the average, is the MDL higher than the concentrations at an average site?).

Hot and Cold Spot Analysis

Example – Benzene (1 of 2)

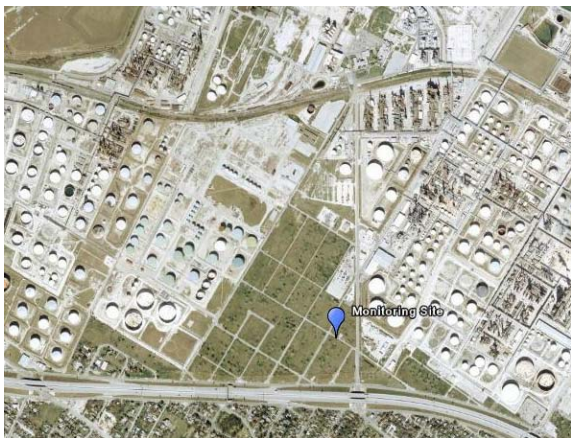


Sites with the 10 highest and 10 lowest benzene concentrations based on 2003-2005 annual averages. The sites ranked lowest were either a result of data reporting or siting issues or were located in rural areas, consistent with our conceptual model of low concentrations.

Hot and Cold Spot Analysis

Example – Benzene (2 of 2)

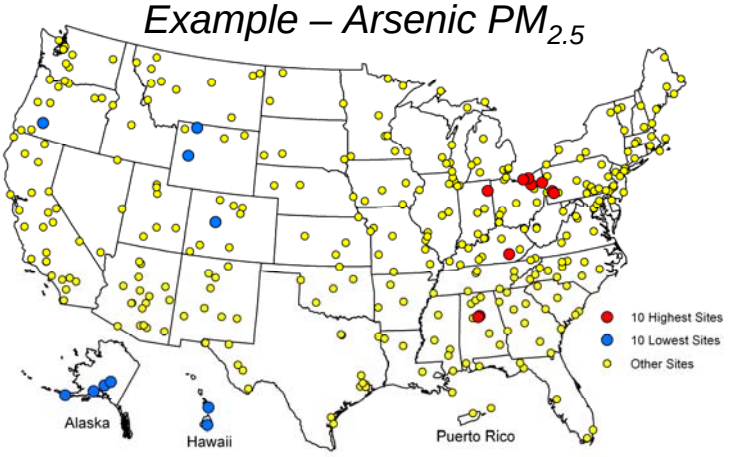
- The sites measuring the highest concentrations in the nation were dominated by nearby point source emissions; the site identified in the figure measured the second highest benzene concentration in the nation.
- This site is very close to two refineries that emit a significant amount of benzene each year according to the NEI.



Google Earth image of the site with the second highest benzene concentrations in the United States. Refineries to the right and left emitted 84,000 and 44,000 lbs of benzene in 2004 (NEI).

Hot and Cold Spot Analysis

Example – Arsenic $PM_{2.5}$



The 10 highest and 10 lowest arsenic $PM_{2.5}$ concentrations based on 2003-2005 annual averages. Conceptually, we would expect Arsenic $PM_{2.5}$ concentrations to be highest in locations dominated by point source emissions, especially smelting and coal combustion. The highest sites are consistent with this conceptual model. The lowest sites are located in extremely remote locations such as Alaska and US national parks which is reasonable for the lowest arsenic $PM_{2.5}$ concentrations.

Urban vs. Rural Analysis

Overview

- Measured concentrations can be highly dependent on individual monitor locations, geography, emissions sources, and meteorological conditions (e.g., prevailing winds).
- **Urban areas – conceptual model**
  - Urban areas contain sources of air toxics that result in increased concentrations and, in some cases, "hot spots" (areas with disproportionately higher concentrations) in the spatial pattern.
  - Urban concentrations vary greatly from day to day due to the mix of local sources and meteorology.
- **Rural areas – conceptual model**
  - Rural areas typically have fewer sources of air toxics. Air toxics concentrations that are transported from urban locations are typically near background levels when they reach rural areas (a function of source strength, distance, and the lifetime of the pollutant).
  - Concentrations do not vary consistently day to day. Daily and seasonal patterns that are dependent on meteorological conditions may still be observed.

Urban vs. Rural Analysis
Approach *(1 of 3)*

- Characterize each site as urban or rural.
  - If available, start with EPA urban/rural designations listed in AQS (note that these designations are not always up to date).
  - Verify the designations using Google Earth—they may be outdated or incorrect.
  - Be wary of defining a site using population density, total county population, or other metrics—local knowledge of the site appears to be the best way to identify site characteristics.
- Identify pollutant availability and time period for each site.
  - The goal is to have a spatially representative mix of urban and rural sites measuring a pollutant over the same time period. This mix can be a challenge since toxics are more commonly measured in urban locations.
- Choose pollutant/site combinations that are spatially and temporally representative.
  - Pollutant-specific monitoring time periods need to be the same for site comparison; otherwise differences in observed concentrations could be biased by seasonal or inter-annual patterns.

Urban vs. Rural Analysis
Approach *(2 of 3)*

- Estimate valid 24-hr averages for the sites, pollutants, and time periods of interest.
  - Characterize all concentration averages that are below the associated average MDL
- Visualize the data by site by preparing plots of data distributions, including some measure of the data below detection.
  - Look for differences in concentrations.
- Identify statistically significant differences in urban vs. rural site concentrations.

Urban vs. Rural Analysis

Approach *(3 of 3)*

- Summarize the results with a focus on neighboring urban vs. rural sites.

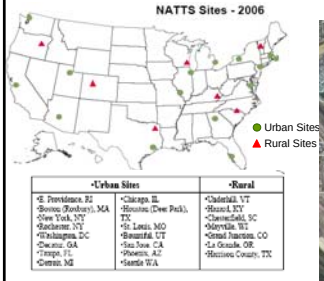
  Which urban and rural sites measured significantly higher or significantly lower concentrations, if either? Which showed no difference?

- Investigate data that do not meet expectations (e.g., concentrations at a rural may be significantly higher than those at a nearby urban site).
  - Are the sites representative of the area (i.e., compare to other urban or rural sites)?
  - Are there monitor location abnormalities (e.g. local terrain, prevailing winds)?
  - Are there measurement methods or MDL differences between the sites?
  - Is there a significant rural emissions source?
  - Are possible data errors or outliers driving the trend?

Urban vs. Rural Analysis

Example – Investigating Urban vs. Rural Sites *(1 of 2)*

- When beginning an urban vs. rural analysis, verify that sites are properly designated "urban" or "rural".
- A map of urban and rural NATTS sites across the United States is shown with Google Earth pictures of two of the rural sites—Grand Junction, CO, and La Grande OR.
- Both sites are designated as rural in AQS, but the Colorado site appears urban in character, and air toxics concentrations may not conform to the model for a rural site.
- The Oregon site is rural based on the observation that the surrounding area is mainly farmland.



Grand Junction, CO

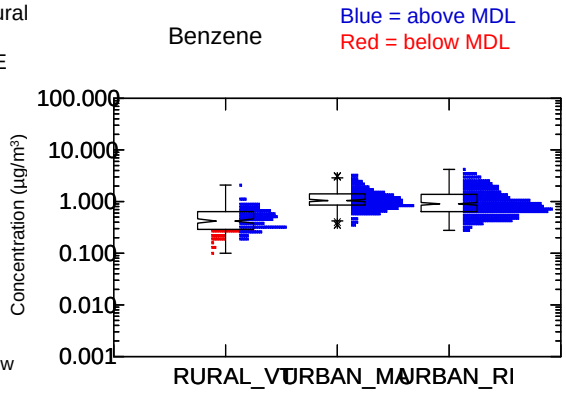
La Grande, OR

Two rural sites in the NATTS network. Images obtained from Google Earth.

Urban vs. Rural Analysis

Example – Investigating Urban vs. Rural Sites (2 of 2)

- Benzene concentrations at a rural Vermont site compared to concentrations at two urban NE sites.
- The rural site shows statistically significantly lower concentrations.
- If a site does not fit an urban or rural definition as expected, check for
  - Measurement method or MDL differences
  - Local emissions sources
  - Time series comparing the two sites with color-coded data below detection.
  - Evaluate data subsets when both sites have measurements above detection. Does this tell a different story?



The example figure is from an analysis of NATTS sites using 2003-2005, 24-hr average, benzene data. The box plots encompass all data while the overlaid dot density shows each data point and whether it is above or below detection (blue vs. red).

Spatial Patterns

Summary

- Benzene, 1,3-butadiene
  - Concentrations vary around the United States and are high in urban areas where there are more mobile sources. The highest concentrations of these two toxics are found in areas influenced by point source emissions in addition to mobile sources.
  - Within- and between-city variability is generally near a factor of 5.
- Carbonyl compounds
  - Carbonyl compounds are measured widely and show very consistent concentrations across the nation. This consistency is due to the dominant secondary formation mechanism.
  - Within and between-city variability is relatively low with few exceptions.
- $PM_{2.5}$ metals
  - The spatial character of $PM_{2.5}$ metals is difficult to determine due to differences in measurement methods and MDLs among monitoring networks.
  - Overall, it seems that concentrations are slightly higher in the eastern half of the United States.

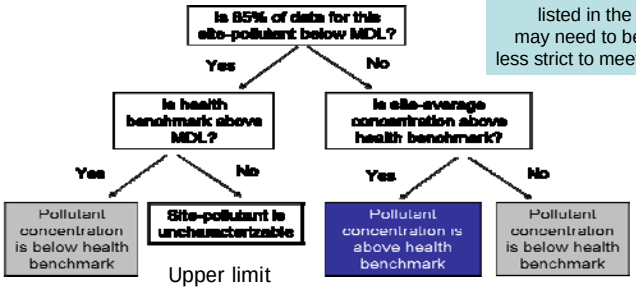
Risk Screening

Overview

- A key use of air toxics data is to compare annual average concentrations to risk levels for context.
- Risk screening can help identify air toxics of concern.
- Information to consider in conducting a risk screening is available, e.g., in "A Preliminary Risk-Based Screening Approach for Air Toxics Monitoring Data Sets"
- For information on a more thorough air toxics risk assessment, see the Air Toxics Risk Assessment Library

Risk Screening

Approach



The % of data below MDL listed in the first box may need to be stricter or less strict to meet your DQOs.

- For this first level of screening, site average concentration data from the most recent year (s) (e.g., 2003-2005) were used to identify the number of sites at which a pollutant was definitively above or below the relevant EPA OAQPS 1-in-a-million cancer risk as found at: http://www.epa.gov/ttn/atw/toxsource/summary.html. Results are ranked by screening level.
- Air toxics were also noted if most site concentrations could not be characterized as above or below this level with certainty.

Risk Screening
Example

Decreasing risk → (Red = Notes)

Concentrations above 1-in-100,000 cancer risk level at >25% of sites	Concentrations above 1-in-1,000,000 cancer risk level at >50% of sites	Concentrations above 1-in-1,000,000 cancer risk level at 10-50% of sites
Benzene	Arsenic ($PM_{2.5}$ and PM_{10})	Tetrachloroethylene
Acrylonitrile [1]	Acetaldehyde [2]	Cadmium (PM_{10} and TSP)
	Carbon tetrachloride	Naphthalene
	1,3-Butadiene	1,4-Dichlorobenzene
	Nickel (PM_{10} only)	Benzyl Chloride
	Chromium (estimated Cr VI from Cr $PM_{2.5}$)	

- This table displays only pollutants whose concentrations were monitored well enough to support a conclusion that they were above the relevant risk levels for pollutants for which at least 20 monitoring sites existed in the United States from 2003-2005.
- We are confident these cancer-risk pollutants are at or exceed the categories of cancer risk (i.e., may be higher, but are not lower)

[1] May have sampling issues biasing concentrations high, magnitude unknown
[2] May have sampling issue biasing concentrations low by a factor of 2 (Herrington et al., 2007)

Risk Screening
Results at a National Level

At a regional, state, or local level, results may differ. This table provides a context for comparing local results.

Higher confidence – chronic cancer risk (ordered by importance)	Lower confidence – chronic cancer risk (ordered by importance)	High confidence – chronic and acute noncancer hazard
Benzene	Ethylene dibromide	Acrolein
Acrylonitrile[1]	1,1,2,2-tetrachloroethane	
Arsenic	1,2-dibromo-3-chloropropane	Local chronic hazard
Acetaldehyde[2]	Ethylene oxide	Formaldehyde
Carbon tetrachloride	Ethylene dichloride	Manganese
1,3-Butadiene	Hexachlorobutadiene	Acrylonitrile[1]
Nickel[3]	1,2-dichloropropane	1,3-Butadiene
Chromium[3]	1,1,2-trichloroethane	Nickel
Tetrachloroethene	Vinyl chloride	
Naphthalene	Trichloroethylene	
Cadmium	Benzo[A]pyrene	
1,4-Dichlorobenzene	Dibenzo[A,H]anthracene	
Benzyl chloride	3-Chloropropene	

[1] May have sampling issues biasing concentrations high, magnitude unknown
[2] May have sampling issue biasing concentrations low by a factor of 2 (Herrington et al., 2007)
[3] Concentrations adjusted to estimate toxicity based on subset expected to be in either Cr VI or Nickel subsulfide.

Summary

Checklist for Ways to Characterize Air Toxics

Temporal Characterization

- ❑ The general procedure for investigating temporal patterns is the same for all aggregates.
  - Prepare valid concentration and normalized temporal aggregates and summary statistics.
    - Normalization allows comparison between sites and pollutants even if absolute concentration values vary widely.
    - Keep track of the amount of data below detection.
  - Plot data with notched box plots or line graphs of multiple statistics (e.g., mean vs. 90th and 10th percentiles) with confidence intervals.
  - Characterize patterns by pollutant
    - Do patterns fit your conceptual model?
    - Are they statistically significant?
  - Investigate unexpected results
- ❑ Diurnal patterns – If alternate sampling schedules are used, calculate the weighted average by the most representative sampling hour; otherwise, diurnal patterns may be obscured.
- ❑ Day-of-week patterns – Examine data availability by day-of-week.
  - If sufficient data exist for each day of the week, examine day-of-week patterns.
  - If insufficient data exist, weekday vs. weekend patterns can be used.
- ❑ Seasonal patterns – Aggregate to the monthly level if sufficient data exist. Use quarterly averages if data are not sufficient or monthly patterns are too noisy.
- ❑ Compare what you have learned from the different temporal aggregates. Do conclusions make sense in the larger temporal picture?

  For example, the diurnal pattern of formaldehyde suggests that concentrations are highly dependant on sunlight. This dependency is confirmed by the seasonal pattern, which shows higher concentrations in summer (i.e., more sunlight).

Summary

Checklist for Ways to Characterize Air Toxics

Spatial Characterization

- ❑ General spatial patterns
  - Create site level average values by pollutant for the time period of interest. Make sure data is temporally comparable at all sites.
  - Investigate spatial variability by calculating and graphing summary statistics of the site averages. The results provide overview information about the magnitude of spatial variation.
  - Visualize spatial variability by creating maps of the site-level average concentrations.
    - Results will provide more specific information about the spatial gradients of air toxics.
    - Including supplementary data such as MDLs, remote background concentrations, and health benchmarks provides a framework for the observed concentrations.
- ❑ Within- and between-city variation
  - Calculate valid annual averages for each site within a city that has more than one monitor.
  - Create notched box plots of annual averages by city.
    - Each box will contain one point for each monitor, so the box will indicate within-city variability.
    - Including multiple cities on one plot will provide a comparison of between city variability.
- ❑ Hot and cold spot analysis
  - Calculate valid annual averages for each site.
  - Rank the averages in order of concentration.
  - Using maps, compare sites with highest and lowest concentrations to all sites.
  - Investigate data and metadata for the sites with highest and lowest concentrations. Do concentrations make sense based on the metadata and conceptual models?
- ❑ Urban vs. rural site analysis
  - Verify the EPA urban/rural designation of each site using Google Earth.
  - Identify pollutant data availability and time period.
  - Create a data set of pollutant/site combinations that are spatially and temporally representative.
  - Plot valid 24-hr average data as a notched box plots for neighboring urban and rural sites.
  - Summarize the results and investigate sites that do not meet the conceptual model of an urban or rural site.

Summary

Checklist for Characterizing Air Toxics

Risk Screening

- Create valid site average concentration data for the most recent years.
- Calculate the percent of sites above the associated risk level of interest and the percent of data below detection.
- Follow the risk screening decision tree to identify the exposure risk for each pollutant.
- More advanced risk analyses should be performed by risk assessment professionals.

A Final Note on Data Below Detection

- Most air toxics have enough data below detection to cause uncertainties and/or biases in aggregated data if not handled properly.
- Note, however, that it is not valid to remove these data because they are representative of true values on the lower end of the concentration spectrum; removal would cause even more significant positive biases.
- It is always important to know the amount of data below detection when looking at any data set. The effects of data below detection should be considered in all analyses.

Resources

- Statistical
  - StatSoft: Background on a variety of statistics http://www.statsoft.com/textbook/stathome.html
  - NIST Engineering Statistics: Background on a variety of statistics http://www.itl.nist.gov/div898/handbook/index.htm
  - SYSTAT: A graphical and statistical tool http://www.systat.com/
  - Minitab: A graphical and statistical tool http://www.minitab.com/Emissions
- Emissions
  - EPA AirData: Air toxics emissions reports to the county level http://www.epa.gov/air/data/reports.html
  - National Emissions Inventory 2002: Emissions inventory for the United States; some Canada and Mexico data also available. http://www.epa.gov/ttn/chief/net/2002inventory.html
  - EPA Toxics Release Inventory (TRI): A variety of emissions data sets http://www.epa.gov/triexplorer/chemical.htm.

Quantifying and Interpreting Trends in Air Toxics

Are air toxics concentrations changing?
Are the ambient concentration changes in response to changes in emissions?

Trends in Air Toxics

What's Covered in This Section

- This section focuses on trends in ambient air toxics over time.
- The following topics are addressed in this section:
  - Quantifying Trends
    - Overview of trends analysis
    - Setting up the data for trend analyses
    - Effect of changes in MDL on trends
    - Summarizing trends
    - Discerning and quantifying trends
      - Quantifying Trends
      - Visualizing Trends
    - Aggregating trends to larger spatial areas
  - Interpreting Trends
    - Evaluating annual trends in the context of control programs
    - *Adjusting trends for meteorology (introductory)*

Trends Overview

Motivation

- Assessing trends is useful.
- Visual inspection of trends is important.
- Understanding data uncertainties is necessary.
- Obtaining consensus (or weight of evidence) among results from different approaches increases our certainty in the observed trends.

Trends Overview

Analysis Questions

- Are concentration levels changing at a monitoring site?
- Are changes consistent across sites, areas, or regions?
- Are changes consistent across pollutants or pollutant groups?
- Are changes consistent across time periods?
- Are changes consistent with expectations (e.g., emissions controls, changes in population)?

Setting Up Data for Trend Analysis

Overview

Steps to prepare data for trend analysis

- Acquire and validate data
- Identify and treat data below detection in preparation for annual averages
- Create valid annual averages or other metrics for trends
- Create valid site-level trends

Setting Up Data for Trend Analysis

Identifying Censored Data (1 of 3)

- Data are typically reported as a concentration value with an accompanying method detection limit (MDL).
  - In AQS, the MDL is either a default value associated with the analytical method (MDL) or a value assigned by the reporting entity for that specific record (alternate MDL).
- NATTS program guidance suggests that laboratories report all values, regardless of the MDL.
  - However, many air toxics data are reported as censored values; i.e., they have been replaced with zero, MDL/2, or MDL (or some other value).
- Identifying censored values is a helpful first step in treating data below detection.

Setting Up Data for Trend Analysis
Identifying Censored Data (2 of 3)

- Data may be identified and separated at or below the detection limit along with the associated MDL and date/time.
  - Use alternate MDLs if available rather than the default MDLs.
- Data may be examined for obvious substitution. Count the number of times each value at or below detection is reported at a given site, parameter, and method. Are the majority of data reported as the same value (e.g., zero or MDL/2)?
  - If data are largely reported as two or more values, investigate the temporal variation of the data. Are there large step changes where reporting methods or MDLs have changed?
  - Do the duplicate values indicate a typical censoring method (e.g., MDL/2, MDL/10)?
  - Alternate MDLs may be different for each sample run causing a distribution of values if MDL/x substitutions were used. Just because values below MDL are not all the same does not mean they are not censored!

Setting Up Data for Trend Analysis
Identifying Censored Data (3 of 3)

- Check for MDL/X substitution.
  - Make a scatter plot of the value vs. MDL to see if the data fall on a straight line.
  - If the data do form a straight line, the slope of the regression line will indicate the value by which the MDL has been divided.
    - Is the value a reasonable number that would be used for MDL substitution (e.g., 1, 2, 5, or 10)?
    - The distribution of the ratios should be highly variable if the data are not censored.

Setting Up Data for Trend Analysis

Treating Data Below Detection (1 of 3)

- If uncensored values (i.e., NOT zero, MDL/2, or MDL) are reported below MDL, use the data "as is" with no substitution.
- If uncensored values are not available, substitute MDL/2 for data below MDL or use more sophisticated methods as described in Section 4.
- If there is a mix of censored and uncensored data,
  - Compare two substitution methods: (1) MDL/2 substitution for censored values and leave uncensored values "as is" and (2) MDL/2 substitution for all data below detection
  - If results are in the same direction using both substitution methods, confidence in the results is increased and substitution method 1 should be retained. If the results do not agree, a more sophisticated method for estimating the data below MDL should be employed.

Setting Up Data for Trend Analysis

Treating Data Below Detection (2 of 3)

- Each annual average should have an associated calculation of the percent below detection.
  - These data provide information about the biases of the annual average when data are below detection.
- When assessing trends over time for a pollutant, assess trends at all sites regardless of the percent of data below MDL.
  - Note, however, that data are below detection for many site/pollutant combinations.
  - To avoid over-interpretation of observed trends, it is recommended the trend values and their associated percent below detection be visually inspected. Consider trends at sites where at least half of the years for a given trend period have at least 15% of their measurements above MDL for that year.

Setting Up Data for Trend Analysis

Treating Data Below Detection (3 of 3)

- For the national-level analyses, a 15% "cut-off" was selected based on review of a small data set in which most data were above detection.
  - Bias in the annual average was investigated for this data set across a range of percent of data below detection. At 15% below detection, the bias in the annual average was 10-40%.
  - A more stringent cut-off may be required if less bias is desirable.
- In all cases, the percent below MDL should be considered as a possible source of bias when interpreting site level trends.

Setting Up Data for Trend Analysis

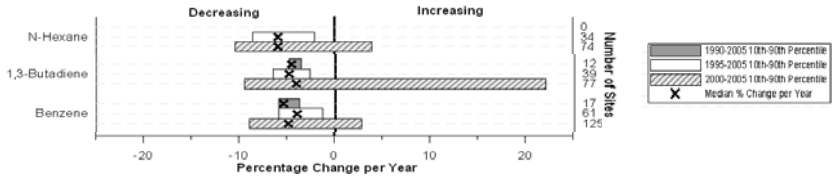
Creating Valid Trends

Trends are investigated for a unique combination of parameter, monitoring location, and method code.

- Initially, it is important to segregate method codes for a given parameter and monitoring location to assess differences (e.g., biases, detection limits) that might result in comparability issues.
- Methods may change over the course of time, perhaps causing significant analytical biases that may affect trends assessments. After investigating individual trends, by method for example, further aggregation may be reasonable.
- At a given monitoring location, sometimes more than one monitor reports the same pollutant, known as a collocated measurement. When collocated measurements are made, data from each monitor are differentiated in AQS using POCs.

Setting Up Data for Trend Analysis
Trend Length and Completeness

- Length and completeness criteria may be used to ensure that trends are representative of the time period of interest and that data are consistent for intercomparison among sites.
- When choosing these criteria, analysts should strive to strike a balance between maximizing available data and creating valid trends in the period of interest.
- More stringent constraints result in a reduction of available data. On the other hand, shorter trend periods are subject to more variability, for example, because of changes in meteorology which often obscure underlying trends.



In the example, three trend periods were investigated: 1990-2005, 1995-2005, and 2000-2005. Only 17 sites in the United States collected benzene data over the 1990-2005 sampling period that met the completeness criteria. In contrast, data from 125 sites met the completeness criteria for the shorter 2000-2005 trend period. Variability for shorter trend periods is much higher.

Setting Up Data for Trend Analysis
Trend Length and Completeness

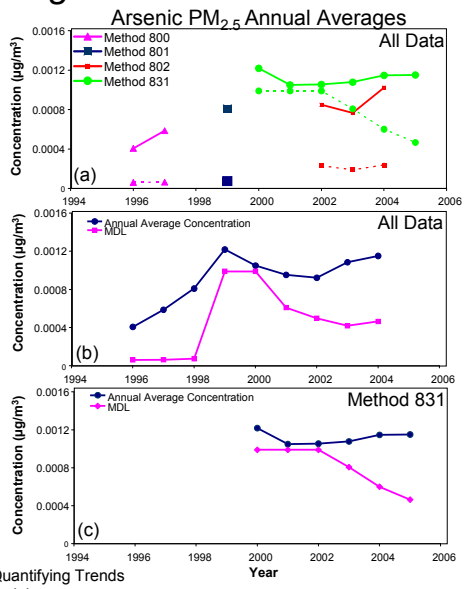
- Trend Length
  - One goal of the NATTS is to provide data with a minimum trend length of six years to be able to compare two 3-yr averages.
  - Of course, other trend periods are acceptable!
- Trend Completeness
  - Of the number of data years in a trend period, at least 75% is suggested for a site to be included (e.g., for a six-year trend period, at least five years of valid annual averages are suggested).
  - Trends with data gaps of more than two years should not be used.

Setting Up Data for Trend Analysis

Example – Creating Valid Trends

Looking at trends by method code is important.

- Figure (a) shows all annual averages for arsenic $PM_{2.5}$ at a site, color-coded by method. Solid lines indicate annual averages and dashed lines show average MDLs.
- Figure (b) shows the trend (blue) and average MDL (pink) for all data at a site regardless of method (i.e., the same data as in Figure (a) connected into one trend). This produces a statistically significantly increasing trend.
- Figure (c) shows the results if data are partitioned by method. Only data with method 831 are reserved because this method is the only one to have a trend period greater than four years. The results show a statistically insignificant decreasing trend, opposite the result obtained using all data.



Setting Up Data for Trend Analysis

Evaluating the Effect of Method Changes

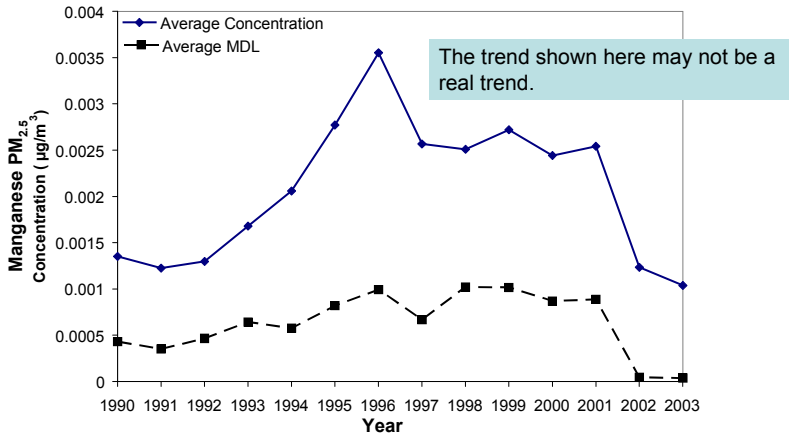
Assessing the comparability of methods will be a case-by-case analysis; no one procedure will provide the answer, but the following is a good start.

- Plot annual averages and associated average MDLs, color-coded by method for each air toxic; tabulate % of data below detection by year.
- Visually assess method changes for unusual patterns in average concentration and MDL.
- If MDL changes occur, investigate the % of data below detection to determine if MDL substitutions are driving the difference. Keep in mind the % of data below detection and effect of MDL substitutions for subsequent analyses.
- Examine trends in air toxics data that are not expected to change significantly between years (e.g., carbon tetrachloride); significant jumps in annual average concentrations for these air toxics may indicate a problem.
- Compare pollutants measured by the same methods that are expected to vary together (e.g., benzene and toluene) and look for discontinuities.
- Investigate collocated data together, if available. In some cases, a measurement method may have changed in the primary monitor, but not in the secondary monitor.

Effect of Changes in MDL on Trends Assessment

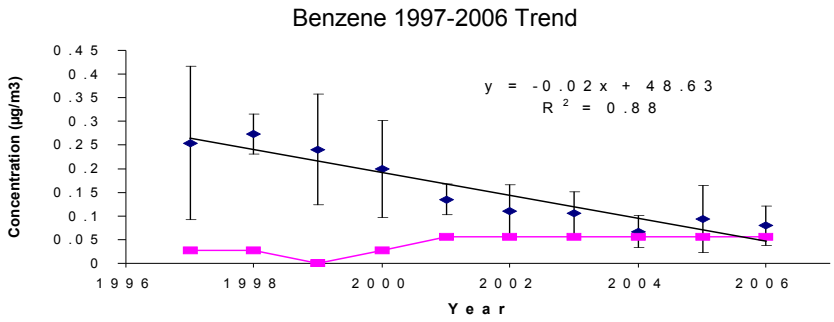
- Another important consideration in preparing data for trend analysis is that detection limits can change over time for a given monitoring site, parameter, and method. At a national scale, some detection limits change by orders of magnitude.
- These changes may influence annual averages, particularly if MDL substitutions are used. Similar trends between MDL and annual average concentrations may indicate that the changes in MDL are strongly influencing the annual average trends.
- Inspect the trends in MDL in addition to the trends in concentration, especially for air toxics with concentrations close to the MDL (i.e., within a factor of 10).
- More sophisticated statistical analysis may be needed to quantify the underlying influence of the MDL changes on the ambient concentrations.

Effect of Changes in MDL on Trends Assessment *Example* *(1 of 2)*



In the national-level investigation of manganese (Mn) trends, MDL trends were similar to concentration trends making us suspicious of the reliability of the overall ambient trend. This example shows average Mn $PM_{2.5}$ concentrations and MDLs from 1990 to 2003. For this data set, Hyslop and White (2007) showed that reported MDLs are much lower than actual detection limits. Current recommendations are to be cautious with data within a factor of 6 to 10 of the reported MDL.

Effect of Changes in MDL on Trends Assessment *Example* *(2 of 2)*



In contrast to the previous Mn $PM_{2.5}$ trend, this benzene trend does not show influence from a change in MDL (i.e., the trends in concentration and MDL show different patterns).

Quantifying Trends

Approach *(1 of 2)*

- Initial investigation of trends
  - Inspect first and last year of the trend period or two multi-year averages for change.
  - Use simple linear regression to determine the magnitude of a trend over the trend period.
- Quantifying trends
  - The percent difference between the first and last year of the trend period provides a rough sense of the change.
  - The difference between two multi-year averages provides another measure of change and helps smooth out possible influences of meteorology.
  - The percent change per year is provided by the slope of the regression line. This "normalized" value allows the analyst to compare changes across varying lengths of time (i.e., sites with different trend periods).

Quantifying Trends

Approach (2 of 2)

- Testing the significance of the observed trends
  - Calculate the significance of the slope using the F-test (see next slide). The F-test provides a statistical measure of the confidence that there is a relationship between the two variables (i.e., the regression line does not have a slope of zero which would indicate that the dependent variable is not related to the independent variable).
  - Other methods can be employed to test for significance including t-tests, nonparametric tests (tests for and estimates a trend without making distributional assumptions such as Spearman's rho test of trend; Kendall's tau test of trend), and analysis of variance.

Quantifying Trends

Interpreting Linear Regression Output

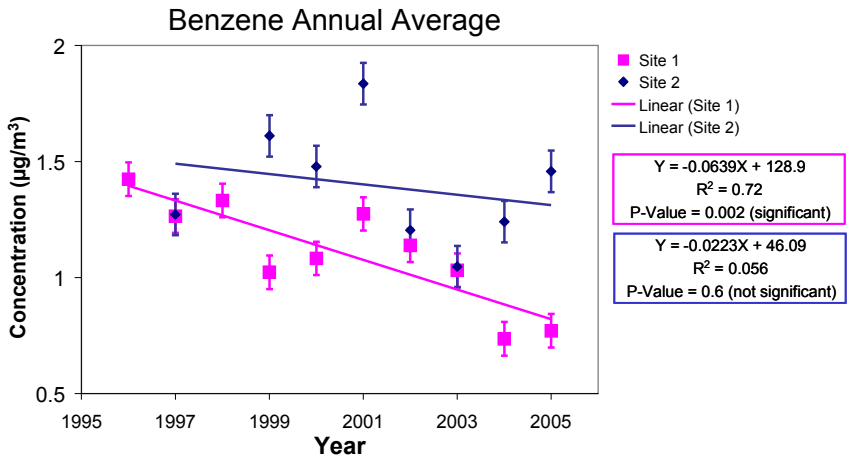
Slope	Intercept	% Change	% Change Per Year
-0.3943	789.562	-69.241021	-6.2946382
R^2			
0.794456			
F-Statistic		P-value	Confidence level
30.92103			99.946575

This example output shows a decline in annual average benzene concentrations over time with 95% confidence and slope not equal to zero.

- **Slope, intercept, % change, % change per year, R^2.** Indicate the slope of the line, y-axis intercept, % change between first and last year of the line, % change divided by number of years, and fraction of variation accounted for.
- **F-statistic or F-ratio.** Use F-ratio to test the hypothesis that the slope is 0. The F-ratio is large when independent variables help explain the variation in the dependent variable. Therefore, large F-ratios indicate a stronger correlation between the two variables (i.e., the slope of the regression line is NOT zero).
- **P-value.** Probability of exceeding the F-ratio when the group means are equal (generally, 95% confidence is used as a cutoff value, corresponding to a P-value of 0.05).

Quantifying Trends

Statistical Significance Example



This example shows benzene trends at two sites. Both sites show a linear regression with a negative slope, but only Site 1 shows a statistically significant decrease. At Site 2, a decrease in concentrations is apparent, but the change is not statistically significant (i.e., failed F-test).

Visualizing Trends

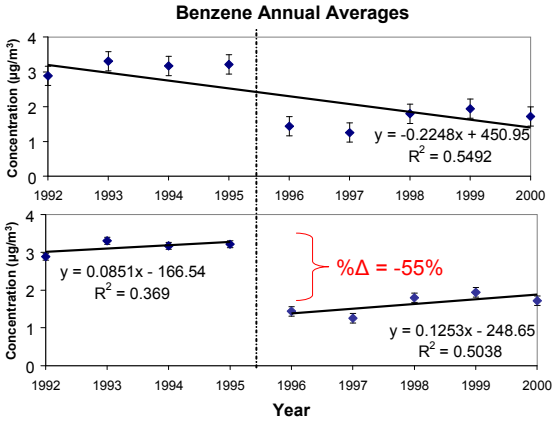
Overview

- Visual inspection of trend data is vital! A linear fit to a trend may not be appropriate; for example, a step change may have occurred due to a major emissions regulation or a nonlinear or exponential fit may be more appropriate.
- Methods for visualizing the data include
  - Line graphs of selected indicators
  - Box plots (high and low values, median values, outliers)
  - Plots of mean or median values with confidence intervals
  - Combination of a map and temporal information

Visualizing Trends

Line Graphs

- Benzene in gasoline was significantly reduced in several urban areas starting in the mid-1990s when reformulated gas (RFG) was introduced. Dramatic reductions were observed in ambient benzene concentrations over this time period.
- Both plots contain the same data. If one trend line is used, the overall trend decreases. If two trend lines are segregated by the RFG year (1995), benzene concentrations are relatively flat before and after RFG implementation.

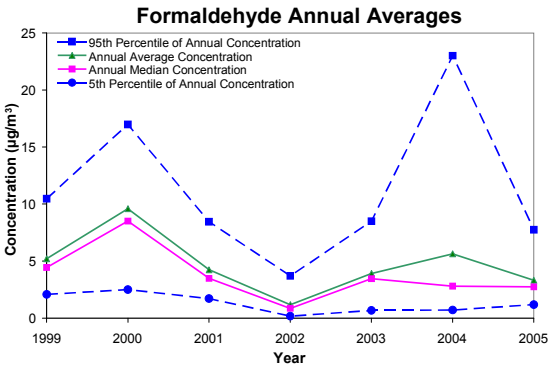


The figure shows the same benzene annual averages fitted with regression lines in two ways. The first fits all data with one regression line and the second takes into account a large step change that occurred from regulations put into effect in 1995.

Visualizing Trends

Using Other Statistical Metrics

- In addition to an annual average, other statistical indicators can be used to verify a trend.
  - These indicators include median, maximum, minimum, and selected percentiles.
  - These metrics are especially helpful in identifying effects of censored data below detection.

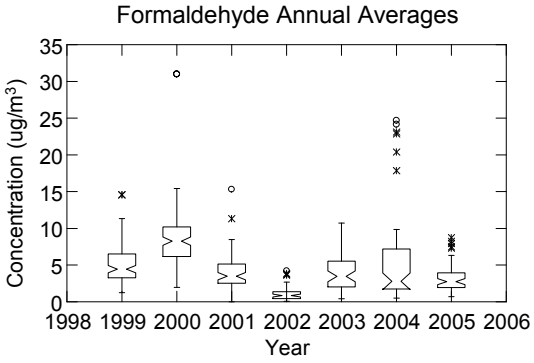


This figure, showing formaldehyde annual data with various statistical measures, demonstrates that the annual pattern in concentration is relatively consistent. 2002 concentrations were low and there is no consistent trend over this 1999-2005 time period.

Visualizing Trends

Box Plots

- Box plots are useful to display multiple statistical metrics and visually asses statistical significance.
- Box plots illustrate the trends in the high and low values, interquartile ranges, median, and confidence intervals of the annual average.

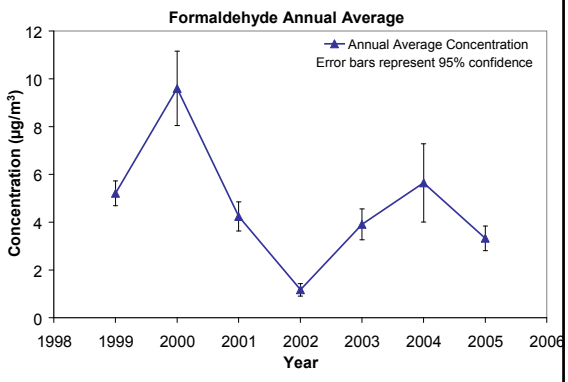


The figure shows annual formaldehyde concentrations represented as box plots. The variability is similar from year to year since the boxes for each year are about the same height. Concentrations in 2002 were statistically significantly lower than in other years because the confidence intervals do not overlap any other year.

Visualizing Trends

Using Confidence Intervals

- Since the plotted CIs overlap in 1999 and 2001 but not in 2000 and 2001, 1999 and 2001 concentrations are not significantly different, but 2000 and 2001 concentrations are significantly different.
- CIs are a function of fewer samples resulting in large CIs. Air toxics data sets are typically small (i.e., only a few samples per month); thus, CIs help analysts understand the range in which the annual mean concentration can statistically fall.
- CI is computed as follows: $\overline{x} \pm z^* \frac{\sigma}{\sqrt{n}}$



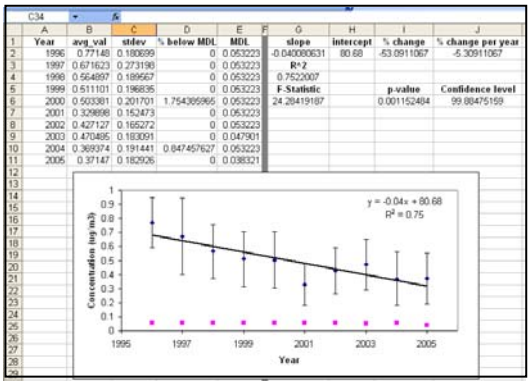
where
x is the mean value, σ is the standard deviation;
n is number of samples; and
z* **is the upper (1-*C*)/2 critical value (use a look up table for the % required) for the standard normal distribution.**

Visualizing Trends

Including Underlying Data

- A trend for each parameter, site, and method was plotted next to the underlying data. The figure shows annual averages with standard deviations in blue and average MDLs in pink. The underlying data include the average MDL, % below MDL by year and calculated regression, and F-value statistics as well as % change per year.
- Data are mostly above detection and show a statistically significant decreasing trend of about 5% per year.

(a)



Year	avg_val	stdev	% below MDL	MDL	slope	intercept	% change	% change per year
1996	0.77148	0.180699	0	0.053223	-0.040000631	80.68	-53.0911067	-5.30911067
1997	0.671623	0.273198	0	0.053223	R^2			
1998	0.564897	0.189567	0	0.053223	0.7522007			
1999	0.511101	0.196835	0	0.053223	F-Statistic		p-value	Confidence level
2000	0.503381	0.201701	1.754365966	0.053223	24.28419187		0.001152484	99.88475199
2001	0.329898	0.152473	0	0.053223				
2002	0.427127	0.165272	0	0.053223				
2003	0.470495	0.183091	0	0.047901				
2004	0.369374	0.191441	0.847457627	0.053223				
2005	0.37147	0.182926	0	0.038321				

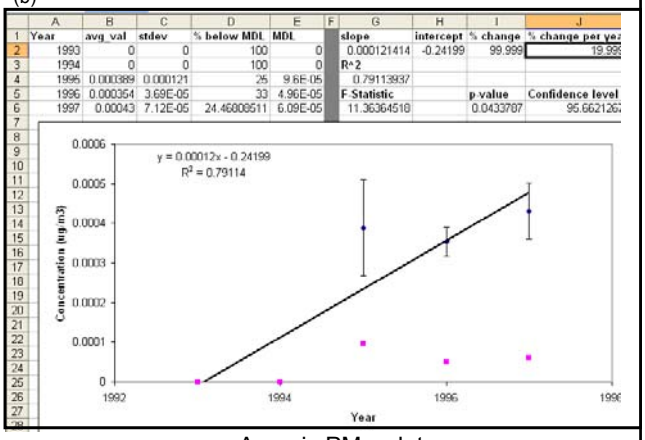
Example of a benzene trend for the 1995-2005 trend period.

Visualizing Trends

Including Underlying Data

- Calculations indicate a statistically significant increasing trend of 20% per year.
- If these statistics were used alone, they would indicate a serious arsenic problem at this site.
- The underlying data shows the first two years are 100% below detection, resulting in values that are entirely MDL/2-substituted.
- The values for these years may, in fact, be significantly lower and should not simply be discarded; we cannot tell from the current data.

(b)



Year	avg_val	stdev	% below MDL	MDL	slope	intercept	% change	% change per yea
1993	0	0	100	0	0.000121414	-0.24199	99.999	19.99
1994	0	0	100	0	R^2			
1995	0.0003889	0.000121	25	9.6E-05	0.79113937			
1996	0.000364	3.69E-05	33	4.96E-05	F-Statistic		p-value	Confidence level
1997	0.00043	7.12E-05	24.46808511	6.09E-05	11.36364518		0.0433707	95.6621267

Arsenic $PM_{2.5}$ data

This trend should be considered suspect and validated by comparison with neighboring sites.

Visualizing Trends

Calculating Trend Period Percent Change

- Four methods for calculating trend-period percentage change and the associated percentage change that would result from applying each method to the benzene data shown:
  1. Using the first and last measured data point (-40.4%).
  2. Using the regression equation (-57.1%).
  3. Using all values before and after a step change (-55.3%).
  4. Using three-year averages before and after a step change (-53.7%).
- Method 1 provides no sense of the underlying pattern for all years of interest, and the results are affected by meteorology of the chosen years.
- Method 3 isolates the two data points having the most impact on the overall trend, but requires visualizing the data first.
- Methods 2 and 4 use values that are weighted by more years of data within the trend period, providing more smoothing of variability from meteorological fluctuations.

Summarizing Trends

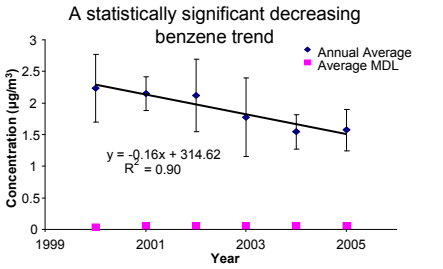
Overview

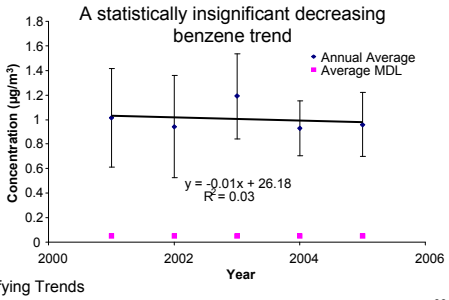
- Investigate trends among sites by pollutant.
  - Similar trends results among the sites makes a compelling argument that change on a larger spatial scale has occurred.
- Characterize the spatial distribution of trends by showing trends at each site on a map.
  - Trends may not agree nationally in direction or magnitude but may show spatial patterns of interest.
- Characterize the distribution of individual site trends by displaying the range of percentage change per year over various trend periods and for all sites meeting minimum trend criteria.

Summarizing Trends

Trends Among Sites

- Site-level trends for benzene from two U.S. sites; average MDLs are plotted in pink for reference.
- The top figure shows a statistically significant decreasing trend, while the bottom figure shows a statistically insignificant decreasing trend.
- Confidence in these results is high. The data are mostly above detection, MDLs are consistent for the whole trend period, and no outliers appear to influence the trend.
- If any of these problems do exist, the underlying trend data should be evaluated more carefully to understand the reliability of the trend.





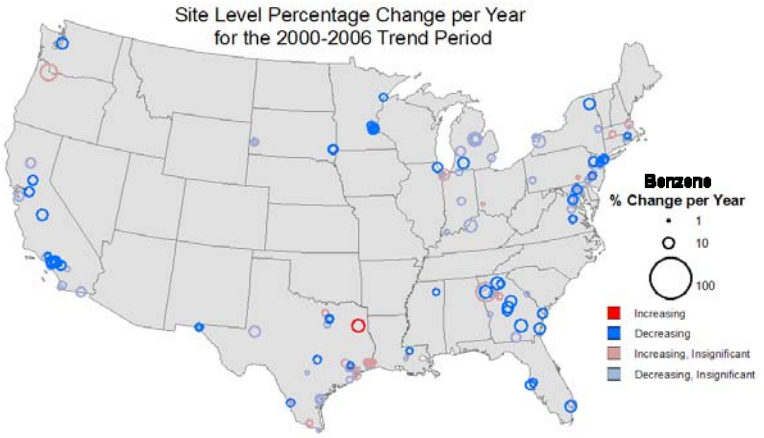
Summarizing Trends

Trends Among Sites

- Next steps in investigating suspect trends
  - If one or more annual averages are outliers, revalidate the underlying data.
    - Is one high concentration event the cause, or is there a distribution of high values? Is there an explanation for the high annual average to prove it valid (e.g., increased local source emissions) or in error (e.g., unit conversion error)?
  - If MDL changes occur and
    - A low percentage of data is below detection, the change in MDL should not have a noticeable effect.
    - A high percentage of data are below detection, there is decreased confidence in the trend. If MDL substitution is used check that the trend does not follow the shape as the MDL changes; if it does the trend is likely unreliable.
  - If a high percentage of data is below detection without an MDL change, the central tendency of the data may still be accessible, but there is lower confidence in the trend.

Summarizing Trends

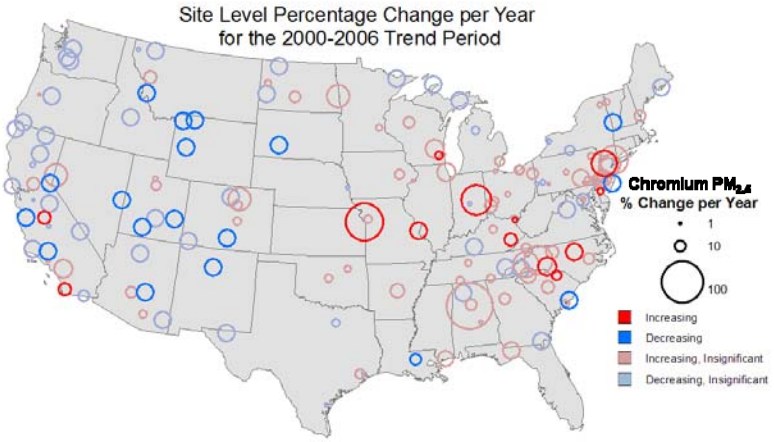
Example – Spatial Distribution (1 of 2)



Benzene site-level percent change per year for 2000-2006. Many sites in the United States show a statistically significant decline in benzene concentrations over the period.

Summarizing Trends

Example – Spatial Distribution (2 of 2)

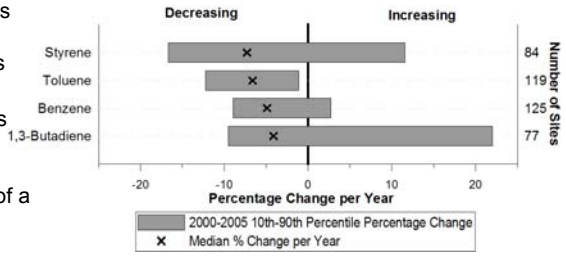


Chromium $PM_{2.5}$ concentrations across the United States in 2000 to 2006. The statistically significant trends are spatially distinct, indicating increasing concentrations in the eastern half of the country and decreasing concentrations in the West.

Summarizing Trends

Example – Percentage Change per Year

- The bar chart summarizes trends in percent change per year for selected mobile source air toxics for 2000-2005 data.
- A range of results is seen across the network (i.e., 10th to 90th percentile sites); however, most sites are experiencing declines of a few percent per year with remarkable consistency (see median); "outlier" (e.g., 95th percentile) sites may be candidates for additional investigation.
- 1,3-butadiene and styrene show a wider assortment of percentage changes by site. The median U.S. monitoring site, however, shows a trend of about -5%, in agreement with the other mobile source air toxics.



- Benzene and toluene show similar ranges in percent change per year and less variability in trends across the United States than 1,3-butadiene and styrene.
- Toluene is decreasing at 90% of sites by about 2% to 12% per year, while benzene is decreasing at most sites and may be increasing at some sites.

Summarizing Trends

Example – Percentage Change per Year

- Site-specific percent change values for benzene used in the bar chart, similar to the proportional maps shown previously.
- Comparing data summaries, such as the bar chart, to more detailed plots, such as the map, offers an overview of the data. The map shows the spatial distribution of data included in the summary statistics. Benzene is increasing in some areas of the United States, but none of the trends are statistically significant. Many of the decreasing trends, on the other hand, are statistically significant.

Aggregating Trends to Larger Spatial Regions (1 of 2)

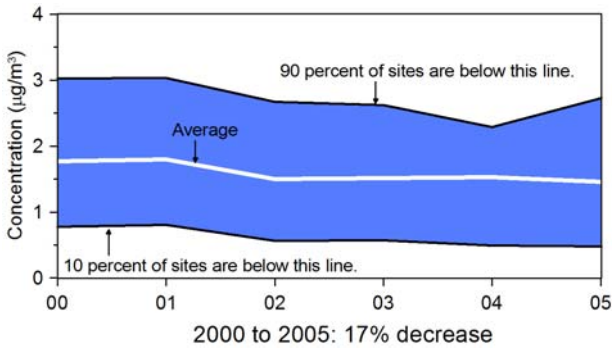
- Aggregated trends for larger spatial regions, such as trends by state or EPA Region, may be of interest to communicate results at a "big picture" level to interested stakeholders.
- As data sets become smaller—i.e., the analyst looks at fewer sites and fewer years—gaps in the data record become more important.
  - For example, some site-level trend periods may meet the minimum criteria but will still have gaps in the data.
  - Problems arise when, in combining data sets, a site, especially one measuring high or low concentrations, has missing data during some time periods.

Aggregating Trends to Larger Spatial Regions (2 of 2)

- To handle these data gaps.
  - For general site-level analyses, leave gaps as is.
  - While not done at a national level, when aggregating to larger spatial regions, data gaps could be filled in, using the following methods, to be consistent with current trends analyses performed for criteria pollutants:
    - Missing the last year – set the missing year equal to the second-to-last year.
    - Missing the first year – set the missing year equal to the second year.
    - Missing any other year – interpolate between the adjacent two years.
    - No more than two years in succession can be missing *(this was applied in the national analyses*).

Aggregating Trends

Example – Using Line Graphs



- Line graphs can be used to assess trends in selected indicators.
- National benzene trends (annual average concentrations) from 2000-2005 are summarized in the graph.

Line graph figures were created with Grapher7.

Accountability

Overview (1 of 2)

- Changes in air quality may be due to a number of factors. Trends in air quality can provide evidence that local, regional, or federal emissions controls have successfully reduced ambient concentrations of pollutants harmful to human health.
- Analysis should include as much information on interpretation of trends as possible including evaluation of other potential sources of the compound in question as well as regulations, and meteorological influences that may impact emissions.
- The evaluation of the impacts of regional control programs (those that affect multiple states) and local control programs (those that affect an urban area) on air quality is complicated and is stepwise and site- and pollutant-specific.
- A major challenge in this type of analysis is the scale of influence of a control and of the impact of that control on air quality. Previous investigations of ambient air quality changes encountered the confounding influences of multiple controls applied within similar time frames and at different spatial scales.

Accountability

Overview *(2 of 2)*

- Use caution – Matching trends to changes in emissions is not sufficient to prove that an emission change actually caused the ambient change.
- Emissions regulations are typically phased in over a period of years, causing a gradual change in ambient concentrations; other factors such as meteorology, local source profiles, and MDL changes may also explain changes. The use of supplementary data (e.g., investigating trends in a pollutant not expected to be influenced by the emission change) is necessary to be sure observed changes are truly emissions-related.
- Two approaches to a trends accountability analysis can be taken depending on the availability of information: an emission control approach (bottom up) and an ambient data approach (top down).

Accountability

Bottom-Up Approach *(1 of 2)*

- Select a control measure.
- Identify the air toxics expected to be affected and the available data, other controls that might have affected the pollutants, and other pollutants that may have been affected.
- Consider the spatial scale, or zone of influence (ZOI), of the control measure. Was the control applied at a single facility (monitor-specific or fence line), at an urban scale (MSA-wide), national scale (e.g., 49-state automobile emission rules), or global scale (e.g., Montreal protocol)?
- Determine the timing and magnitude of the changes. Was the control phased in over a period of time, applied to specific emitters? Phasing in a control makes it more difficult to discern the relationship between the ambient concentration change and the control change.

Accountability

Bottom-Up Approach *(2 of 2)*

- Consider the magnitude of the expected air quality changes relative to the variability in the ambient data. If the inherent variability in the ambient data is very large, a small change in emissions may not be observable.
- Select the appropriate statistical metrics or approach for the analysis. Data treatments may help reduce the variability in the data so that trends can be observed.
- Develop hypotheses of expected changes, identify supporting evidence of changes, and investigate corroborative evidence of the changes. It is often helpful to test for changes in data sets or pollutants in which changes were not expected (i.e., check the null hypothesis).

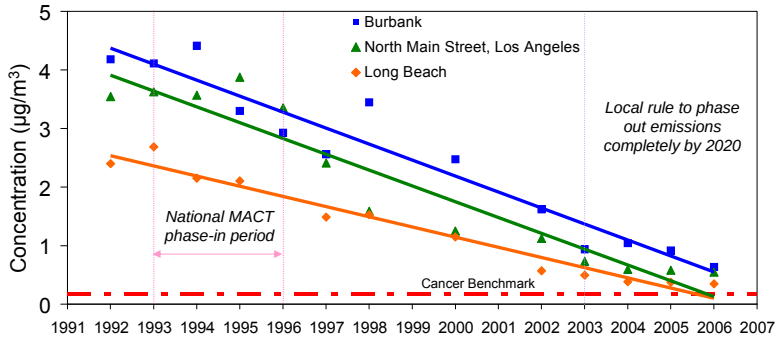
Accountability

Top-Down Approach

- Quantify the change observed in the ambient data. This approach could also be applied to a pollutant for which a change was not observed but expected.
- Identify and assess other data sets and sites that may have also been affected by a similar control measure or emission change to understand the spatial scale of the ambient change. If the control was applied across a broad area, changes at additional sites might be expected.
- Identify potential emissions changes or control measures that could have contributed to the ambient trends. Local knowledge is often a key component of this part of the analysis.
- Compare the control measure implementation schedule with the ambient trends. Do the timing of the control implementation and the change in ambient concentrations coincide?
- Investigate corroborative evidence of the change and test for changes in pollutants for which a change was not expected. It is important not to over-interpret changes in ambient data.

Bottom-Up Example

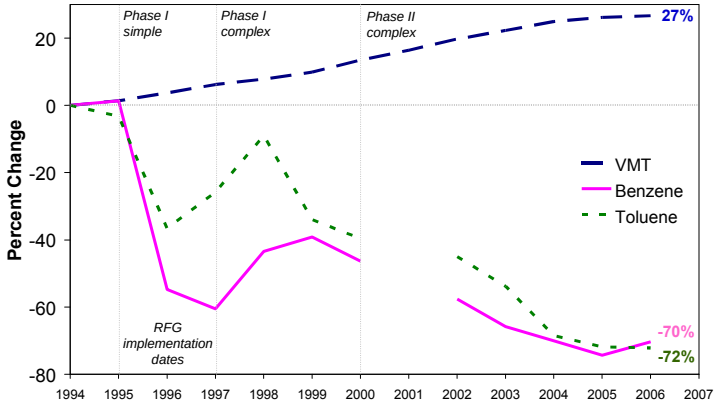
Tetrachloroethene Controls in Los Angeles



- Tetrachloroethene is the chemical most widely used by the dry cleaning industry, with over 85% of facilities using it as the primary cleaning agent. In 1993, the EPA promulgated technology-based emissions standards to control tetrachloroethene emissions from dry cleaners.
- The MACT standards implemented in 1993 resulted in drastic reductions in tetrachloroethene concentrations in the Los Angeles area where monitoring data have been available from three sites since 1992.

Bottom-Up Example

Ozone Precursor Controls in Baltimore, MD



- Air toxics, such as benzene and toluene, that are emitted by motor vehicles are significant contributors to ozone formation. Reformulated gasoline (RFG) was introduced in the United States in phases to reduce motor vehicle emissions of benzene and other ozone precursors in order to reduce ambient ozone concentrations.
- Benzene and toluene concentrations decreased after the 1995 implementation of RFG despite an increase in the number of vehicle miles traveled by cars and trucks in the Baltimore area.

National Level Top-Down Example
Method

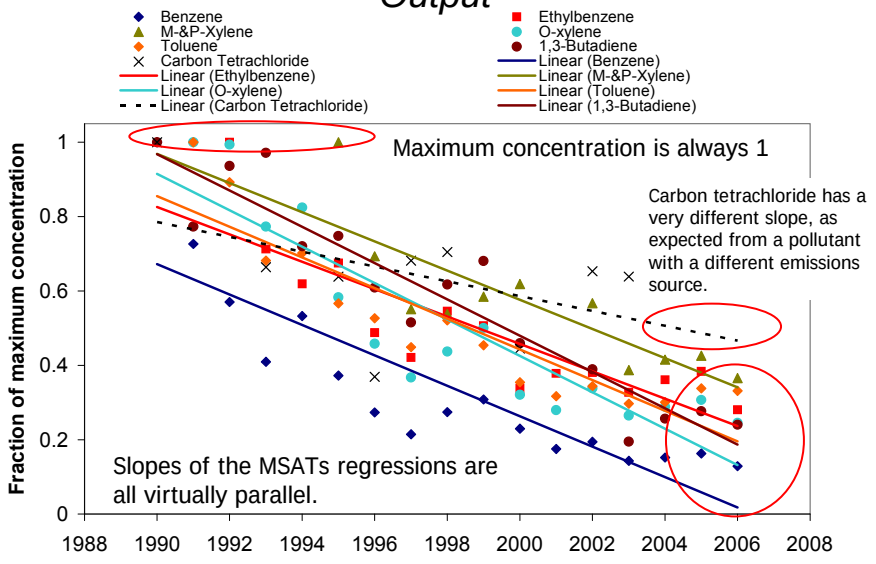
- The hypothesis is that if pollutants are emitted by the same source, emissions should covary over long time scales (i.e., trends should be parallel if normalized).
- Nationally, the goal was to identify covariant trends in MSATs as an indicator of sites dominated by mobile source emissions.
- Site-specific trends for six MSATs (benzene, 1,3-butadiene, toluene, ethylbenzene, o-xylene, m-&p-xylenes) were investigated using carbon tetrachloride as a control.
- Trends were normalized by the maximum annual average concentration within the trend period by site and pollutant (i.e., annual average concentrations each year were divided by the highest annual average in the time period for each pollutant and at each site). Normalization creates a data set that is easier to compare across sites and pollutants and shows the relative change in concentration.

National Level Top-Down Example
Method

- Linear regression was used to create trend lines for each pollutant.
- The sites were visually grouped into various categories by the behavior of pollutant trends.
  - If all MSAT trends had a similar slope, we expect the change in concentration at that site to be a consequence of mobile source reductions.
  - If one MSAT exhibited a slope very different from the others, we would conclude that another source of that pollutant impacting the site was likely.
- For this analysis, only the site and parameter were required to be consistent over the trend period (method and POC were allowed to float between years). Sites with more than five annual averages were included.
- Sites were then investigated using Google Earth to see if our hypotheses were correct.

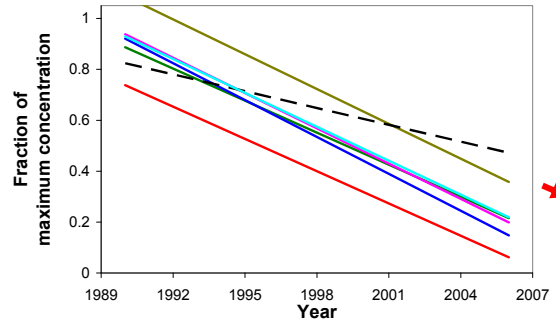
National Level Top-down Example:

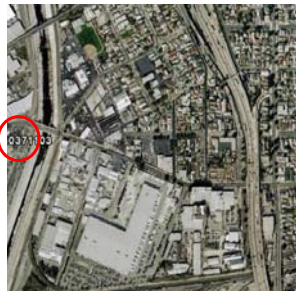
Output



National Level Top-Down Example

Normalized Site-Specific Regression Lines





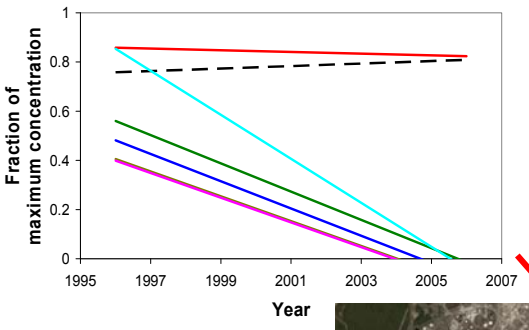
Benzene
Ethylbenzene
M&P-Xylene
O-Xylene
Toluene
Carbon Tetrachloride
1,3-Butadiene

All MSATs show a similar declining slope. This site is primarily mobile source-dominated (it is located very near a major freeway).

National Level Top-Down Example

Normalized Site-Specific Regression Lines

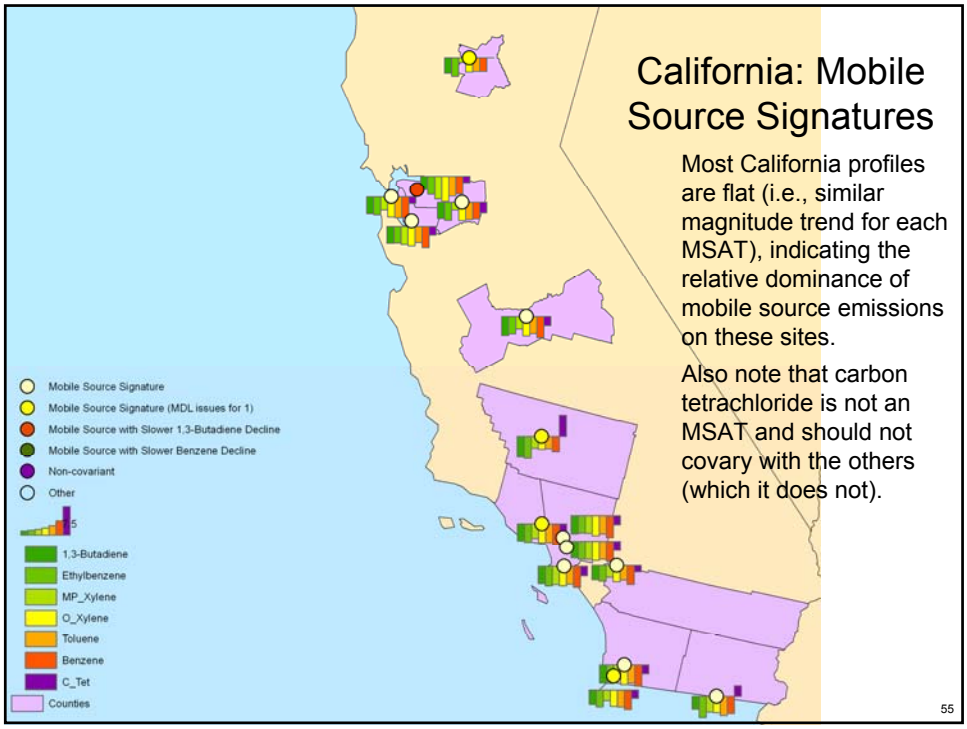
- Similar slopes are seen for all MSATs except benzene and 1,3-butadiene.
- Benzene shows a much slower decline in concentration than the other MSATs while 1,3-butadiene shows a slightly faster decline.
- This monitor is located near a large refinery with both benzene and 1,3-butadiene emissions which may explain this divergent behavior.



National Level Top-Down Example

Spatial Characterization of Trend Profile "Signatures"

- Visual inspection of the slopes of trends provides useful information on the covariance of pollutant concentrations over time.
- The percentage change in concentrations per year can also be plotted on maps for each pollutant shown in the scatter plots to spatially investigate the trends profiles.
- Mobile source signatures have MSAT profiles of similar magnitudes; other signatures have increasing or varying magnitudes among the pollutants.



California: Mobile Source Signatures

Most California profiles are flat (i.e., similar magnitude trend for each MSAT), indicating the relative dominance of mobile source emissions on these sites.

Also note that carbon tetrachloride is not an MSAT and should not covary with the others (which it does not).

55

National Level Top-Down Example

Summary

- The top-down approach is a useful way to investigate site-level trends of pollutants commonly emitted by the same source.
- Most sites in the United States conformed to our expected mobile source trend profile signature.
- The technique also allows identification of sites at which trends do not conform to expectations.
- Some sites showed increasing trends or noncovariant trends in multiple MSATs. Nearby emissions sources may be influencing trends at these sites, and they may be good candidates for case study analyses of other emissions sources.
- The top-down approach may be applicable to other pollutants from mobile sources (CO, NO_x, black carbon) or other emissions sources of multiple co-emitted pollutants.

Meteorological Adjustment of Air Toxics
Introductory Thoughts

- Meteorology can impact air quality.
  - Meteorology can vary significantly among years (e.g., El Niño), and meteorology can have a considerable effect on air quality.
  - To understand changes in air quality that are attributed to emission controls, we need to be able to adjust the data to account for meteorological conditions that were very different from average conditions.
  - By properly accounting for the portion of the variability in the data attributable to changes in meteorology, we can compare air quality among years with widely different meteorological conditions.
  - This assessment is important because we do not have control over meteorological changes.
- Using meteorological adjustment of air toxics is still being explored.
- Application of meteorological adjustment is likely at site-level, and each site and pollutant will need to be treated discretely.

Resources
Tools Available for Trend Analysis

- Examples in this section were created with
  - ArcInfo and ArcView <http://www.esri.com/>
  - SYSTAT
  - Grapher
  - Microsoft Excel
- Air toxics guidance
  - http://www.epa.gov/ttn/fera/risk_atra_main.html
- Computing 95% upper confidence limit (95% UCL) for use in risk assessment
  - ProUCl 4.0 available at http://www.epa.gov/nerlesd1/tsc/software.htm

Trends Summary (1 of 2)

- Setting up data for trends analysis
  - Acquire and validate data. See *Preparing Data for Analysis*, Section 4, for a complete discussion.
  - Identify censored data. Separate data at or below detection for each parameter, site and method.
    - Count the number of occurrences by value. Do the values indicate a specific substitution method?
    - Make scatter plots of data below detection vs. the detection limit for each value. The slope of the line will indicate the denominator if MDL/x substitutions were used, even if alternate MDLs are available.
  - Treat data below detection.
    - If uncensored values are used, include them "as is".
    - If censored values are used, substitute MDL/2.
    - If a mixture of censored and uncensored data is used, compare the methods of all substituted vs. only censored substituted to see if results agree. If not, more advanced methods to treat data below detection may be necessary.
  - Calculate valid annual averages. See *Preparing Data for Analysis*, Section 4, for a complete discussion.
  - Create valid trends.
    - Segregate trends by parameter, site and method.
    - Consider and apply trend completeness criteria depending on data needs.
      - Minimum trend length of 6 years
      - 75% yearly completeness within trend period
      - Data gaps longer than 2 years not allowed
    - Consider yearly aggregated percent of data below detection.
      - Look at all data regardless of percent below detection
      - Remove trends where more than half the year's data are less than 15% of data above detection

Trends Summary (2 of 2)

- Quantifying Trends
  - Magnitude of change
    - Use simple linear regression to calculate first and last year values to determine the percent change over the trend period.
    - Calculate percent change per year for intercomparison of trend periods.
  - Significance of change
    - Quantify the statistical significance of the slope using the F-test.
    - Typically, a trend is considered significant at or above the 95% confidence level.
  - Visualize trends; always include annual percent below detection as a measure of uncertainty.
    - Line graphs
    - Box plots
    - Spatial representations
  - Summarize trends
    - Characterize the distribution of percentage change per year for all sites and investigate mean, median and percentiles.
    - Characterize the spatial distribution of the percentage change per year.
    - Look for consensus in results among methods.
- Accountability – tie annual trends to control programs
  - Acquire background information on control programs; compare this information to site-level metadata keeping in mind local sources, site location etc.
    - Implementation date or time period
    - Pollutants affected and expected magnitude of reduction
    - Types of sources affected
  - Acquire emissions inventory data
    - Toxics release inventory data (TRI) (does not include mobile source emissions!)
    - National emissions inventory data (NEI)
  - Compare ambient data to emission inventories and control programs—correlation is not enough to prove causation
    - Compare similar pollutants that should experience concentration reductions resulting from the control programs.
    - Compare similar pollutants that should NOT experience concentration reductions for the control program.

Additional Reading

Meteorological Adjustment Techniques *(1 of 2)*

Methods for adjusting pollutant concentrations to account for meteorology

- Expected peak-day concentration (California Air Resources Board, 1993)
- Native variability (California Air Resources Board, 1993)
- Filtering techniques (e.g., Rao and Zurbenko, 1994)
- Probability distribution technique (Cox and Chu, 1998)
- Classification and Regression Tree (CART) analysis (e.g., Stoeckenius, 1990)
- Linear regression (e.g., Davidson, 1993)
- Nonlinear regression (e.g., Bloomfield et al., 1996)

Additional Reading

Meteorological Adjustment Techniques *(2 of 2)*

- PAMS ozone adjustment techniques, http://www.epa.gov/air/oaqps/pams/analysis/trends/txtsac.html#meteorological
- Thompson M.L., Reynolds J., Lawrence H.C., Guttorp P., and Sampson P.D. (2001) A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmos. Environ.* **35**, 617-630. Available on the Internet at www.nrcse.washington.edu/pdf/trs26_ozone.pdf
- Data Quality Objectives for the Trends Component of the PM Speciation Network (includes meteorological adjustment techniques in Appendix), http://earth1.epa.gov/ttn/amtic/files/ambient/pm25/spec/dqo3.pdf

Advanced Analyses

What else can I do with my air toxics data?

Advanced Analyses

What's Covered in This Section?

- Overview of selected advanced data analysis techniques that may be useful in further understanding air toxics data.
- Not all of these analyses have yet been thoroughly applied to air toxics data, but approaches that have been applied to $PM_{2.5}$ and PAMS VOC data, for example, should be applicable to air toxics data sets.
- The following topics are covered
  - Source apportionment
  - Trajectory analysis
  - Emission inventory evaluation
  - Model evaluation
  - Monitoring network assessment

Advanced Analyses

Motivation

After basic data validation and "display and describe" analyses have been performed, more can be done with the data if sufficient resources (time, expertise) are available and more sophisticated analyses are needed because basic analyses did not sufficiently answer questions.

- **Source Apportionment**. Understanding the sources impacting your monitors can be explored with source apportionment techniques and tools.
- **Trajectory Analyses**. In addition to better understanding high and low concentrations, source apportionment results can be enhanced with trajectory analyses.
- **Evaluation of Emissions Inventories and Models**. A primary goal of national monitoring networks is to compare ambient data to emission inventories and model output. These evaluations can lead to improvements in the inventories and model performance.
- **Network Assessment**. The pollution sources impacting a site, nearby demographics, and monitoring purpose can change over time. EPA's air toxics monitoring plan includes regular network assessment.

Source Apportionment

Why Perform? (1 of 2)

- Also known as receptor modeling, source apportionment is defined as a specified mathematical procedure for identifying and quantifying the sources of ambient air pollutants at a monitoring site (the receptor) primarily on the basis of concentration measurements at that site.
- Source apportionment relates source emissions to their quantitative impact on ambient air pollution.
- Receptor models can be used to address the following questions:
  - What emissions sources contribute to ambient air toxics concentrations?
  - How much does each source type contribute?
  - Which sources could be targeted with control measures to effect the highest reduction of air toxics concentrations (or risk)?
  - What are the discrepancies between emission inventories and sources identified by receptor models?
  - Are known control strategies affecting the source contributions to air toxics?

Source Apportionment

Why Perform? (2 of 2)

- Many emitters have similar species composition profiles.
  - The practical implication of this limitation is that one may not be able to discern the difference between benzene emitted from light-duty vehicles (LDV) versus benzene from gasoline stations or refineries. One solution to this problem is to add additional species to reduce collinearity. These profiles might help to qualitatively identify mobile sources.
- Species composition profiles change between source and receptor.
  - Most source-receptor models cannot currently account for changes due to photochemistry. Since carbonyl compounds such as formaldehyde and acetaldehyde have significant secondary sources, current methods cannot link these compounds to their primary emission sources.
- Receptor models cannot predict the consequences of emissions reductions.
  - However, source-receptor models can check if control plans achieve their desired reductions using historical data.

Source Apportionment

Single-Sample Models

In *single-sample* models, the analysis is performed independently on each available pollutant.

- The simplest example is the “tracer element” method, in which a particular property (e.g., chemical species) is known to be uniquely associated with a single source. The impact of the source on the ambient sample is estimated by dividing the measured ambient concentration of the property by the property's known abundance in the source's emissions. This method is not often available because of the difficulties of finding unique tracers or knowing their abundances. However, even if a pollutant is not uniquely associated with a source of interest, knowledge of the abundance from that source can be used to provide an upper limit for the source's impact.
- The best-known example of single-sample receptor modeling is the chemical mass balance model (CMB). CMB eliminates the need for unique tracers of sources but still requires the abundances of the chemical components of each source (source profiles) input.

Source Apportionment

Multivariate Models

Multivariate receptor models use data from multiple pollutants and extract source apportionment results from all the sample data simultaneously.

- The reward for the extra complexity of these models is that the models attempt to estimate not only the source contributions (i.e., mass from each source) but also the source compositions (i.e., profiles).
- Several tools described in the literature are available to perform multivariate receptor modeling. EPA supported the development of two modeling platforms: Unmix and positive matrix factorization (PMF). These models are based on factor analysis, or the closely related principal component analysis (PCA).
- There is extensive literature available describing CMB and PMF applications to speciated PM data, less available literature describing applications to VOC data, and very little research on air toxics specifically.

Source Apportionment

Positive Matrix Factorization

- PMF was originally developed by Paatero (1994, 1997) with additional development by Hopke et al. (1991, 2003). PMF can be used to determine source profiles based on the ambient data and associated uncertainties.
- PMF has been applied to many data sets to determine sources of $PM_{2.5}$, ozone precursors, and air toxics.
- PMF uses weighted least squares fits for data that are normally distributed and maximum likelihood estimates for data that are log normally distributed. Concentrations are weighted by their analytical uncertainties.
- PMF constrains factor loadings and factor scores to nonnegative values and thereby minimizes the ambiguity caused by rotating factors.
- Model input includes ambient monitoring data and associated analytical uncertainties (see Wade et al., 2007).
- Model output includes
  - Factor loadings expressed in mass units which allows them to be used directly as source signatures.
  - Uncertainties in factor loadings and factor scores which makes the loadings and scores easier to use in quantitative procedures such as chemical mass balance.

Source Apportionment

Unmix

- Unmix was developed by Ron Henry (1997) using a generalization of the self-modeling curve resolution method developed in the chemometric community.
- The EPA, along with Ron Henry, developed EPA Unmix and documentation that uses MATLAB features but is a standalone model (i.e., MATLAB not needed).
- Unmix is a multivariate receptor modeling package that inputs ambient monitoring data and seeks to find the composition and contributions of influencing sources or source types. UNMIX also produces estimates of the uncertainties in the source compositions.
- Unmix requires many samples to extract potential sources, similar to PMF.
- It assumes that sources have unique species ratios, i.e., "edges" that can be observed in a scatter plot between species; uses these edges to constrain the results and identify factors; and does not need to weigh data points.
- Model input includes ambient monitoring data; uncertainty information and source profiles are not necessary.
- Model output includes source profiles with uncertainties.

Source Apportionment

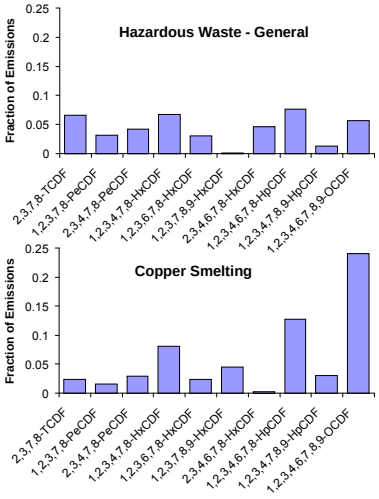
Chemical Mass Balance

- The premise of chemical mass balance (CMB) is that source profiles from various classes of sources are different enough that their contributions can be identified by measuring concentrations of many species collected at the receptor site.
- To apportion sources, CMB uses an effective variance-weighted, least squares solution to a set of linear equations which expresses each receptor species concentration as a linear sum of the products of the source profiles and source contributions.
- Model input includes
  - Source profile species (fractional amount of species in emissions from each source type).
  - Receptor (ambient) concentrations.
  - Realistic uncertainties for source and receptor values. Input uncertainty is used to weigh the relative importance of input data to model solutions and to estimate uncertainty of the source contributions.
- Model output includes contributions from each source type and species to the total ambient concentration along with uncertainty.
- CMB has been used in a number of air pollution studies that examine particulate and VOC source apportionment, but few, if any, specific air toxics studies.

Source Apportionment

Source Profiles *(1 of 2)*

- Source profiles provide information about the relative contribution of pollutants to emissions from a given source.
- Understanding source profiles is important because receptor modeling tools typically output source profile information that needs to be interpreted or requires user-input source profiles as a starting point for analysis.
- Polychlorinated dibenzofuran (PBDF) source profiles for hazardous waste incinerators and copper smelting compiled by the EPA show that though the same compounds are present, the relative abundances are not the same, providing a mechanism for source identification.



Source Apportionment

Source Profiles *(2 of 2)*

- For CMB applications and for interpretation of PMF output, it is important to use source profiles that are representative of the study area during the period when ambient data were collected.
- In CMB, try available source profiles in sensitivity tests to determine the best ones for use (i.e., minimize collinearity).
- Source profiles can be obtained from
  - EPA SPECIATE, recently updated (version 4.0) and available at http://www.epa.gov/ttn/chief/software/speciate/index.html.
  - Literature review, source measurements made in your region during the period for which ambient data are available.
  - Local, state, and federal agencies.
  - Source profiles can also be procured via analysis of ambient data using tools such as PMF and UNMIX.

Source Apportionment

Approach (1 of 2)

Before beginning source apportionment, it is important to "know the data" in order to identify and assess the receptor model outputs. Understanding the data will be achieved in the process of data validation and analysis.

- Understand airshed geography and topography using maps, photographs, site visits, etc.
- Investigate the composition and location of emission sources.
- Understand the typical meteorology of the site, including diurnal and seasonal variations.
- Investigate the spatial and temporal characteristics of the data, including meteorological dependence.
- Investigate the relationships among species using scatter plot matrices, correlation matrices, and other statistical tools.

Source Apportionment

Approach (2 of 2)

- Apply cluster and factor analysis techniques using standard statistical packages to get an overall understanding of pollutant relationships and groupings by season, time of day, etc.
- If there are sufficient samples (e.g., more than two years of 1-in-6 day samples for more than 20 species and more than 50% of data above detection), Unmix and/or PMF may be applied to obtain "source" profiles with more species and further investigate data relationships.
- If samples are few and source profiles are available, CMB may be applied to obtain source contribution estimates.
- Compare source contributions estimates and source profiles from Unmix and PMF to the emission inventory.

Source Apportionment
Example

- PMF receptor modeling was performed for speciated VOC data collected at two PAMS sites, Hawthorne and Azusa, in the Los Angeles area during the summers of 2001-2003.
- Both toxic and non-toxic VOCs were investigated in order to provide as much data as possible for apportionment (Brown et al., 2007a).
- Air toxics included in the analysis were typically grouped as MSATs, though they have industrial sources as well.
- Data were collected as part of the PAMS network providing the advantage of subdaily data and speciated-versus-total mass measurements (total non-methane organic compounds, TNMOC).
- Uncertainty estimates were enhanced from the original analytical uncertainties by reducing the weighting of data below detection and missing data. Uncertainties for missing data were estimated with four times the median concentration, data below detection were given uncertainties of 1.5*MDL, and all other data were given the analytical uncertainty plus 2/3*MDL.

Source Apportionment
Example Preliminary Analyses

- Preliminary data analyses were performed including investigation into data quality, local emissions, species relationships, temporal patterns, etc.
- Findings
  - VOC concentrations were typically higher at Azusa compared to Hawthorne, a result consistent with site locations relative to the ocean.
  - The Azusa air mass was more aged, as indicated by loss of reactive species (except during rush hour); this is also consistent with the sites' locations in the air basin.
  - The Hawthorne site seemed to have constant, fresh emissions, with little change in the relative abundance of VOCs throughout the day, consistent with nearby industrial emissions.
  - Both sites are significantly influenced by mobile sources.

Source Apportionment

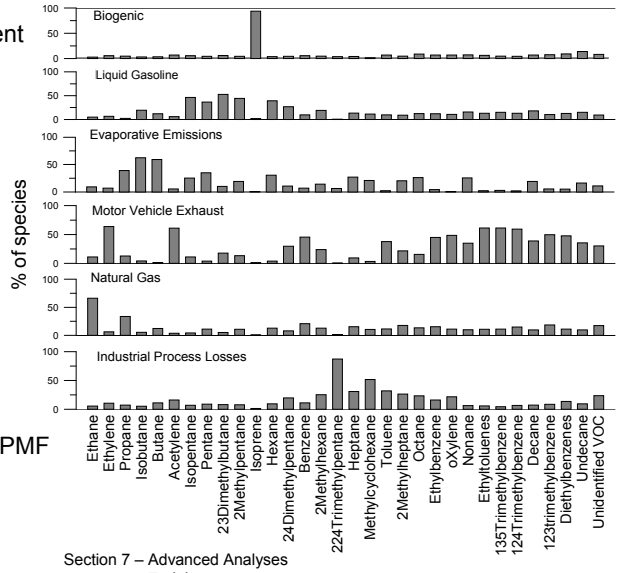
Example Hawthorne Site PMF Profiles (1 of 2)

- Six factors were identified by PMF at the Hawthorne site following protocols discussed in the Multivariate Workbook (Brown and Hafner, 2005).
- Profile names indicate analyst-identified source types. Some rationale for source identification:
  - Biogenic. Isoprene is the only marker for biogenic sources measured in this data set and anthropogenic sources of isoprene are insignificant; temporal patterns match expectations.
  - Liquid Gasoline. Abundance of C5 alkanes agrees with previous work; temporal patterns are consistent with mobile sources.
  - Evaporative Emissions. C3-C6 alkanes and temporal patterns are similar to diurnal temperature patterns.
  - Motor Vehicle Exhaust. Typical exhaust profile and temporal patterns are consistent with rush-hour traffic.
  - Natural Gas. Natural gas is mostly ethane and propane. These are also long-lived species that accumulate in the atmosphere.
  - Industrial Process. Losses. Consistent with nearby industrial emissions.

Source Apportionment

Example Hawthorne Site PMF Profiles (2 of 2)

- The relative percent of species mass attributed to each profile is shown.



Source profiles from PMF

Source Apportionment

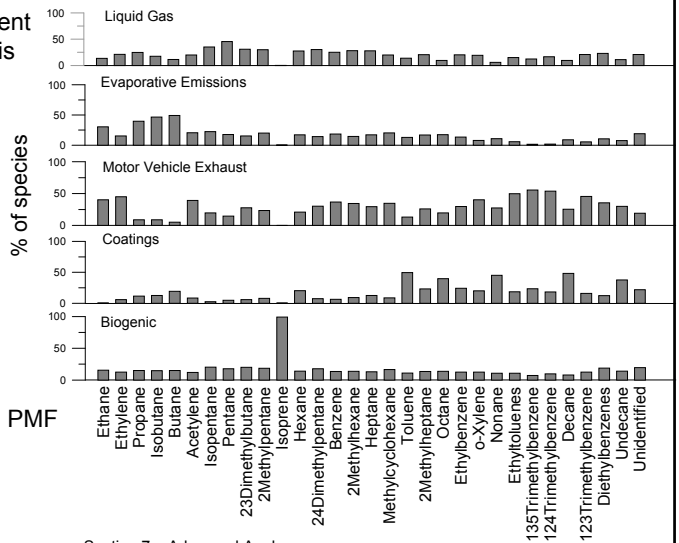
Example Azusa Site PMF Profiles (1 of 2)

- Five factors were identified by PMF at the Azusa site.
- Apportionment of these profiles to specific sources was performed by the analyst based on knowledge of source profiles and other investigations into the data.
- Some of the rationale for source identification
  - Coatings. Presence of C9-C11 alkanes is consistent with previous results; temporal pattern showed a daytime peak consistent with industrial operations.
  - Other profiles are similar to those observed at the Hawthorne site.

Source Apportionment

Example Azusa Site PMF Profiles (2 of 2)

- The relative percent of species mass is shown.

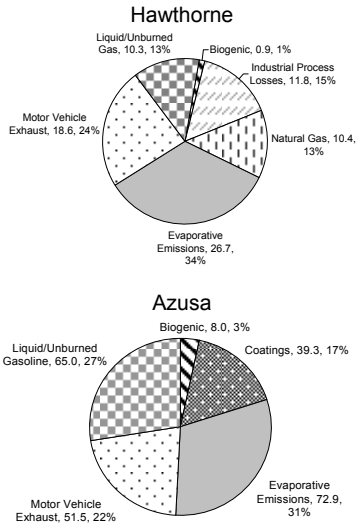


Source profiles from PMF

Source Apportionment

Example Percent of Total Mass

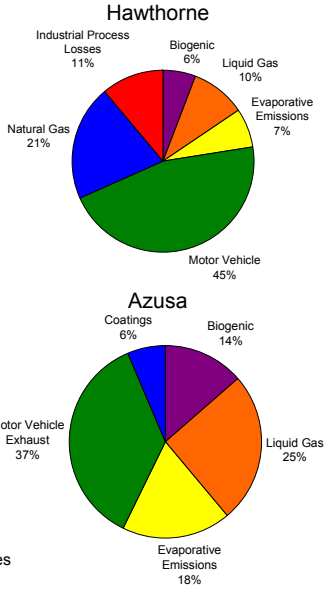
- The pie charts show the importance of each source profile by quantifying the amount of TNMOC mass represented by each profile. For example, in Hawthorne, evaporative emissions accounted for 34% of TNMOC mass during the summers of 2001-2003.
- Mobile source emissions are dominant contributors to TNMOC at both Hawthorne and Azusa with 71% and 80% of total mass, respectively (sum of liquid/unburned gasoline, motor vehicle exhaust, and evaporative emissions).
- The remaining VOC mass is attributed to coatings at the Azusa site and is split between industrial processes and natural gas at the Hawthorne site.



Source Apportionment

Example Apportionment of Benzene

- Percentage of benzene (by mass) attributed to each source profile identified by PMF at the Hawthorne and Azusa sites.
- As expected, both sites show a significant percentage of benzene mass attributed to mobile sources and gasoline evaporation. Interestingly, almost one-fourth of the benzene at the Hawthorne site is attributed to natural gas. While benzene is not emitted in natural gas, a significant fraction of ambient benzene is associated with air parcels containing ethane and propane. Since benzene is relatively long-lived, it is possible that benzene in this profile represents urban background. The same observation can be made for the benzene in the biogenic profile.
- A reduction in benzene emissions might be sought through addressing mobile sources. This type of reduction would likely reduce the urban background concentrations as well.



Source Apportionment

Summary of Source Apportionment Steps

- Review data quality and spatial/temporal characteristics.
- Prepare data for source apportionment.
  - Processing the necessary data differs among the tools, but typically the analyst needs to select pollutants with sufficient data above detection and understand/quantify uncertainty for each concentration. Guidance is provided in the EPA's Multivariate Receptor Modeling workbook (Brown et al., 2007b).
  - Uncertainty estimates for air toxics data have been developed as part of Phase V analyses and continuing work with EPA.
- Understand the air shed by assessing likely emissions sources and local meteorology. This helps set expectations for what the source apportionment results should show.
- With guidance from literature and workbooks, apply source apportionment tools. This is an iterative process!
- Evaluate results for reasonableness.
- Compare results to emission inventories.

Trajectory Analysis

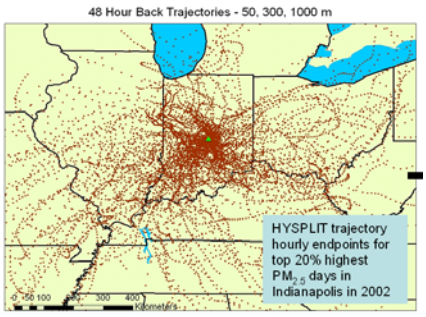
Introduction

- Trajectory analysis uses knowledge of air mass movement to trace the most likely areas of influence on high pollutant concentrations.
- The use of trajectory analysis after source apportionment helps analysts better understand, interpret, and verify source apportionment results.
- Analysis techniques
  - Backward trajectories
  - Trajectory densities
  - Potential Source Contribution Function (PSCF)
  - Conditional Probability Function (CPF)

Trajectory Analysis

Backward Trajectories

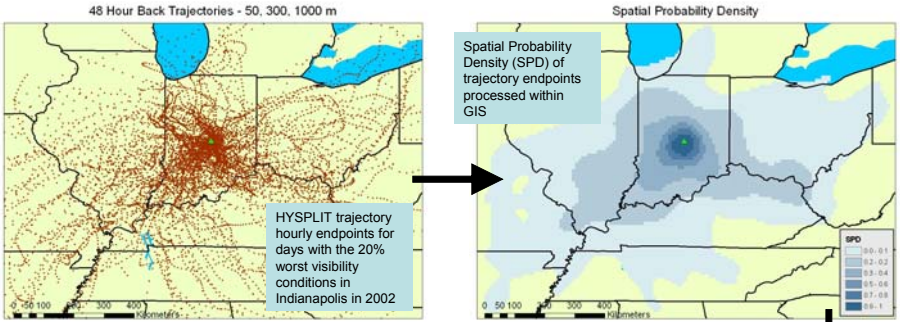
- Backward air-mass trajectories estimate where air parcels were during previous hours.
- Air-mass trajectories can be employed to investigate long-term, synoptic-scale meteorological conditions associated with high concentrations of individual factors.
- Estimates grow less certain as time elapses.



Trajectories are often plotted as single points for every hour backwards from the start point as shown here (also called a spaghetti plot). However, they should not be viewed as specific points, but rather as a small area around that point and with the last and next point.

Trajectory Analysis

Trajectory Densities



Trajectories are often processed into density, rather than “spaghetti”, plots. Higher density corresponds to more trajectories passing through that grid square. This plotting enables a number of useful analysis techniques, such as Potential Source Contribution Function (PSCF) analysis.

Trajectory Analysis

Potential Source Contribution Function (PSCF)

PSCF uses HYSPLIT backward trajectories to determine probable locations of emission sources.

$$PSCF = \frac{m_{ij}}{n_{ij}}$$

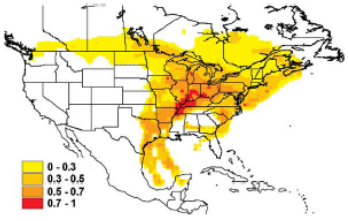
n_{ij} = number of times trajectory passed through cell (i,j).

m_{ij} = number of times source contribution peaked while trajectory passed through cell (i,j).

Top 10%-20% source contributions are used for m_{ij}.

In the example, all five-day backward trajectories, for every two hours were applied to the corresponding 24-hr source contributions.

PSCF calculated for each cell sized 1°×1° and results displayed in the form of maps on which PSCF values ranging from 0 to 1 are displayed in a color scale.



PSCF function plot for sulfate affecting Philadelphia. Higher probability is associated with an area of high SO_2 emissions.

(Source: Begum et al., 2005)

Trajectory Analysis

Conditional Probability Function (CPF)

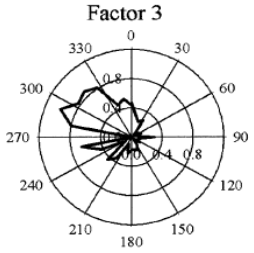
CPF uses wind direction, rather than trajectories, to determine the likely direction of sources. CPF compares days when concentrations were highest to the average transport pattern (i.e., the climatology).

$$CPF = \frac{m_{\Delta\Theta}}{n_{\Delta\Theta}}$$

$n_{\Delta\Theta}$= number of times wind direction is from sector $\Delta\Theta$.

$m_{\Delta\Theta}$= number of times source contributions are high while wind direction was from sector $\Delta\Theta$.

A CPF value close to 1.0 for a given sector ($\Delta\Theta$) indicates a high probability that a source is located in that direction.



Example CPF plot for the highest 25% contribution from a PMF factor pointing to the northwest of site as a possible source region.

(Source: Kim et al., 2004)

Trajectory Analysis

Interpretation

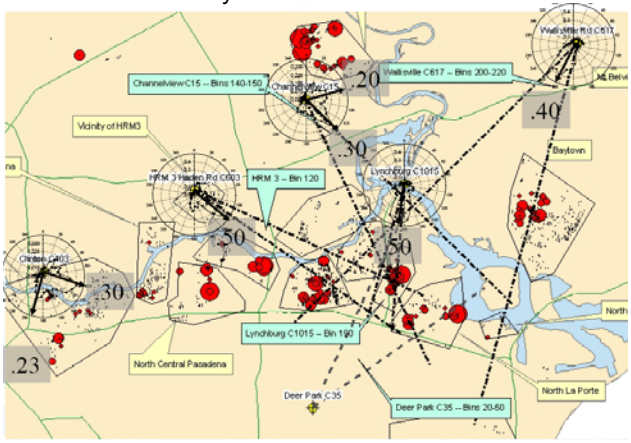
- No matter which trajectory analysis is used, interpretation of results is similar. These methods are all complementary to source apportionment or can be standalone to assess source regions. No one method shown is superior.
- The following questions may be investigated for verification of results:
  - Do results meet the conceptual model of emissions and removal of air toxics?
  - Are these the areas from which emissions influence would be expected?
  - Does the transport pattern make sense with respect to the age/chemistry of a given factor (i.e., more transport and chemistry are associated with secondary pollutants such as formaldehyde)?

Trajectory Analysis

Using CPF Results

This approach is based on the assumption that wind direction and trajectory analysis results should be consistent with the spatial distribution of the sources in the emission inventory.

The directions of source regions from the CPF plots agree with the locations of propene sources in the area (red circles), giving more confidence to the source apportionment results.



(Source: Berkowitz et al., 2004)

Emission Inventory Evaluation

Why Bother Evaluating Emissions Data?

- Emission inventory development is an intricate process that involves estimating and compiling emissions activity data from hundreds of point, area, and mobile sources in a given region.
- Because of the complexities involved in developing emission inventories and the implications of errors in the inventory on air quality model performance and control strategy assessment, it is important to evaluate the accuracy and representativeness of any inventory that is intended for use in modeling.
- Furthermore, existing emission factor and activity data for sources of air toxics and their precursors are limited and the quality of the data is questionable.
- An emission inventory evaluation should be performed before the data are used in modeling.

Emission Inventory Evaluation

What Tools are Available for Assessing Emissions Data?

- Several techniques are used to evaluate emissions data including "common sense" review of the data; source-receptor methods such as PMF; bottom-up evaluations that begin with emissions activity data and estimate the corresponding emissions; and top-down evaluations that compare emission estimates to ambient air quality data. Each evaluation method has strengths and limitations.
- Based on the results of an emissions evaluation, recommendations can be made to improve an emission inventory, if warranted. Local agencies responsible for developing an inventory can then make revisions to the inventory data prior to modeling.
- $PM_{2.5}$ and PAMS data analysis workbooks provide some example analyses and approaches that are applicable to air toxics data (Main and Roberts, 2000; 2001).

Emission Inventory Evaluation

Using Ambient Data

- Ambient air quality data can be used to evaluate emission estimates ("top-down"); however, the following issues should be considered:
  - Proper spatial and temporal matching of emission estimates and ambient data is needed.
  - Ambient background levels of air toxics need to be considered.
  - Meteorological effects need to be considered.
  - Comparisons are only valid for primarily emitted air toxics.
  - To compare ambient concentrations to emissions estimates, a pollutant or total value (such as total VOC) is needed to create a ratio. Typically, NO_x or CO is used.

Emission Inventory Evaluation

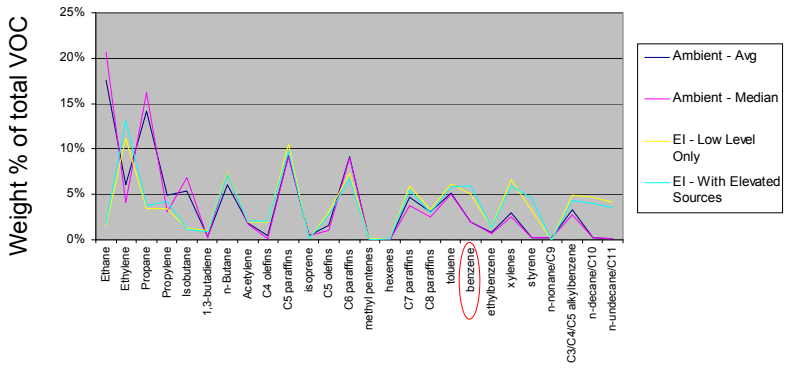
Top-down Approach

- **Top-down emissions evaluation** is a method of comparing emissions estimates with ambient air quality data. Ambient/emission inventory comparisons are useful for examining the relative composition of emission inventories; they are not useful for verifying absolute pollutant masses unless they are combined with bottom-up evaluations. The top-down method has demonstrated success at reconciling emission estimates of VOC and NO_x.
- **Top-down approach**: Compare ambient- and emissions-derived primary air toxic/NO_x, CO, or VOC ratios.

If early morning samples are available (such as with PAMS data), these sampling periods are the most appropriate to use because emissions are generally high, mixing depths are low, winds are light, and photochemical reactions are minimized.

Emission Inventory Evaluation

Example



- At this PAMS site, the EI-derived compositions of benzene are significantly higher than the ambient-derived compositions. Examination of point source records near the source indicates that the sources of these emissions are chemical manufacturing operations. It appears that the chemical speciation profiles used to speciate the point source inventory over-represent the relative amount of benzene (by about a factor of 2 to 5). Similarly, xylenes are overestimated.
- Toluene and 1,3-butadiene are only slightly overestimated in the EI at this site.

Evaluating Models

Introduction

- Air quality models have been used for decades to assess the potential impact of emission sources on ambient concentrations of criteria and toxic air pollutants.
- In the past decade, air quality models have also been used as planning tools for criteria pollutants, e.g., SIP development and attainment demonstration.
- However, until recently, air quality models have not been used as planning tools for air toxics, due to the lack of measurements with which to evaluate the models.
- The need to assess the usefulness of these models in air quality planning and to improve both modeling and evaluation methods has been identified – How well are we modeling air toxics?
- Reasonable agreement between model and monitor concentrations was set by EPA as “within a factor of 2”.
- Example of model-to-monitor comparisons for NATA and methodology for comparisons are provided at:
  http://www.epa.gov/ttn/atw/natamain/index.html
  http://www.epa.gov/ttn/atw/nata1999/99compare.html

Evaluating Models

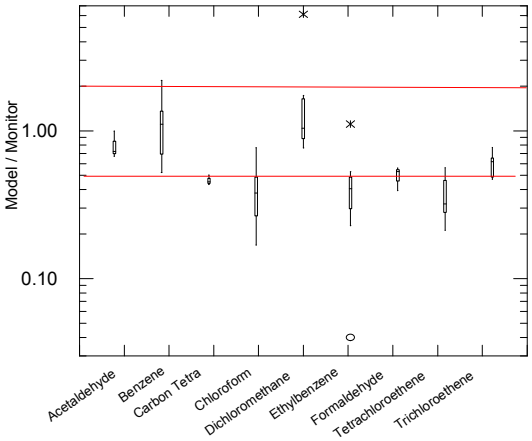
Methodology

- **Modeled Data**. Modeled data of interest for air toxics include publicly available and widely used NATA data. For this example, NATA99 model results were used.
- **Monitored Data**. In order to reduce perturbations from meteorology and other data biases in monitored data, the site average of 1998-2000 valid annual averages was used for comparison to model output.
- The lowest spatial resolution of NATA99 data is census tract level, so NATA99 modeled results should be related to ambient monitoring data at this level. If multiple sites fall into one census tract the sites should still be individually evaluated.
- **Analyses**. If data from many sites are available, box plots of modeled/monitored data can be examined; fewer sites lend themselves to a scatter plot approach of model-to-monitor data. Model-to-monitor ratios within a factor of 2 are considered to be within the acceptable limits of a good comparison; see http://www.epa.gov/ttn/atw/natamain/index.html.

Evaluating Models

Using Box Plots

- Red lines indicate the cutoff for modeled-to-monitored concentrations within a factor of 2.
- Acetaldehyde, benzene, dichloromethane, and trichloroethene typically agreed within a factor of 2, consistent with national level comparisons of model and monitor data.
- However, ethylbenzene, formaldehyde, carbon tetrachloride, chloroform and tetrachloroethylene showed monitored concentrations more than a factor of 2 higher than model estimates.

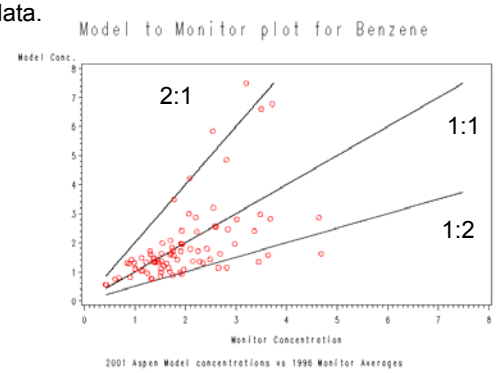


Ratio of NATA99 modeled data to monitored data at several urban sites

Evaluating Models

Using Scatter Plots (1 of 2)

- Modeled and monitored concentrations can also be compared using scatter plots, plotting each data pair (ambient site-average, model output) separately. For NATA 1999, benzene data compared well to the modeled data.



Model-to-monitor scatter plot for benzene. Most points fall within the factor of 2 wedge, and none are far outside the wedge.

Evaluating Models

Using Scatter Plots (2 of 2)

- There are several reasons to expect good agreement between model prediction and monitor results for benzene.
  - It is a widely distributed pollutant emitted from point, area, and mobile sources. Thus, if the model is biased in the way it handles any one of these source categories, the bias will likely be dampened by one of the other sources.
  - An estimated background concentration was available for benzene in the modeling effort.
  - There is a large number (87) of monitoring sites for benzene for this comparison, resulting in an adequate sample size for the statistics in the comparison.
  - Monitoring technology for benzene has a long history, suggesting that the monitoring data reflects actual ambient concentrations.
  - Benzene emissions have been tracked for many years, so there is some confidence in emission estimates.

Network Assessment

Introduction

- Air quality agencies may choose to re-evaluate and reconfigure monitoring networks because
  - Air quality has changed;
  - Populations and behaviors have changed;
  - New air quality objectives have been established (e.g., air toxics reductions, $PM_{2.5}$, regional haze); and
  - Understanding of air quality issues and monitoring capabilities have improved.
- Network assessments may include
  - Re-evaluation of the objectives and budget for air monitoring;
  - Evaluation of a network's effectiveness and efficiency relative to its objectives and costs; and
  - Development of recommendations for network reconfigurations and improvements.
- Network assessment guidance is available from EPA at http://www.epa.gov/ttn/amtic/cpreldoc.html.

Network Assessment

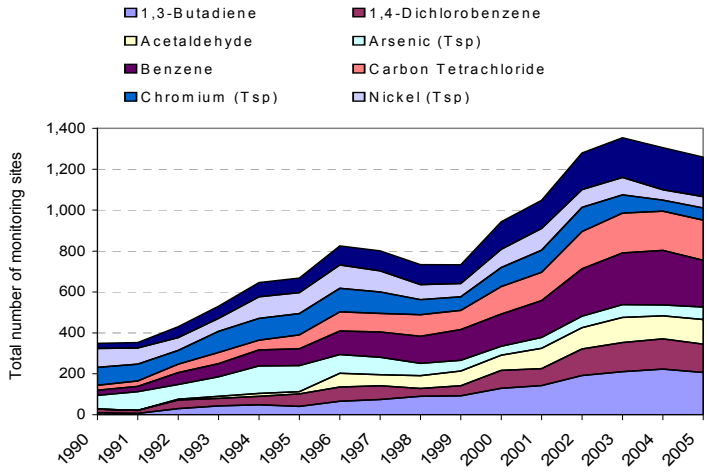
Methodology

Some things to consider when performing a network assessment:

- Length of monitoring. Takes into account a site's monitoring history because long data records can be highly useful in trends and accountability analyses.
- Suitability analyses. Combines many data sets such as population or population change, meteorology, topography, and emissions to asses suitability of current or future monitoring locations.

Network Assessment

Period of Operation (1 of 2)

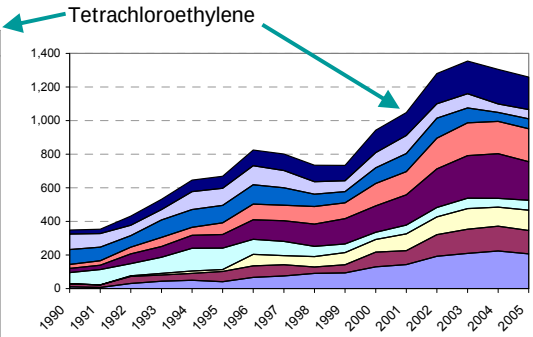


The figure shows the number of monitoring sites per year for a variety of air toxics. The number of air toxics monitoring sites has increased dramatically since 1990.

Network Assessment

Period of Operation (2 of 2)

City, State	AQS SiteID	Years
Stockton, CA	06-077-1002	13
Baltimore, MD	24-510-0040	12
Los Angeles, CA	06-037-1002	11
San Francisco, CA	06-001-1001	10
Fresno, CA	06-019-0008	10
Baltimore, MD	24-005-3001	10
Los Angeles, CA	06-037-1103	9
Los Angeles, CA	06-037-4002	9
San Diego, CA	06-073-0003	9
San Francisco, CA	06-075-0005	9
San Jose, CA	06-085-0004	9
Baltimore, MD	24-510-0006	9
Sacramento, CA	06-061-0006	8
San Diego, CA	06-073-0001	8
Oxnard, CA	06-111-2002	8
Chicago, IL-IN-WI	18-089-2008	8
Baltimore, MD	24-510-0035	8



The table lists the number of annual averages available for tetrachloroethylene at toxics monitoring sites from 1990 to 2003. For this analysis, sites with the longest record would be rated higher than those with shorter records.

Network Assessment

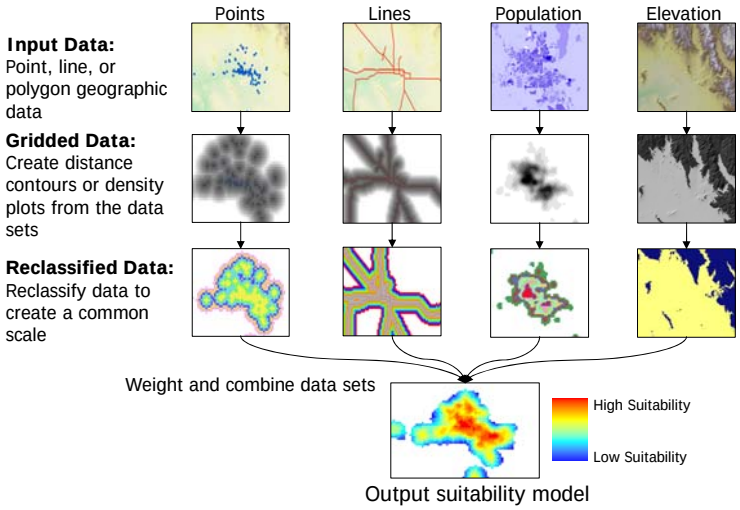
Suitability Modeling/Spatial Analysis (1 of 2)

- Motivation
  - This method may be used to identify suitable monitoring locations based on user-selected criteria.
  - Geographic map layers representing important criteria, such as emissions source influence, proximity to populated places, urban or rural land use, and site accessibility, can be compiled and merged to develop a composite map representing the combination of important criteria for a defined area.
  - The results indicate the best locations to site monitors based on the input criteria and may be used to guide new monitor siting or to understand how changes may impact the current monitoring network.
- Resources needed
  - GIS, site locations, population and other demographic/socioeconomic data, emission inventory data
  - Meteorology and concentration data may be helpful, but are not necessary
  - Skilled GIS analyst

Network Assessment

Suitability Modeling/Spatial Analysis (2 of 2)

A representation of the process of suitability modeling and spatial analysis



Network Assessment

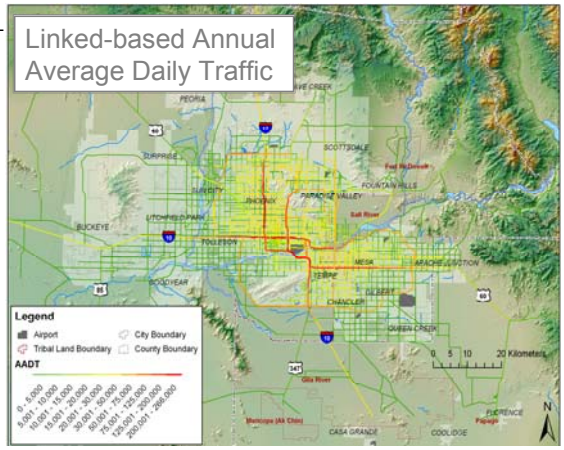
Suitability Modeling Example

- The goal of this analysis was to use GIS technology to identify locations within an area potentially suitable for placing air toxics and/or particulate monitors to better assess diesel particulate matter (DPM) emissions impacts on population.
- The emission inventory was assessed to determine
  - predominant sources of DPM; and
  - the best available geographic data to represent the spatial pattern of the identified emission sources in the region.
- The relative importance of each geographic data set was determined based on its potential DPM contribution.
- The input layers were weighted accordingly and combined to produce a suitability map using the Spatial Analyst GIS tool.

Network Assessment

Example Suitability Modeling Data Layers

1. Traffic volume (Annual Average Daily Traffic, AADT)
2. Heavy-duty truck volume (from AADT data)
3. Locations of railroads and transportation depots
4. Residential and commercial development areas
5. Golf courses and cemetery locations (lawn and garden equipment usage)
6. Airport locations
7. $PM_{2.5}$ point source locations (weight assigned to each source depends on the source's relative EC contribution)
8. Total population and sensitive population (e.g., under 5 and over 65 years of age) density
9. Annual average gridded wind fields representing predominant wind direction throughout the region



Network Assessment

Example Suitability Modeling Weighting

Weighting Scheme – two model scenarios were used:

1. Proximity to diesel emission sources (hot spot)
2. Proximity of population to diesel sources

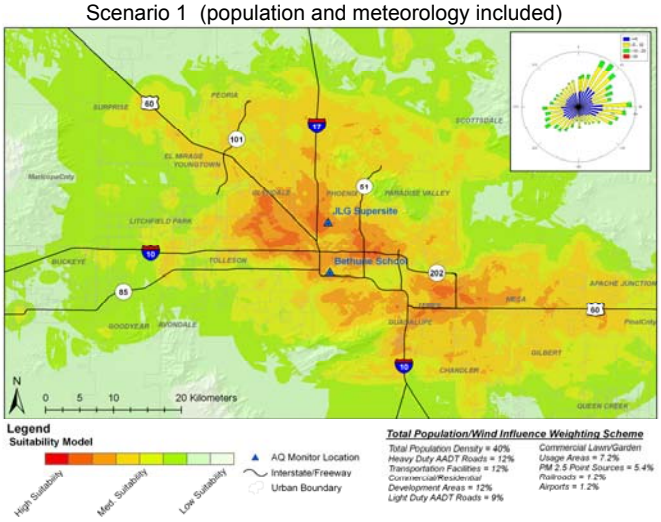
Layer	(1) Hot Spot	(2) Total Population	Weighting Criteria
Density of total population	–	40%	High population density = more suitable
Heavy-duty vehicle activity	20%	12%	High traffic density = more suitable
Light-duty vehicle activity	15%	9%	High traffic density = more suitable
Transportation distribution facility	20%	12%	Close to facility = more suitable
Lawn/garden activity areas	12%	7.2%	High activity density = more suitable
Commercial/residential construction activity areas	20%	12%	High activity density = more suitable
Distance to airports	2%	1.2%	Close to airport = more suitable
Distance to railroads	2%	1.2%	Close to railroad = more suitable
$PM_{2.5}$ point source activity	9%	5.4%	High non-EC $PM_{2.5}$ emissions density = less suitable

Network Assessment

Example Results of Suitability Modeling

- The map shows the results of combining all data layers in Scenario 1 (table on previous slide).
- The map indicates that the Glendale area is a hot spot for both diesel influence and population, as well as the area around the Phoenix Supersite.
- The area between Guadalupe and Mesa is also suitable for monitoring to better understand DPM impacts.

Scenario 1 (population and meteorology included)



Network Assessment

Suitability Analysis Summary

- Results of this analysis assisted decision makers in
  - Assessing the utility of current monitors;
  - Selecting locations for new monitors;
  - Setting monitoring priorities; and
  - Investigating a range of monitoring objectives and considerations.
- Suitability analysis can improve the effectiveness of monitoring decisions

Resources

- PMF, Unmix, and CMB:
  http://www.epa.gov/scram001/receptorindex.htm
- EPA's Multivariate Receptor Modeling Workbook:
  http://www.sonomatechdata.com/sti_workbooks/#MVRMWB
- NOAA HYSPLIT model:
  http://www.arl.noaa.gov/ready/hysplit4.html
- EPA SPECIATE, recently updated (version 4.0):
  http://www.epa.gov/ttn/chief/software/speciate/index.html.
- Network assessment guidance:
  http://www.epa.gov/ttn/amtic/cpreldoc.html

Suggested Analyses

What types of analyses could be done with my air toxics data?

Motivation

- Ambient air toxics have been monitored since 2001/2002 as part of NATTS and even longer as part of other monitoring programs. While national-level analyses have been conducted, it is important that these data be investigated at a local, state, and regional level to better understand an area's air toxics issues.
- Regular data analysis may be conducted annually to identify potential problems with the data at the site level. Adjustments can then be made in collection or analysis to improve data quality before several years of potentially poor quality data have been collected.
- Key areas of interest
  - Is the quality of data sufficient for analysis?
  - How would air toxics be characterized in the area?
  - What are local sources of air toxics?
  - Are there changes in toxics concentrations over time?

Suggested Analyses

What's Covered in This Section

A set of potential analyses using Arizona data is used as an example.

- A sample analysis of an urban data set is outlined from start to finish to provide a thorough example. These data were previously assessed and readily available.
- Note that this analysis is an example and is not intended to show the only way air toxics analyses should be performed. Deviations or additional analyses may be necessary depending on the data or the analyst's objectives.

Introduction to the Data

Overview

- The sample data set used throughout this section is from an air toxics study performed in Arizona as part of the Joint Air Toxics Assessment Project (JATAP).
- The purpose of the study was to determine which air toxics are of most concern to the area and tribal communities.
- Twenty-four-hour air toxics samples were collected every sixth day. On some days at some sites, two 12-hr samples were collected; for this analysis, these samples were 24-hr averaged. Only gaseous air toxics were collected and discussed here.
- A considerable quality assurance effort was made
  - Duplicate samples (collocated)
  - Replicate data (additional chemical analysis on canister)
  - Interlaboratory comparisons (more than one laboratory was involved)
  - Data validation
- For the trend assessment, we used historical data at two longer-term sites in the study area to illustrate air toxics concentrations over time in the area.

Introduction to the Data

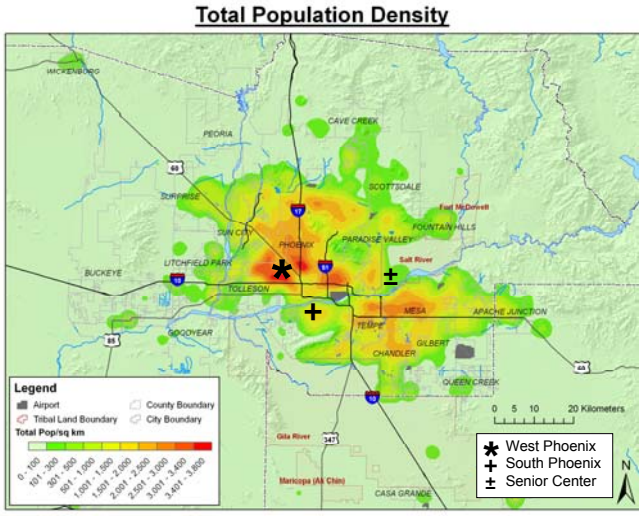
Monitoring Site Locations

The map shows the eight monitoring sites in the study. The West Phoenix, South Phoenix, and Senior Center sites are used most frequently in the sample analyses. The St. Johns site was operated by the Gila River Indian Community. The Senior Center site was operated by the Salt River Pima-Maricopa Indian Community.

Understanding Sources

Population Density

- The map shows population density in the study area. The three focus sites are indicated.
- Data from these sites help identify the most populated areas and potential air toxics source locations (e.g., high population density ≈ higher emissions).
- 2000 population density data were obtained from the U.S. Census Bureau.

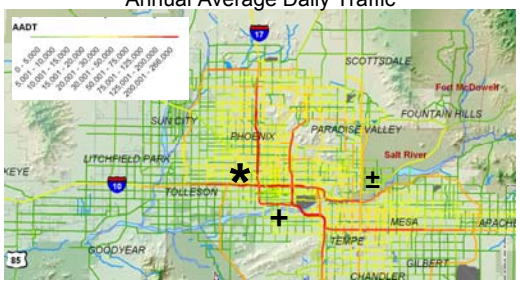


Understanding Sources

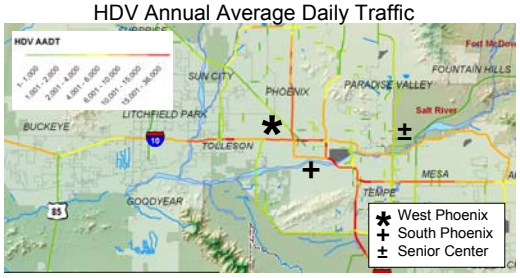
Mobile Sources

- The map shows annual average daily traffic (AADT) and heavy-duty vehicle (HDV) daily traffic for the study area (number of vehicles per day). The three sites of interest for this example are shown.
- AADT is an indicator of the relative on-road mobile source activity, and corresponding emissions levels, in the study area.
- Traffic data were obtained from the Arizona Department of Transportation (ADOT).

Annual Average Daily Traffic



HDV Annual Average Daily Traffic

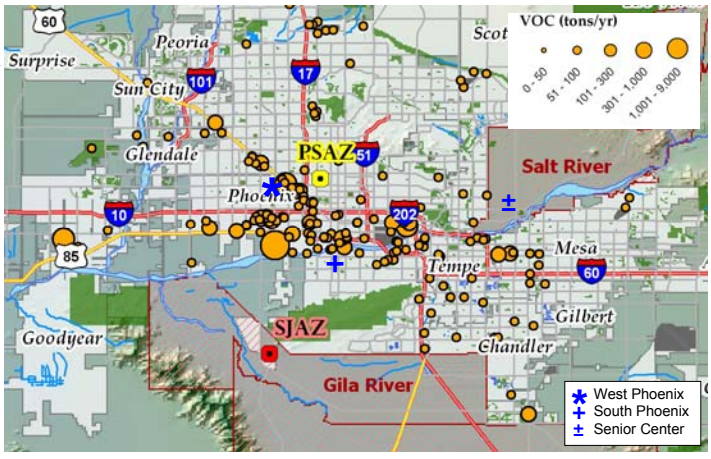


Understanding Sources

Point Sources

- The map shows point source emissions for total VOCs in the study area. The three sites of interest are shown on the map. Other sites in the area are also shown (Supersite [PSAZ] and St. Johns [SJAZ]).
- Note that mobile source emissions are not included in this data set (*see the average daily traffic maps on previous slide).*
- Emissions data were obtained from the 2002 NEI.

Point Source Emissions of VOCs



Using Quality Assurance Data

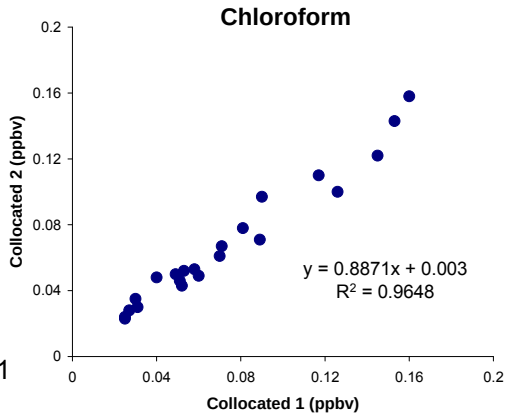
Overview

- Quality assurance (QA) is performed during sample collection and analysis to provide additional information about data quality and usefulness.
  - Collocated samples indicate agreement between *sample collection*
  - Replicate samples indicate agreement between *sample analysis*
- These data provide insight into biases and error that may occur in the process of collecting and analyzing samples.

Using Quality Assurance Data

Visual Inspection of Collocated Samples (1 of 2)

- Visual inspection of collocated samples is important to identify outliers and understand sampler performance.
- Collocated data for chloroform are plotted in the figure.
- The data indicate that chloroform is consistently measured; however Sampler 2 reported slightly lower values than Sampler 1 at higher concentrations.

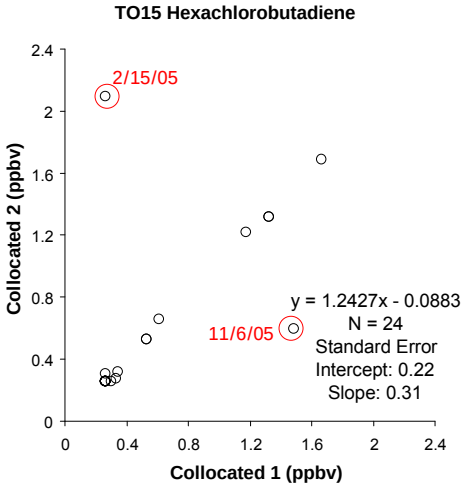


The figure shows collocated chloroform samples collected in the study. It was created with Microsoft Excel.

Using Quality Assurance Data

Visual Inspection of Collocated Samples (2 of 2)

- Collocated data for hexachlorobutadiene are plotted; outliers are circled in red. Outliers identified from collocated samples should be excluded from further data analyses.
- The data indicate that hexachlorobutadiene is not consistently measured; Sampler 2 reported lower values than Sampler 1 at high concentrations. This is consistent with observations of collocated chloroform data.



Using Quality Assurance Data

Summarizing Sample Problems for Analysis

- In site-level analyses, we typically exclude any of these failures. We flagged as suspect the pollutant identified as a problem in the indicated sample and did not use this pollutant/sample combination in subsequent analyses (e.g., toluene on 7/26/03).
- Flag 1 indicates that the percentage error was greater than 50%. Flag 2 indicates that the absolute difference in the two species was greater than three times MDL. Flag 3 indicates that the replicate or collocated average was suspect.

Date	Species Name	Flag 1	Flag 2	Flag 3	Suspect
7/26/2003	Toluene	x	x		x
7/26/2003	1,3,5-trimethylbenzene	x			x
7/26/2003	1,2,4-trimethylbenzene	x			x
8/25/2003	MTBE			x	x
8/25/2003	Methyl ethyl Ketone			x	x
8/25/2003	n-octane			x	x
8/25/2003	1,3,5-trimethylbenzene			x	x
8/25/2003	1,2,4-trimethylbenzene			x	x
9/24/2003	Methyl ethyl Ketone		x		x

Data Completeness

Overview

- For the site-level analysis, we summarized available data and calculated data completeness based on expected samples.
- This step included calculating the number of valid samples versus the expected number of samples based on collection frequency.
- In general, 75% data completeness is required to calculate valid aggregated values (e.g., monthly, quarterly, and annual averages).

Data Completeness

Site Level Summary

Site	Sampling	Sampling Duration	Samples Expected	Samples Available	Valid Samples	Percent Valid
Greenwood	Cartridges[a]	24-hr	61	60	60	98
	Canisters	24-hr	61	61	59	97
JLG Supersite	Cartridges[a]	24-hr	61	61	49	80
	Canisters	24-hr	61	61	55	90
Queen Valley	Canisters	24-hr	31	31	30	97
St. Johns	Canisters	24-hr and 12-hr	30 (24-hr) 62 (12-hr)	37 (24-hr) 44 (12-hr)	79	95[b]
Senior Center	Canisters	24-hr and 12-hr	30 (24-hr) 62 (12-hr)	37 (24-hr) 46 (12-hr)	83	98[b]
South Phoenix	Cartridges[a]	24-hr	61	60	52	85
	Canisters	24-hr	61	60	59	97
West Phoenix	Canisters	24-hr	61	60	59	97

- The table shows data necessary to calculate the data completeness and the percent of valid data. The number of valid samples was computed after data validation steps but shown here for a complete summary.
- A high percentage of samples from all sites were valid.
- Additional samples may be marked as suspect during the process of data analysis.

[a] Carbonyls only.

[b] This percentage is based on 24-hr average sample days.

Assessing Data Above Detection

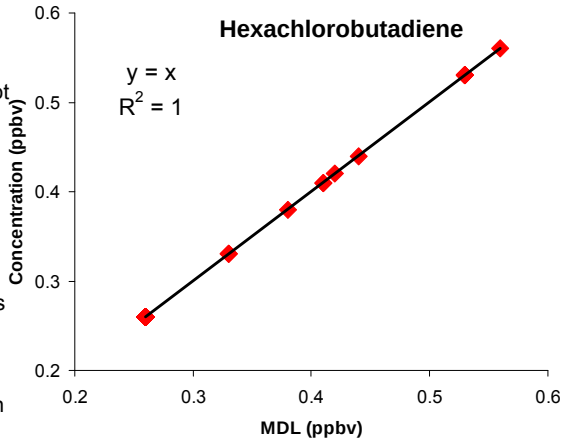
Species	2005 Percent Above MDL						
	St. Johns	Senior Center	South Phoenix	West Phoenix	Green-wood	JLG Supersite	Queen Valley
Benzene	100	99	100	100	100	100	100
Bromomethane	40	36	37	49	24	33	23
Carbon tetrachloride	89	89	89	83	100	100	100
Chloroform	43	90	77	83	98	100	53
Dichloromethane	76	94	97	98	100	100	97
Ethylbenzene	71	92	92	94	100	100	93
Hexachloro-butadiene	0	0	0	0	2	4	0

- The percent of data above detection should be calculated for each pollutant, site and year; additional calculations will be needed if monthly or seasonal aggregates are produced. The table shows an excerpt of the entire data set - the percent of data above detection for 2005. This example spans the range of data above detection observed in the data set.
- More data were below detection at St. Johns and Queen Valley, consistent with their location away from sources. Hexachlorobutadiene was typically below MDL at all sites.

< 25% Above MDL
25% to 75% Above MDL
>= 75% Above MDL

Identifying Censored Data

- Alternate MDLs were included with the study data. Because alternate MDLs are often different for each sample, it is not always clear from the data that censoring (e.g., substitution with MDL or MDL/2) has occurred. We need to ensure that all samples are treated similarly when data are aggregated.
- The agreement between concentration and MDL indicates that the alternate MDL was substituted for values below detection. These samples were identified and MDL/2 substitution was subsequently applied for data aggregation.



The graph shows the comparison of concentration values to their MDL for data at or below detection.

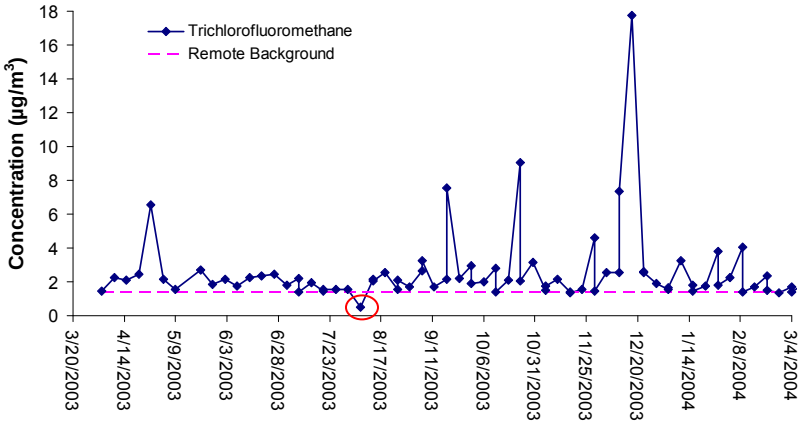
Validation Techniques

Overview

- Once data are received from the laboratory, or a data repository such as AQS, apply screening criteria during the early stages of data validation to identify suspect data that may not be representative of actual ambient concentrations.
- Perform basic visual analyses to identify potential problems in the data and to begin to understand data characteristics.
- Use knowledge of similarity of sources, lifetime, and reactivity to assist in data validation.
- The following screening checks are typically used
  - Comparison to remote background concentrations. Urban air toxics concentrations should not be lower than remote background concentrations.
  - Range checks. Check minimum and maximum concentrations for anomalous values.
  - Buddy site check. Compare concentrations at one site to nearby sites to look for anomalies.
  - Sticking check. Check data for consecutive equal data values which indicate the possibility of censored data not flagged appropriately.
  - Scatter plots. Investigate the relationship between species to identify sources and suspect data.
  - Fingerprint plots. Investigate the pattern of species concentrations and relationships among species to identify sources and suspect data.

Validation Techniques

Remote Background Check

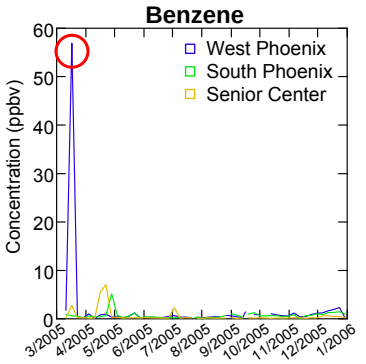


- Time series of concentrations of trichlorofluoromethane compared to background concentrations measured at remote sites in the Northern Hemisphere.
- A significant dip in concentrations is circled in red. Concentrations at this monitor were typically equal to or greater than background concentrations,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

Validation Techniques

Buddy Site Check

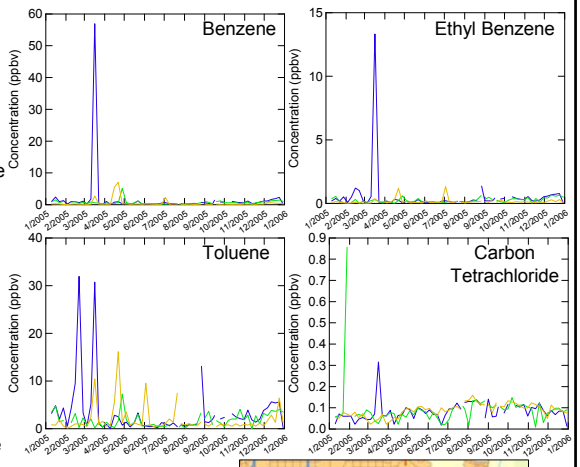
- Time series of benzene concentrations for three Phoenix sites.
- There is a suspect data point at the West Phoenix site in March 2005, which is not corroborated by the other sites. This indicates that the data point should be considered suspect because a concentration spike of that magnitude should 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.



Validation Techniques

Time Series

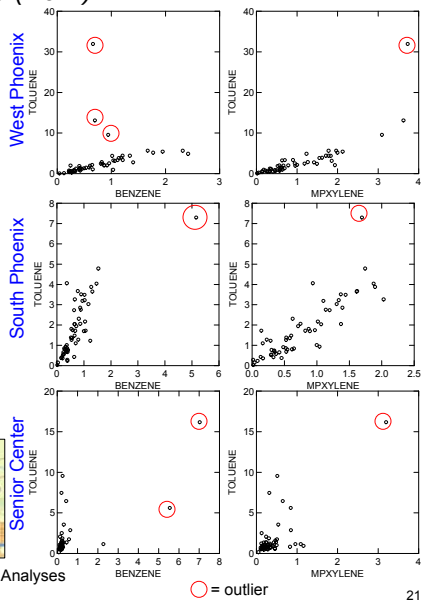
- The fact that these species peak at the same time is suspicious, because an increase of that magnitude from typical mobile source emissions is unlikely. However, an unusual event may have occurred, such as a gasoline spill very near the West Phoenix site that could have led to the high concentrations.
- Examining the time series of carbon tetrachloride helps confirm or reject this theory because there are no likely sources that would cause a spike of that magnitude. The time series of carbon tetrachloride does show a spike on the same day indicating that the event is in fact an instrument or analysis error. All data for that date and site should be flagged as suspect and not used in subsequent analyses.



Validation Techniques

Scatter Plots (1 of 2)

- At the West Phoenix site, the correlation between toluene, benzene, and m,p-xylene is strong, indicating that this site is highly mobile source-dominated.
- Outlier data points may point to data issues or other source influences. For toluene outliers, high toluene concentrations are often associated with solvent use or surface coatings; thus, the samples are likely valid.
- The correlations at the South Phoenix site are not quite as strong, but still indicate that the site is likely mobile source-dominated.
- The Senior Center site, on the other hand, shows a weak correlation between the three species as expected for a site farther from fresh emissions.

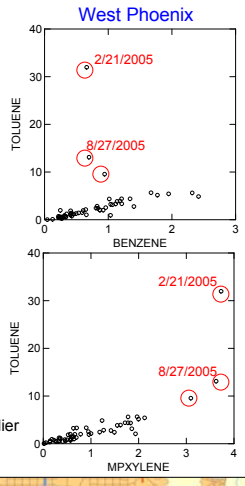


◯ = outlier

Validation Techniques

Scatter Plots (2 of 2)

- The outlier values all correspond to the unusually high toluene concentrations. Significantly, the three toluene outliers correspond with the three highest m,p-xylene events.
- These correlations indicate that the high concentrations may not be due to collection or analysis errors, but may indicate solvent or surface-coating emissions impacting the site. Further exploration might include assessing the importance of these concentrations on the annual average and looking for possible sources of toluene in the emission inventory.
- The table shows emission profiles for surface coating from EPA's SPECIATE. Xylenes and toluene account for almost one- third of this source profile supporting the hypothesis that the high concentration events are solvent-driven.



◯ = outlier

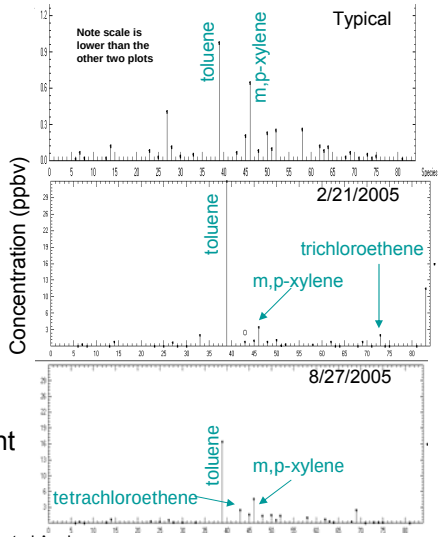
Profile Number: 6002
Profile Name: Surface Coating Operations (Industrial)
Percent Total: 100

POLLUTANT	CAS No.	Percent
ISOMERS OF XYLENE	1330207	15.800
TOLUENE	108883	14.700
METHYL ETHYL KETONE	78933	8.100
DIETHYLENE GLYCOL	111466	6.600
N-BUTYL ALCOHOL	71363	6.400

Validation Techniques

Fingerprint Plots

- A typical fingerprint can be quantitatively determined (e.g., median sample composition) or qualitative (e.g., visual inspection of all fingerprints).
- The figures to the right show a typical fingerprint plot and fingerprint plots for 2/21/2005 and 8/27/2005 (the two dates of the highest outlier events in the previous slides).
- A review of fingerprints listed in EPA's SPECIATE shows that toluene and xylenes are prominent components of surface coatings.



Validation Techniques

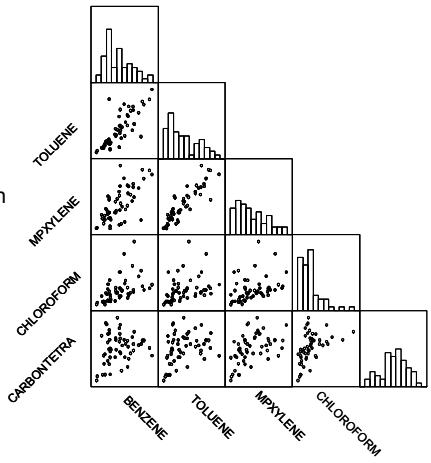
Summary

- What have we learned from applying these validation techniques?
  - Additional invalid and suspect data points were identified.
  - Data quality and limitations are better understood.
  - Spatial and temporal characteristics of the data are more thoroughly indicated.
  - Hypotheses about possible source influences for further investigation can be formed.
- These are a few examples of the data validation process that would be performed on the data set.
- Remember, data validation continues as part of data analysis.

Basic Understanding of Data

Scatter Plot Matrices

- The graph to the right shows scatter plot relationships for five pollutants at the South Phoenix site. Note that previously identified outliers have been removed.
- The data show a clear correlation between toluene, m,p-xylene, and benzene, indicating that these pollutants are likely from mobile sources.
- Chloroform also shows a slight correlation with the mobile source pollutants (across the second row from the bottom) but the bifurcated relationship indicates a secondary source.
- Carbon tetrachloride shows little correlation with any species and shows a histogram that is roughly Gaussian, as expected for background pollutants.



Putting Data In Perspective

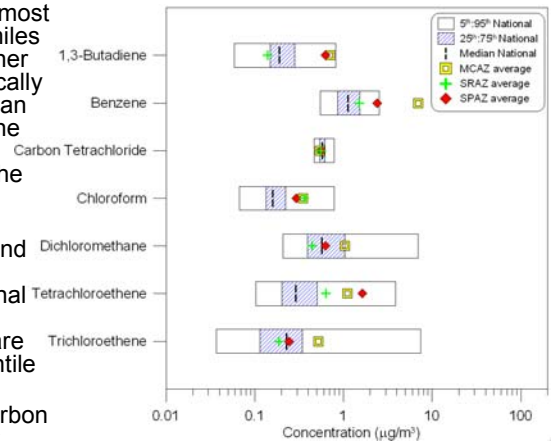
Overview

- Putting concentrations and MDLs into perspective provides a framework for comparing site-level concentrations to national levels and to other sites in the area.
- This information is useful in assessing whether concentrations are typical, low, or high and can help explain the impact of local source emissions on monitored concentrations.

Putting Data In Perspective

National Concentrations

- Though Senior Center is the most rural (although within a few miles of urban emissions) of the other sites, concentrations are typically higher than the national median and sometimes higher than the national 75th percentile concentration, showing that the site is impacted by urban emissions.
- Concentrations at the West and South Phoenix sites are also typically well above the national median. Concentrations of benzene and 1,3-butadiene are near or above the 95th percentile of national concentrations.
- National concentrations of carbon tetrachloride fall within a very small range due to its ubiquitous background concentration.

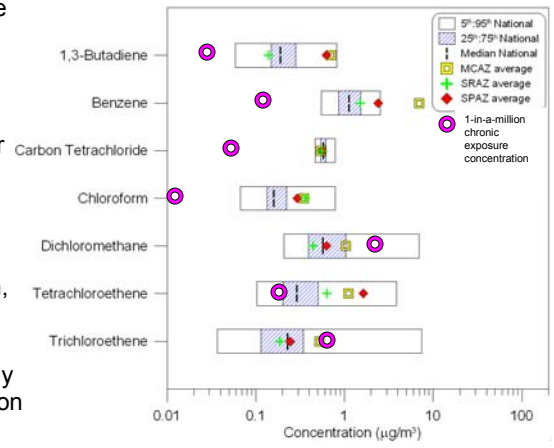


MCAZ = West Phoenix
SPAZ = South Phoenix
SRAZ = Senior Center

Putting Data In Perspective

Cancer Risk

- The figure shows the same data as the previous slide, with the addition of the chronic exposure concentration associated with a 1-in-a-million cancer risk to place health risks in perspective.
- Concentrations could be compared to other cancer risk levels: 0.1-in-a-million, 10-in-a-million, 100-in-a-million, etc.
- Concentrations are typically higher than the 1-in-a-million cancer risk level shown except for dichloromethane and sometimes trichloroethene.

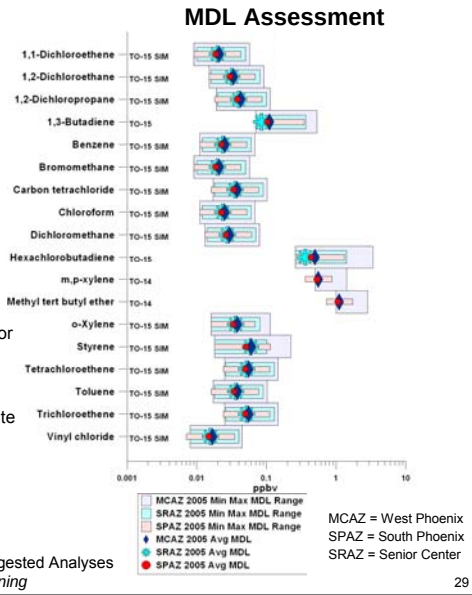


MCAZ = West Phoenix
SPAZ = South Phoenix
SRAZ = Senior Center

Putting Data In Perspective

MDLs

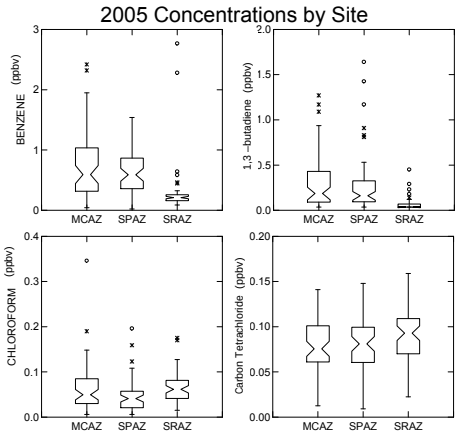
- Examining the relationship between MDLs at multiple sites is imperative to verify that MDL/2 substitutions are not biasing the data differently at different sites.
- Average MDL and minimum-to-maximum MDL range for three study sites.
- This graphical method allows the analyst to quickly confirm that MDLs are very similar between sites.
  - MDLs at the West Phoenix site (light purple bar) are sometimes higher than at other sites.
  - The difference is not enough to cause a major bias unless a high % of data < MDL. For example, hexachlorobutadiene is typically below detection so MDL/2 substitution may cause concentrations at the West Phoenix site to appear higher than at the other sites. However, hexachlorobutadiene has such a large portion of data below detection that it cannot be reliably used for many analyses.



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

Spatial Patterns

- Understanding spatial patterns is important and can provide insight into
  - Improving monitoring networks
  - Verifying and improving emission inventories
  - Verifying and improving models
  - Identifying sources
- Benzene and 1,3-butadiene concentrations are higher and more variable at the West and South Phoenix sites.
- Chloroform and carbon tetrachloride concentrations are relatively consistent at all sites.



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

Temporal Patterns

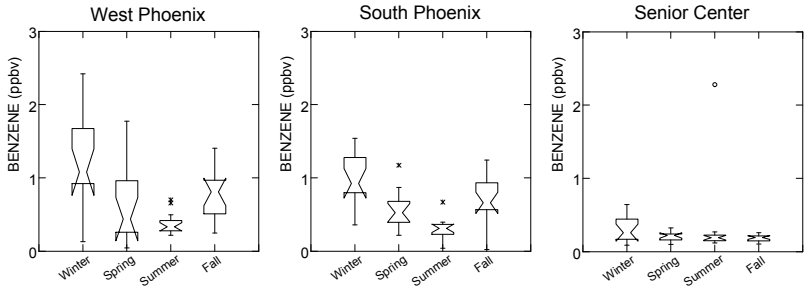
Overview

- Characterization of temporal patterns can provide information on sources, physical or chemical processes affecting air toxics concentrations, and additional data validation.
- Before beginning temporal characterization, it is recommended to create valid aggregated data sets (examples in *Characterizing Air Toxics,* Section 5) to ensure the data are representative.
- There are sufficient data records in the example data set (i.e., one year of samples collected every sixth day) to characterize seasonal and weekday/weekend patterns.
- There are too few records in this data set to create day-of-week patterns (i.e., 95% confidence intervals on the means will overlap too much across the days because of the small sample size).
- 1- to 3-hr samples were not collected so diurnal patterns cannot be investigated.

Temporal Patterns

Seasonal Patterns of Benzene

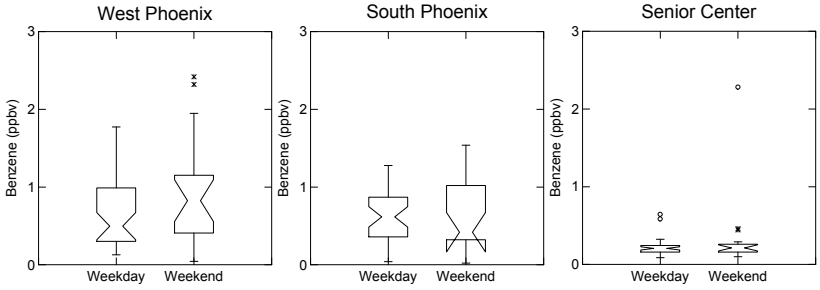
- The South and West Phoenix sites show typical benzene seasonal patterns with lower concentrations during warm months and higher concentrations during cooler months - a result of mixing height differences and reactivity with season as opposed to changes in sources.
- At the Senior Center site, benzene shows an invariant seasonal pattern. While we expect higher concentrations in winter, note that the concentrations are generally lower during all seasons at this site. All samples are well-mixed upon arriving at the Senior Center and are similar to summer concentrations at the other sites.
- These data follow expectations for urban and downwind sites. The seasonal variability for these pollutants shows that for the urban data, computed annual averages without the winter quarter would be biased low and vice versa for a missing summer quarter.



Temporal Patterns

Weekday/Weekend

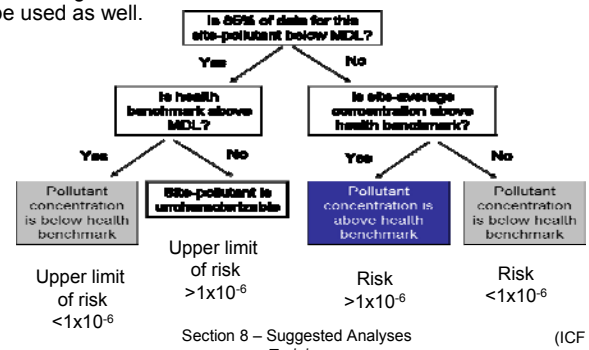
- We would expect lower MSAT concentrations on weekends, but in practice this is not always observed.
- The West Phoenix site shows higher weekend concentrations, but the difference is not statistically significant at 95% confidence. This difference may indicate that additional weekend events near the site are causing benzene emissions. For example, monitors placed near a facility with high use on weekends, such as a recreational facility, may cause this pattern. Additional investigation of the surrounding area may be warranted but was not done.
- The South Phoenix site shows slightly lower weekend concentrations (but not statistically significant). This pattern is more typical of urban sites at a national level.
- The Senior Center site shows invariant weekday/weekend patterns consistent with the well-mixed and aged nature of samples arriving at the site.



Risk Screening

Overview

- Risk screening may provide a summary of ambient concentrations of air toxics that may be of concern.
- To identify species which may indicate higher risk, follow the decision tree below for each pollutant.
- After risk species have been identified, you may wish to create risk-weighted annual averages.
- The screening here uses the 1-in-a-million cancer risk level – one could select a higher or lower risk level and define the benchmark depending on the purpose of the screening. Other health effects, such as non-cancer threshold values, could be used as well.



(ICF Consulting, 2004)

Risk Screening

West Phoenix Site

West Phoenix data necessary for risk screening

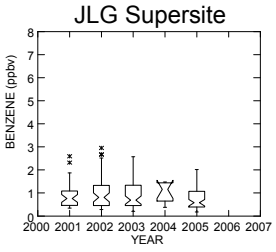
Pollutant	% Below Detection	1-in-a-million cancer risk (ppbv)	Average Method Detection Limit (ppbv)	West Phoenix Site Average Concentration (ppbv)
Benzene	0	0.040	0.50	1.7
Hexachlorobutadiene	100	0.0043	0.13	0.17

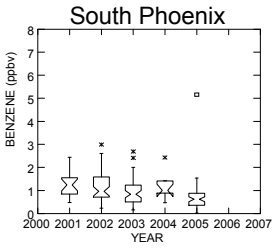
- Perform risk screening by applying all the data listed in the table to the risk-screening decision tree. Screening may be performed on a range of risk levels and also for non-cancer levels of concern.
- Benzene
  - More than 85% of data is above detection so there is high confidence in measured concentrations.
  - The site average concentration is above the chronic exposure concentration associated with a 1-in-a-million cancer risk.
- Hexachlorobutadiene
  - 100% of data is below detection so we have no confidence that the measured concentrations accurately reflect ambient concentrations. However, we know that concentrations are below the MDL (note that MDLs varied by sample and the average is shown).
  - The chronic exposure concentration associated with a 1-in-a-million cancer risk is below the MDL.
  - We know that both the data and the cancer risk level of 1-in-a-million are below the MDL-- improved data collection methods are necessary to more accurately characterize risk. The upper limit of risk is based on the MDL.

Trends

Five-Year Trends

- Inter-annual trends were investigated for all pollutants with sufficient data.
- The notched box plots show benzene concentrations at two sites with data available from 2001 to 2005.
- Benzene concentrations have remained relatively flat at the JLG Supersite and South Phoenix site. However, there is a statistically significant difference between the 2001 and 2005 concentrations at the South Phoenix site.
- Trends for other air toxics showed similarly consistent concentrations from year to year for this time period.
- Once six years of data are available, two 3-yr averages should be compared (i.e., average of 2001, 2002, and 2003 vs. 2004, 2005, and 2006; see *Quantifying Trends*, Section 6).

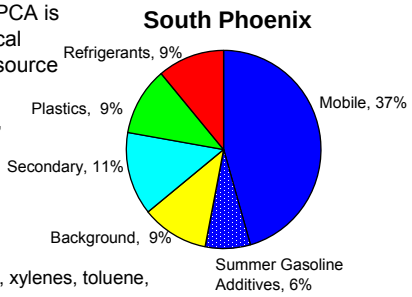




Source Apportionment

Example (1 of 2)

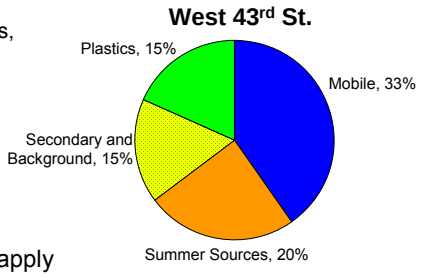
- Principal component analysis (PCA) was applied to air toxics data from two sites, South Phoenix and West 43rd St., as part of an exploratory analysis.
- PCA uses correlation or covariance between each pair of variables to estimate relationships. PCA is relatively easy to perform with basic statistical packages; however, the analyst must infer source types from the factors.
- In South Phoenix, PCA resolved six factors, accounting for 81% of the variance. These data are illustrated in the top pie chart (note that the percentages are percent of variance explained in the data, not percent of the mass).
  - 37%: Mobile sources (benzene, 1,3-butadiene, xylenes, toluene, ethylbenzene)
  - 9%: Background (carbon tetrachloride, methyl ethyl ketone)
  - 11%: Secondary (formaldehyde, acetaldehyde)
  - 6%: Summer gasoline additives (MTBE)
  - 9%: Plastics (methylene chloride)
  - 9%: Refrigerants/AC (dichlorodifluoromethane, trichlorofluoromethane)



Source Apportionment

Example (2 of 2)

- PCA resolved four factors at the West 43rd Phoenix site, accounting for 82% of the variance; carbonyl compound data were not available at this site (so fewer factors were resolved).
  - 33%: Mobile sources (benzene, xylenes, toluene , ethylbenzene)
  - 20%: Summer sources, e.g., BBQs, air conditioning (trichlorofluoromethane, acetylene, propylene)
  - 14%: Secondary/background (MEK, MTBE, dichlorodifluoromethane)
  - 15%: Plastics (trimethylbenzenes)
- Next steps in this analysis may be to apply CMB or PMF to estimate source contributions.



Model-to-Monitor Comparisons

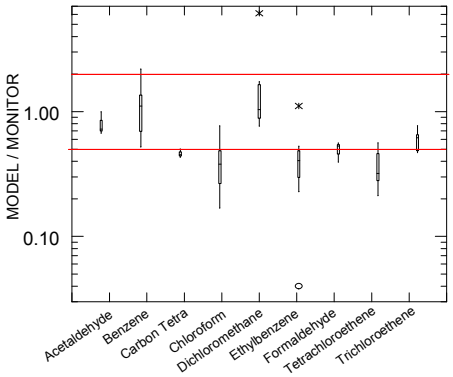
Overview

- EPA periodically performs national-scale air toxics assessment (NATA) to identify and prioritize air toxics emissions source types and locations which are of greatest potential concern in terms of contributing to population health risk. Modeled concentration estimates for 177 air toxics and DPM are provided by county. For more information on NATA see http://www.epa.gov/ttn/atw/natamain/.
- As part of an evaluation of how models used in NATA performed, EPA conducted a monitor-to-model evaluation to evaluate modeled values.
- A comparison of monitored and modeled data may help in checking the uncertainty of modeled values.

Model to Monitor Comparisons

Example

- The figure shows the ratio of NATA99 modeled data to annual averages computed from monitored data at the study area sites to indicate the accuracy of modeled data.
- When comparing modeled-to-monitored concentrations, results within a factor of 2 are considered reasonable agreement (U.S. Environmental Protection Agency, 2006b).
- Acetaldehyde, benzene, dichloromethane, and trichloroethene typically agreed within a factor of 2, consistent with national-level comparisons of modeled and monitored data.
- However, ethylbenzene, formaldehyde, carbon tetrachloride, chloroform, and tetrachloroethylene showed monitored concentrations more than a factor of 2 higher than model estimates at study area sites.



The graph shows the comparison of modeled to monitored annual averages at the study area sites.

Summary

What We Learned from this Data Analysis

- **Data Validation – were data of sufficient quality for analysis?**
  - Overall data completeness was sufficient for analysis.
  - Species data above detection were sufficient to perform most analyses, while a significant percent of some species' data was below detection.
  - QA analyses showed that agreement between collocated data was typical of conclusions from other studies.
  - Data were validated using time series, buddy site checks, scatter plots, and fingerprint plots. Invalid data points were identified and removed.

Data were determined to be of sufficient quality for most analyses.

Summary

What We Learned from this Data Analysis

- **Data Characterization – How would air toxics in the area be characterized?**
  - Air toxics concentrations in the study area were compared to national concentrations and cancer benchmarks; concentrations of most air toxics are above the national median concentration at all study sites and are typically above cancer benchmarks. It is not clear why, and an evaluation/ development of the air toxics emission inventory is planned
  - MDLs at study sites were found to be similar across sites indicating that data are comparable.
  - Spatial analyses showed concentrations were similar at the South and West Phoenix sites while significantly lower concentrations of MSATs at the Senior Center site were consistent with the sites' proximity to emissions.

Summary

What We Learned from this Data Analysis

- **Data Characterization – How would air toxics in the area be characterized? (Cont.)**
  - Temporal patterns were investigated.
    - Seasonal patterns showed expected trends at the West and South Phoenix sites. Senior Center site benzene concentrations were low and showed no seasonal trend consistent with aged air impacting the site.
    - There were no significant weekend/weekday patterns, a typical result as truck traffic or weekday carryover often cause increased Saturday concentrations. There were not enough data points to reliably investigate trends by day-of-week.
  - Ambient concentrations were compared to NATA 1999 modeled data. About half the species monitored at study area sites were more than two times above their modeled concentration values.
  - Risk screening was performed and the species of most concern were found to be benzene, 1,3-butadiene, acetaldehyde, carbon tetrachloride, chloroform, and tetrachloroethene. Hexachlorobutadiene is likely a significant contributor to risk, but is not measured well enough to quantify the risk.

Summary

What We Learned from this Data Analysis

- **Trends – Are there changes in air toxics concentrations over time?**
  - Five-year trends (2001-2005) showed no significant change at the study sites
- **Advanced Analyses – What are local sources of air toxics?**
  - PCA was performed for South Phoenix and West 43rd St. Mobile sources contributed to about one-third of the variance at both sites. Pollution related to plastics, background species, and secondary species contributed about another third. Both sites showed significant influence from "summer" pollutants related to BBQs, air conditioning/ refrigerants, and summer fuel additives.
  - Mobile source influences were confirmed by other analyses.
    - Scatter plots showed strong correlation between mobile source air toxics.
    - Spatial patterns revealed higher mobile source concentrations near busy roadways and much lower concentrations in remote areas
  - Short-term solvent emissions events were identified during the process of data validation.