Database Preparation

Prepared by:
Michael C. McCarthy
Hilary R. Hafner
Eric A. Gray
Sonoma Technology, Inc.
Petaluma, CA

Presented to:
Air Toxics Monitoring Data Analysis Workshop
Raleigh, NC
September 28, 2005

Overview (1 of 2)

- Database goals
- Initial database
- Cleaning, validating, and flagging
- Averaging procedures
- Final database suitability analysis
- Conclusions
- References and acronyms

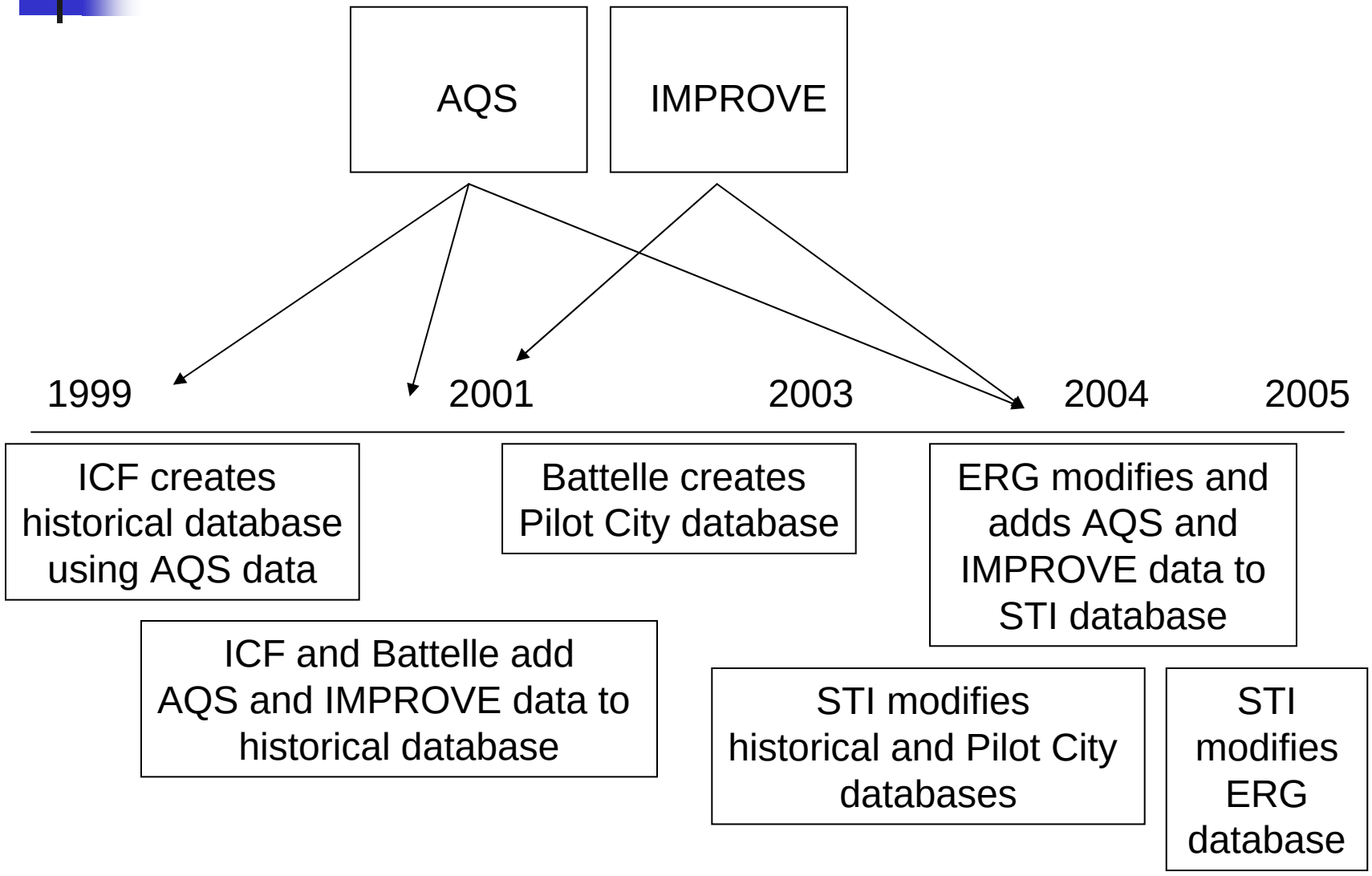
Overview (2 of 2)

- Details on data validation, flagging, and averaging
- Details on what species were available for subsequent analysis
- Sets the stage for the next three talks
- Big Picture questions:
  - Which air toxics species are adequately represented in the database for temporal and spatial analysis?
  - How should missing data and data below detection levels be treated?
  - What is our confidence in the data?

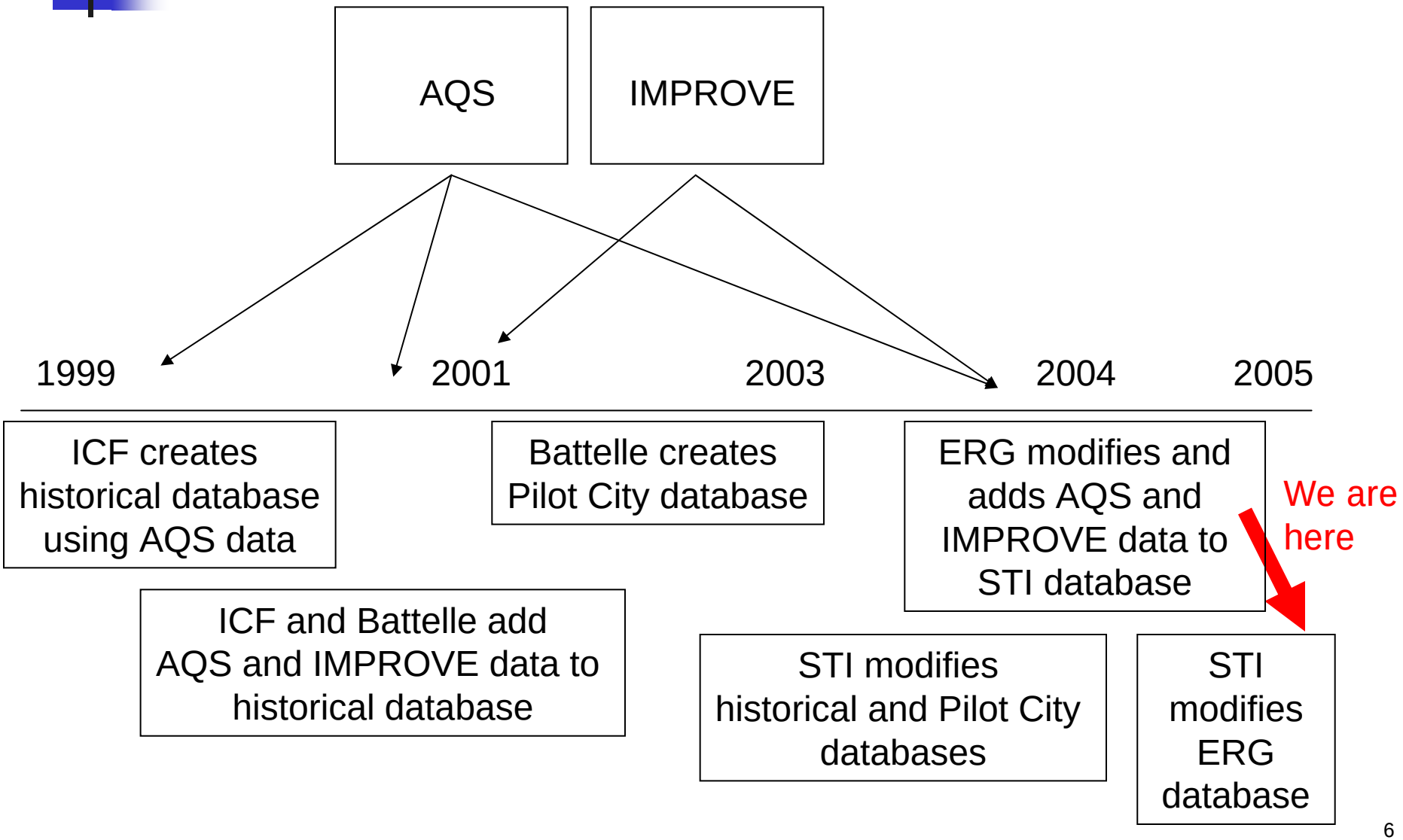
Database Goals

- Create a database with only reported values
- Create defensible temporal averages that can be used to investigate temporal and seasonal variability
- Continue to refine database validation and averaging steps

Database History (1 of 2)



Database History (2 of 2)



Database Details

- 39 air toxics (including size fractions)
- 1990 through 2003
- 7,000,000 individual concentration records
- Species with largest number of samples included benzene, lead (tsp), o-xylene, ethylbenzene, and toluene
- Data were of varying sample duration (e.g., 1-hr, 3-hr, 24-hr)

Air Toxics in the Database

Volatile Organic Compounds (VOCs)

Benzene
1,3-Butadiene
Xylenes (o-, m-, p-, and sum)
Ethylbenzene
Toluene
Formaldehyde
Acetaldehyde
Carbon Tetrachloride (Tetrachloromethane)
Chloroform (Trichloromethane)
1,2-Dichloropropane
Methylene Chloride (Dichloromethane)
Tetrachloroethylene (Perchloroethylene, Tetrachloroethene)
Trichloroethylene (Trichloroethene)
Vinyl Chloride (Chloroethene)

Particulate Matter (PM) metals

Arsenic $PM_{2.5}$ and (tsp)
Beryllium (tsp)
Cadmium $PM_{2.5}$ and (tsp)
Chromium VI, $PM_{2.5,}$ and (tsp)
Lead $PM_{2.5}$ and (tsp)
Manganese $PM_{2.5}$ and (tsp)
Mercury $PM_{2.5}$ and (tsp)
Nickel $PM_{2.5}$ and (tsp)

Mobile source-dominated
Carbonyl compounds
Chlorinated
Metals
PAHs

Polycyclic Aromatic Hydrocarbons (PAH, SVOCs, POM)

(Total tsp and vapor)
Acenaphthene
Acenaphthylene
Anthracene
Benzo(a)anthracene
Chrysene
Fluoranthene
Naphthalene
Phenanthrene
Pyrene
(Total PM_{10} and vapor)
Benzo(b)fluoranthene
Benzo(k)fluoranthene
Dibenz(a,h)anthracene
Indeno(1,2,3-cd)pyrene

Which Important HAPs Are Not in the Database? 1999 NATA Assessment Ranking

Green = In Database
Yellow = Poorly represented in Database
Red = Not in database

	Rank of the number of people exposed to cancer risk greater than…			
Pollutant	**>1*10^{-4}**	**>1*10^{-5}**	**>1*10^{-6}**	**National measurements available?**
Benzene	3	1	3	Yes
Chromium VI	2	6	10	Some
Coke oven emissions	1	3		No
Naphthalene	10	4	8	Some
1,3-butadiene		2	7	Yes
Carbon Tetrachloride		8	2	Yes
POM (total)	5	7		Some
Hydrazine (probably bogus)	4	9		No
Tetrachloroethene		5	9	Yes
Ethylene dibromide		10	6	Some
Bis(2-ethylhexyl)Phthalate			1	No
Acetaldehyde			4	Yes
1,1,2,2-tetrachloroethane			5	Some
Arsenic	6			Yes
Ethylene oxide	7			Some
Cadmium	8			Yes
Benzidine	9			No

U.S. EPA, 2005

Data Cleaning Steps (1 of 2)

1. Imported database into Structured Query Language (SQL)
2. Performed initial quality control steps
   - Checked for duplicate records (There were none.)
   - Removed invalid records. Invalid records included those with
     - previous invalid flags (from AQS or IMPROVE)
     - Missing units
     - units that had been assigned as default AQS units
     - missing site, date, duration, or pollutant information
     - Null concentration and MDL fields
     - concentrations equal to MDL/2 or MDL

Data Cleaning Steps (2 of 2)

3. Flagged records as suspect that met either of the following criteria:
   - Concentration was below MDL, but not zero, MDL, or MDL/2
   - No MDL value was reported
4. Removed records with concentrations that were below remote background values

Examples from McCarthy et al., 2005, *in press*

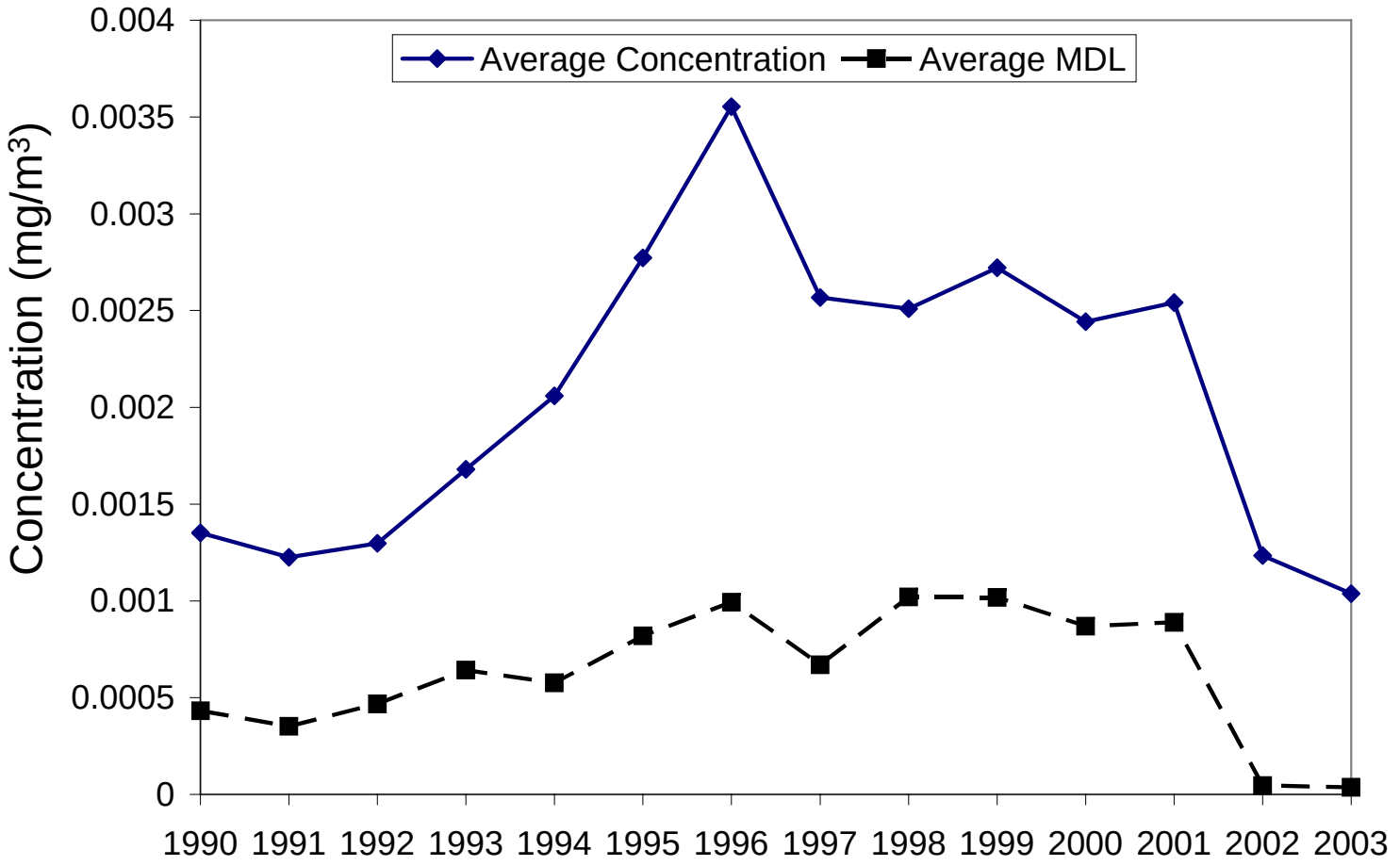
Pollutant	Minimum background ($\mu g/m^3$)	70% of minimum background ($\mu g/m^3$)
benzene	0.047	0.033
carbon tetrachloride	0.616	0.431
formaldehyde	0.14	0.098
tetrachloroethylene	0.015	0.011

MDL Issues

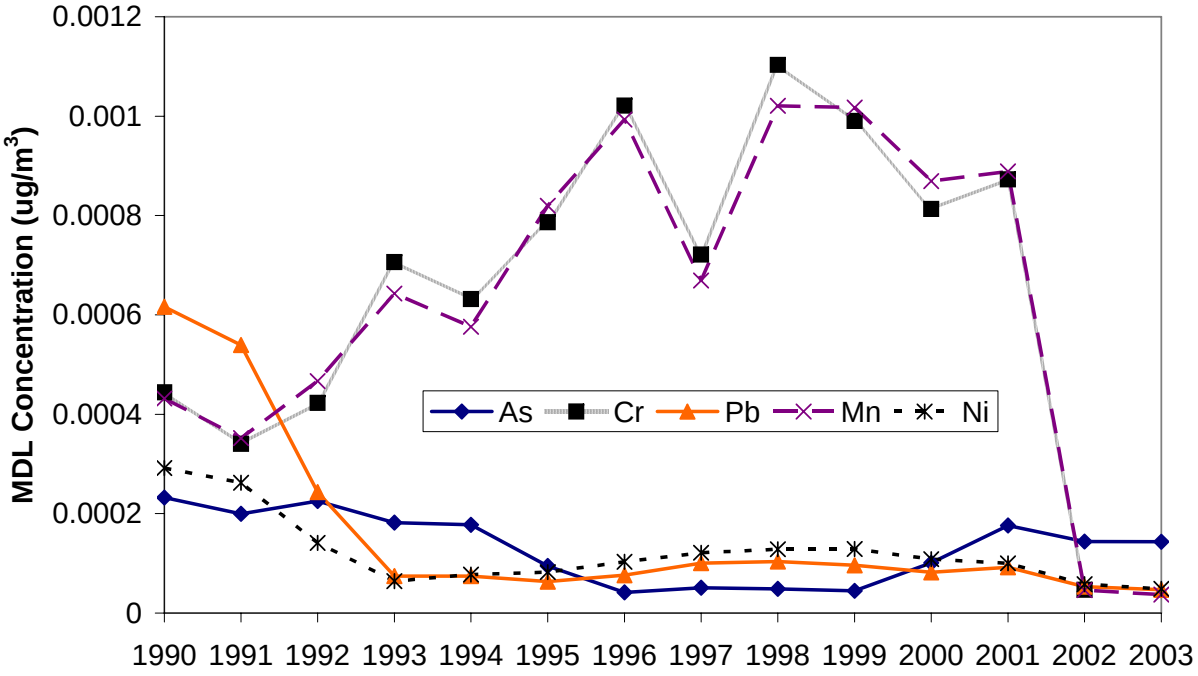
- About 40% of the data records had reported concentrations that were either zero, MDL, or MDL/2.
- These records were considered invalid for this study. This approach
  - creates a positive bias in mean concentrations for some species
  - reduces the amount of data available
  - enhances our confidence in the remaining average concentrations

IMPROVE Metal MDLs: Manganese $PM_{2.5}$

Changes in average concentration track changes in MDL. Trends are suspect.



IMPROVE Metal MDLs: Changes Over Time



Between 2001 and 2002, chromium and manganese measurements were switched from PIXE to XRF.

PIXE = Particle-Induced X-ray Emission spectroscopy
XRF = X-Ray Fluorescence spectroscopy

For all elements, it was suggested that trends only be analyzed using concentrations that are at least 10 times MDL by experts from the IMPROVE network (White, personal communication, 2005)

IMPROVE metals were not used for annual trend analysis; however, they were used for seasonal analysis.

Database Validation

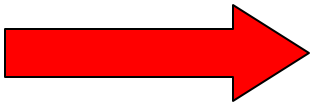
- More than 98% of toluene and ethylbenzene measurements were invalidated because records were assigned default units.
- The database contains far fewer m-&p-xylene than o-xylene records because of analytical difficulties in separating the m- and p-xylene isomers. *The combined m-&p-xylene (AQSID=45109) concentrations should be added to the database for future work.*
- Some pollutants appeared to have substituted values, such as MDL/2. Because MDL information was missing, these data were not removed during validation steps. Two air toxics, vinyl chloride and 1,2-dichloropropane, were most affected and should be considered suspect.

Data Averaging Steps (1 of 3)

- Average concentrations were generated for
  - Daily averages: 24-hr averages of subdaily sample durations (e.g., 1-hr or 3-hr samples)
  - Seasonal averages: mean concentrations of daily averages collected during
    - Quarter 1 (Jan-Mar) —Cool
    - Quarter 2 (April-June) —Warm
    - Quarter 3 (July-Sep) —Warm
    - Quarter 4 (Oct-Dec) —Cool
  - Annual average: mean of seasonal averages, rather than daily averages

Data Averaging Steps (2 of 3)

- Data completeness criteria applied to >75% of expected samples
  - For daily averages, for example, at least 18 of 24 hourly samples were required
  - For annual averages, at least three of four valid seasonal averages were required
  - For seasonal averages, sampling frequency is required to calculate the expected number of samples—this was not available in the database



Data Averaging Steps (3 of 3)

- Therefore, two minimal criteria were used to assess the completeness of seasonal averages:
  - At least 6 valid measurements (i.e., >75% completeness for 1-in-12 day sampling)
  - At least 58 days between the first and last measurement (i.e., the data must span at least 2 months)

Database Factoids

- 1,883,065 valid subdaily records
  - Benzene, o-xylene, and 1,3-butadiene had the highest numbers of subdaily records
  - Over 99.4% were 1-hr or 3-hr sample duration
- 1,387,141 valid 24-hr average records
  - Lead (tsp), $PM_{2.5}$ metals (lead, manganese, arsenic, and chromium), and benzene had the highest number of records
  - PAHs had the lowest number of records
- 82,638 valid seasonal averages
- 17,623 valid annual averages

Diurnal Variability Pollutants

Pollutant	1-hr samples	3-hr samples
Benzene	679,650	72,957
o-Xylene	657,382	67,706
1,3-Butadiene	169,380	20,365
Formaldehyde	22,717	57,372
m-Xylene	6,075	12,082
Ethylbenzene	5,869	1,226
Toluene	4,996	17
Trichloroethylene	2,523	5,664
Tetrachloroethylene	1,732	6,831
Acetaldehyde	729	35,723
p-Xylene	350	12,578
Methylene Chloride	310	6,070
Chloroform	192	6,018

Only VOCs had sufficient subdaily samples to explore diurnal variability.

Other HAPs with subdaily samples included vinyl chloride, 1,2-dichloropropane, and the sum of o-,m-, and p-xylenes. These HAPs did not have sufficient quality measurements to investigate their diurnal variability.

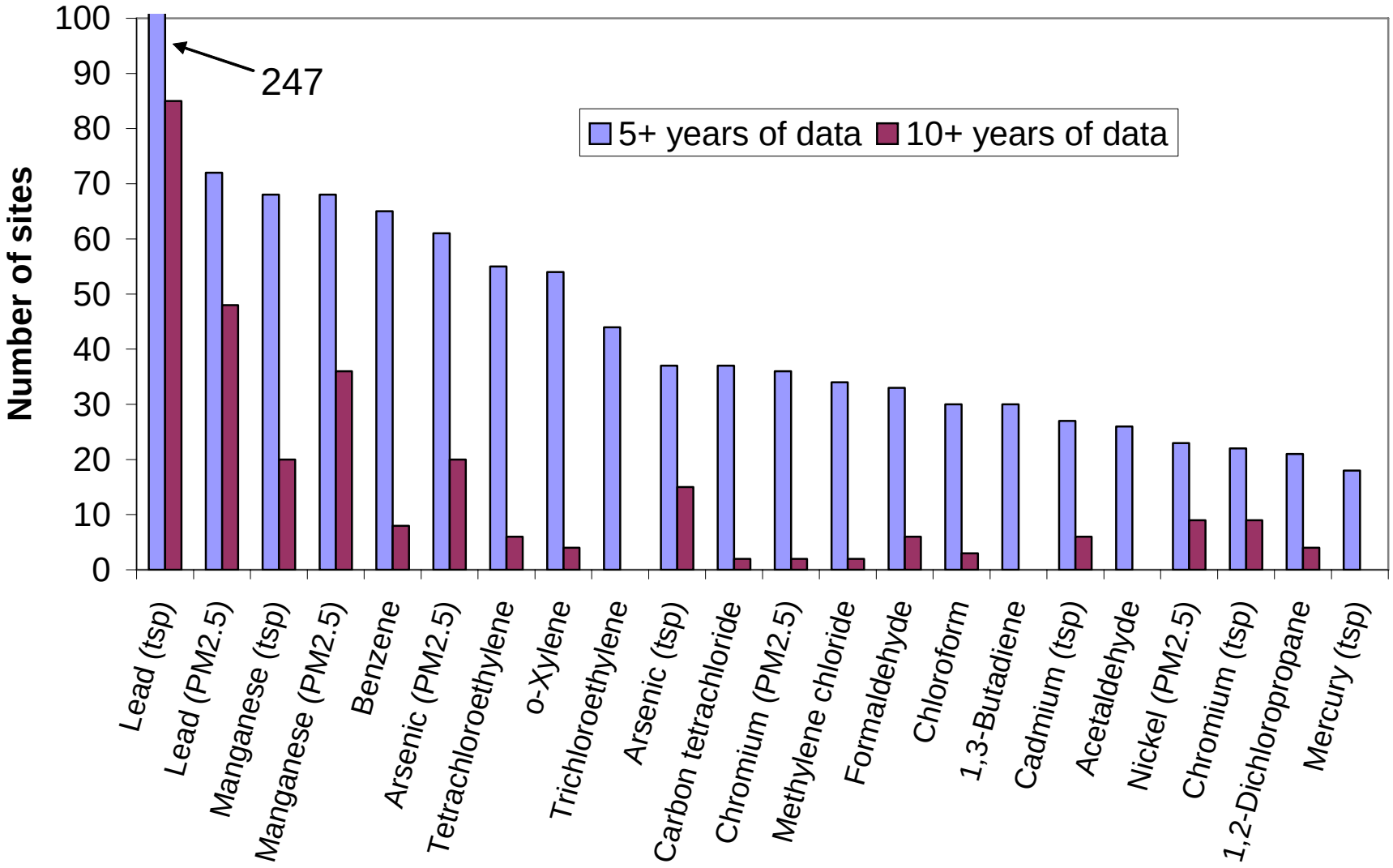
Seasonal Variability Pollutants

Pollutant	Seasonal Averages	Pollutant	Seasonal Averages
Lead (tsp)	12,433	1,3-Butadiene	2,325
Lead $PM_{2.5}$	6,147	Arsenic (tsp)	2,181
Benzene	5,931	Nickel $PM_{2.5}$	2,106
Manganese $PM_{2.5}$	5,688	Chromium (tsp)	1,807
Arsenic $PM_{2.5}$	5,164	Mercury $PM_{2.5}$	1,581
o-Xylene	4,894	Cadmium (tsp)	1,562
Chromium $PM_{2.5}$	4,400	Cadmium $PM_{2.5}$	1,540
Manganese (tsp)	3,638	Mercury (tsp)	1,052
Tetrachloroethylene	3,489	1,2-Dichloropropane	1,018
Carbon Tetrachloride	3,296	Vinyl Chloride	738
Formaldehyde	2,952	Indeno(1,2,3-cd)pyrene	465
Methylene Chloride	2,837	Benzo(b)fluoranthene	408
Acetaldehyde	2,637	Beryllium (tsp)	391
Trichloroethylene	2,629	Nickel (tsp)	370
Chloroform	2,525	m-Xylene	278

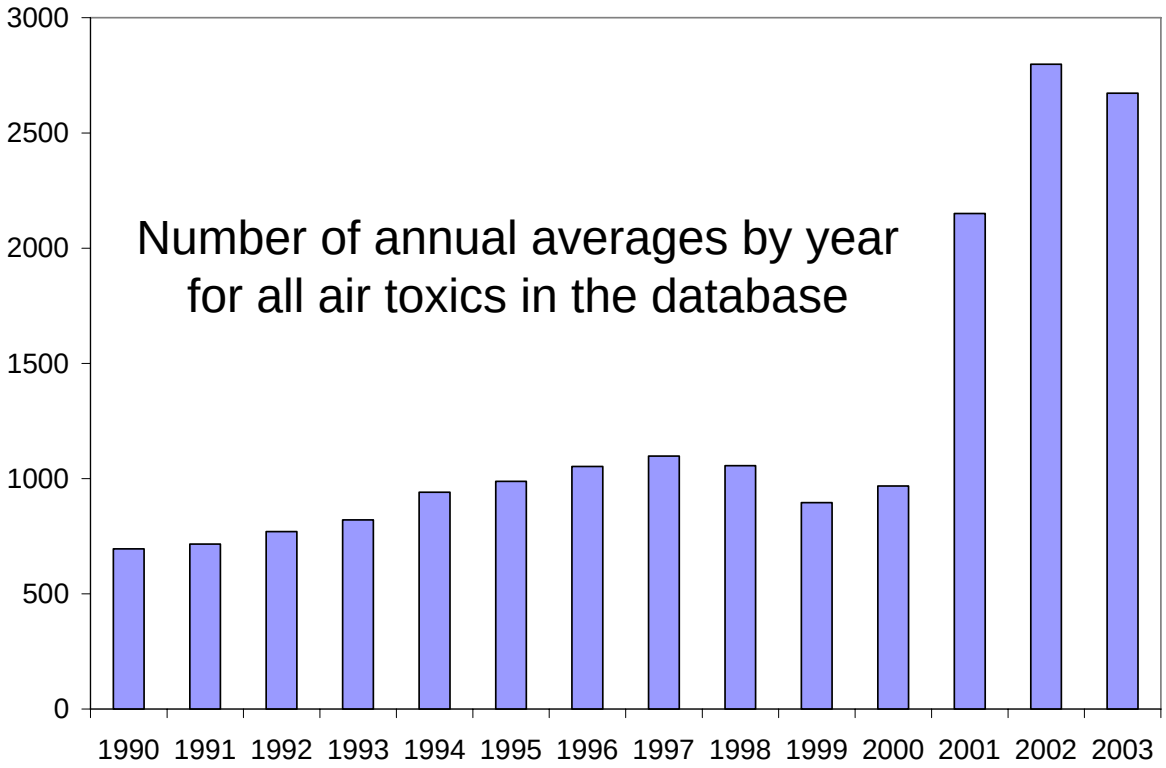
Other HAPs with sufficient seasonal average concentrations for seasonal analysis include p-xylene, chromium VI, toluene, benzo(k)fluoranthene, and ethylbenzene.

In total, 35 air toxics had sufficient seasonal average concentrations for analysis of seasonal variability.

Length of Annual Trend



Annual Trends in the Future



By the end of 2006, more sites should be available for trends analysis.

Annual Trends Pollutants

Volatile Organic Compounds (VOCs)	Particulate Matter (PM) Metals
Benzene	Arsenic (tsp)
1,3-Butadiene	Cadmium (tsp)
o-Xylenes	Chromium (tsp)
Formaldehyde	Lead (tsp)
Acetaldehyde	Manganese (tsp)
Carbon Tetrachloride	
Chloroform	
Methylene Chloride	
Tetrachloroethylene	
Trichloroethylene	

15 of the original 39 pollutants had sufficient data to perform annual trend analysis.

National Spatial Variability Pollutants

Volatile Organic Compounds (VOCs)	Particulate Matter (PM) metals
Benzene	Arsenic $PM_{2.5}$ and (tsp)
1,3-Butadiene	Chromium $PM_{2.5}$ and (tsp)
o-Xylenes	Lead $PM_{2.5}$ and (tsp)
Formaldehyde	Manganese $PM_{2.5}$ and (tsp)
Acetaldehyde	Nickel $PM_{2.5}$ and (tsp)
Carbon Tetrachloride	
Chloroform	
Methylene Chloride	
Tetrachloroethylene	
Trichloroethylene	

All pollutants shown had at least 20 citywide averages from 2000 to 2003.

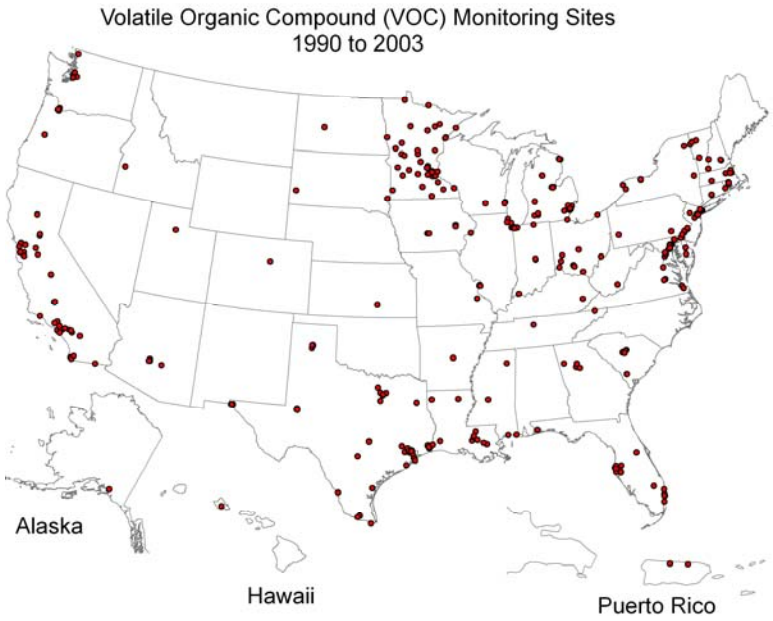
All pollutants shown were monitored at 2 or more sites in at least 10 cities between 2000 and 2003 except chromium (tsp) and nickel (tsp).

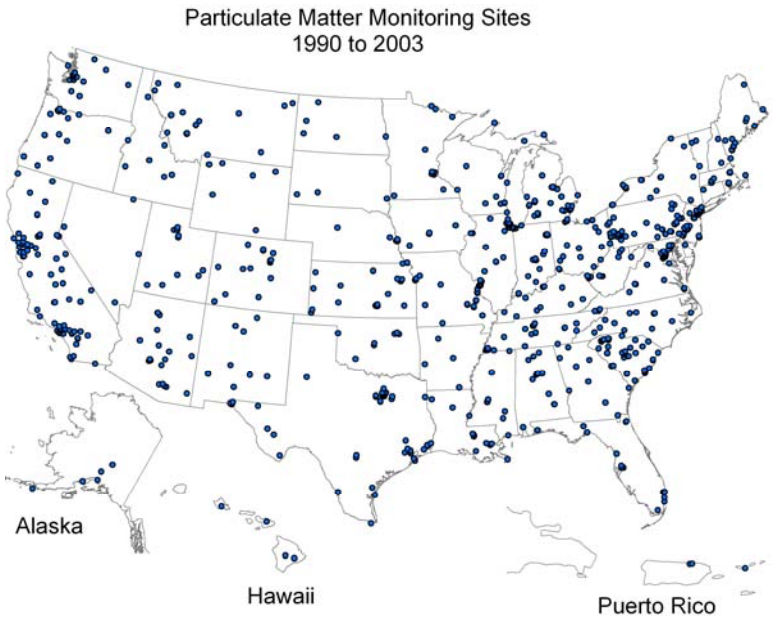
Citywide Averages by Pollutant

Pollutant	Number of citywide averages	At least two site-average cities
1,3-Butadiene	59	26
Acetaldehyde	73	26
Arsenic (tsp)	29	12
Arsenic $PM_{2.5}$	209	48
Benzene	87	34
Carbon tetrachloride	63	28
Chloroform	43	14
Chromium $PM_{2.5}$	210	45
Chromium (tsp)	25	
Formaldehyde	73	26

Pollutant	Number of citywide averages	At least two site-average cities
Lead (tsp)	62	30
Lead $PM_{2.5}$	217	51
Manganese (tsp)	37	14
Manganese $PM_{2.5}$	220	52
Methylene chloride	59	22
Nickel $PM_{2.5}$	67	10
Nickel (tsp)	24	
o-Xylene	74	31
Tetrachloroethylene	55	24
Trichloroethylene	47	17

Spatial Coverage of Monitoring Data





Sites with at least one valid annual average of a VOC or PM metal pollutant between 1990 and 2003.

There is better spatial coverage for PM metals, primarily due to the IMPROVE network (rural sites).

Conclusions (1 of 2)

- Which toxics species are adequately represented in the database for temporal and spatial analysis?
  - 14 gaseous HAPs had sufficient diurnal measurements
  - 35 HAPs had enough measurements to assess seasonal variability
  - 15 HAPs had sufficiently long records to assess annual trends
  - 20 HAPs had sufficient measurements to assess between-city spatial variability
  - 18 HAPs had sufficient measurements to assess within-city spatial variability
  - Some of the HAPs that are predicted to be important for cancer risk are not in the current database

Conclusions (2 of 2)

- How should we treat missing data and data below MDL?
  - Missing data and substituted data below MDL were not used.
- What is our confidence in the data?
  - It depends on what questions we want to answer.
    - MDLs are still too high to have quantitative confidence in many air toxics.
    - For those pollutants with data above the MDL, we can answer many policy-relevant questions.

References

McCarthy M.C., Hafner H.R., and Montzka S.A. (2005) Background concentrations of 18 core air toxics in the northern hemisphere. J. Air & Waste Manage. Assoc. (in press) (STI-903550-2589).

U.S. Environmental Protection Agency (2002) The National Air Toxics Assessment. Technology Transfer Network. Available on the Internet at <http://www.epa.gov/ttn/atw/nata/index.html> web sites and databases (last accessed September 1, 2005).

White W. (2005) IMPROVE network at the Crocker Nuclear Laboratory at U.C. Davis, Davis, CA. Telephone conversation with Michael McCarthy, Sonoma Technology, Inc., Petaluma, CA. July 12.

Acronyms

- AQS = Air Quality System, EPA's data archive
- HAP = Hazardous air pollutant (i.e., air toxics)
- IMPROVE = Interagency Monitoring of Protected Visual Environments
- MDL = Method Detection Limit (sometimes minimum detection limit)
- NATA = National Air Toxics Assessment
- PAH = Polycyclic Aromatic Hydrocarbon
- PM = Particulate matter
- POM = Poly cydic organic matter
- SVOC = Semi-volatile organic compound
- tsp = total suspended particulate
- VOC = Volatile organic compound