DESIGN AND FEASIBILITY ASSESSMENT OF A RETROSPECTIVE EPIDEMIOLOGICAL STUDY OF COAL-FIRED POWER PLANT EMISSIONS IN THE PITTSBURGH PENNSYLVANIA REGION

FINAL SCIENTIFIC REPORT

Reporting Period: March 18, 2005 - December 20, 2006

Richard A. Bilonick

Daniel Connell

Evelyn Talbott

Jeanne Zborowski

Myoung Kim

March 2007

DOE PROJECT NUMBER: DE-FC26-05NT42302

University of Pittsburgh

Pittsburgh, Pennsylvania 15260

CONSOL Energy Research & Development

4000 Brownsville Road

South Park, Pennsylvania 15129-9566

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Abstract

Eighty-nine (89) percent of the electricity supplied in the 35-county Pittsburgh region (comprising parts of the states of Pennsylvania, Ohio, West Virginia, and Maryland) is generated by coal-fired power plants making this an ideal region in which to study the effects of the fine airborne particulates designated as $PM_{2.5}$ emitted by the combustion of coal. This report demonstrates that during the period from 1999-2006 1) sufficient and extensive exposure data, in particular samples of speciated $PM_{2.5}$ components from 1999 to 2003, and including gaseous co-pollutants and weather have been collected, 2) sufficient and extensive mortality, morbidity, and related health outcomes data are readily available, and 3) the relationship between health effects and fine particulates can most likely be satisfactorily characterized using a combination of sophisticated statistical methodologies including latent variable modeling (LVM) and generalized linear autoregressive moving average (GLARMA) time series analysis. This report provides detailed information on the available exposure data and the available health outcomes data for the construction of a comprehensive database suitable for analysis, illustrates the application of various statistical methods to characterize the relationship between health effects and exposure, and provides a road map for conducting the proposed study. In addition, a detailed work plan for conducting the study is provided and includes a list of tasks and an estimated budget. A substantial portion of the total study cost is attributed to the cost of analyzing a large number of archived $PM_{2.5}$ filters. Analysis of a representative sample of the filters supports the reliability of this invaluable but as-yet untapped resource. These filters hold the key to having sufficient data on the components of $PM_{2.5}$ but have a limited shelf life. If the archived filters are not analyzed promptly the important and costly information they contain will be lost.

Project Information

PRINCIPAL AUTHOR	Richard A. Bilonick, PhD University of Pittsburgh School of Medicine
DOE PROJECT NUMBER	DE-FC26-05NT42302
CONTRACTOR	Graduate School of Public Health University of Pittsburgh Pittsburgh, PA 15261
SUBCONTRACTORS	CONSOL Energy Inc. Research & Development South Park, PA 1512 Ohio University Athens, OH 45701
PARTICIPANT	Allegheny County Health Department Pittsburgh, PA 15213

Research Team

University of Pittsburgh School of Medicine	Richard A. Bilonick
University of Pittsburgh Graduate School of Public Health:	Environmental and Occupational Health
	Nancy Sussman
	Larry Keller
	Chris Myers
	Epidemiology
	Evelyn Talbott
	Jeanne Zborowski
	Judith Rager
	Juley Rycheck
CONSOL Energy Inc., Research and Development:	Daniel Connell
	Steve Winter
Ohio University:	Kevin Crist
	Myoung Kim
	Sudhin Devarchetty
	JB Hoy
	Darren Cohen
Allegheny County Health Department:	Samuel Schlosberg

Table of Contents

Index of Tables

Index of Figures

Index of Listings

Executive Summary

The overall goal of the PITT-PM project was to design a feasible retrospective epidemiological study that would satisfactorily characterize and measure the association between airborne fine particulates ($PM_{2.5}$) emitted from coal-fired power plants and human health in the Pittsburgh region. The central element of this proposed study is a times series design requiring the construction of daily time series for health impacts (the responses) and potential explanatory exposure factors for the period from 1999 to 2006 and focusing on the seven-county metropolitan statistical area.

The primary explanatory exposure factors of interest include $PM_{2.5}$ mass and species concentrations, source-related latent factors that explain the observed species concentration correlations along with confounding factors such as weather and co-pollutants. The source-related latent factors would be determined by using a latent variable multivariate receptor model along with block bootstrapping to account for time dependencies. Rather than constructing just a single set of daily time series, it is proposed that as many series as possible be constructed for regions as small as ZIP code areas limited only by the availability of detailed health impact and exposure data and combined together in an overall random effects type model. The proposed design requires the construction of a data base containing health impact data on mortality, morbidity, physician visits, and emergency department data along with data from as many as 70 exposure monitoring sites. In addition, it is essential that existing archived $PM_{2.5}$ filters from monitoring networks be analyzed and included. Where monitors used more than one method of analysis, measurements will be appropriately calibrated and their precision assessed using the method of latent variable modeling. The optimum construction of each exposure time series would be guided by the use of space-time geostatistical methods to allow the proper weighting of multiple monitoring site information distributed both in time and space. Once the series are constructed, a generalized autoregressive moving average model would be constructed for each health outcome (for example, cardiovascular hospital admissions) and the parameters characterizing the relationship between health impact and exposure estimated. The design is similar to a longitudinal mixed effects model where the subjects are ZIP code areas with each ZIP code area having a time series as the repeated measure. Having multiple series available at various spatial locations would allow the overall effects to be determined along with the heterogeneity of the effects over the region. In order to provide realistic confidence intervals for the estimated parameters, spatial bootstrap sampling will be needed to accommodate the expected spatial correlation among the ZIP code area responses.

The majority of the effort in designing this study was devoted to ascertaining and describing what exposure data and health outcomes data are available for the region comprised of the city of Pittsburgh, Pennsylvania, the county of Allegheny and the surrounding 35 counties. Additional effort was expended in determining the optimum time period and most appropriate and innovative statistical methods for characterizing the health impacts. The following report provides the design details, characterizes the available data, and illustrates the statistical methods that would be used.

1 Study Design

1.1 Introduction

The overall goal of the PITT-PM project was to design a feasible retrospective study that would satisfactorily characterize the nature of any association between fine particulates ($PM_{2.5}$) emitted from coal-fired power plants and human health in the Pittsburgh region. Ideally, by statistically controlling for any plausible confounding factors such as weather, various co-pollutants, vehicle traffic, and so forth, support for a causal connection could be produced if it existed. A retrospective study is restricted to using data that likely was collected for purposes other than looking for health impacts from coal-fired power plant emissions. The optimal data set would contain complete, accurate, and precise information for individual people such that their health status, exposure to coal-fired particulates and other confounding factors would be known. Such ideal data was not be expected to be available. Health data is available readily for hospital admissions and emergency department visits, including dates but limited in spatial resolution to ZIP code. Health data concerning physician visits, use of medications, and so forth would be more difficult to obtain. Exposure factors for subjects must be estimated from centrally located ambient monitoring sites. The actual individual exposure will likely differ not only randomly but systematically over time. Additionally, pollution monitoring only measures particulate emissions from coal-fired power plants indirectly thus requiring extra effort to identify and measure the emissions from sources hidden in the monitoring data. The differences between the optimal and what is actually available would likely tend to attenuate any relationship so that it will be important to use the available data carefully with the appropriate statistical methodology to mitigate the attenuation.

In designing this study, it was necessary to identify potential health impact data sets and exposure data sets that cover the city of Pittsburgh, the surrounding Allegheny County, and nearby counties in Pennsylvania, Ohio, Maryland, and West Virginia. A small amount of the available data was collected, examined and used in the exploration of potentially useful statistical methods. The first step in conducting the actual study will be to construct data bases containing the health impact data and the exposure data. The exposure data exists at multiple sites often collected using different analytical techniques when measuring a particular variable. Calibration between different methods will need to be performed to make the measurements comparable. These data bases will be used to construct a data base of daily health impact variables and daily exposure variables for ZIP code subareas because of the limitation in spatial resolution due to the health outcomes data. To construct the daily data base for exposure, space-time geostatistical methods will be used (as demonstrated in this study) to take measurements from multiple spatially distributed monitoring sites over time and construct optimal estimates of daily exposure for each ZIP code area as dictated by the availability of health impact information. To separate and identify emissions due to coal-fired power plants from other sources, latent source factors will need to be identified and measured using a latent variable multivariate receptor model. Finally, the use of generalized autoregressive moving average (GLARMA) time series models would be employed to characterize the relationships between various sets of explanatory factors and a number of health impact variables so as to fully characterize the impact of coal-fired power plant emissions. These results will be compared to results obtained using non-time series models such as generalized additive models (GAM) and generalized linear models (GLM) and also to a case crossover analysis.

When exploring the best possible statistical methodologies to accomplish the tasks discussed above, it was clear that previous studies have not always used the best available methodologies, or even appropriate methodologies, nor have appropriate methodologies always been used correctly. In particular, the calibration of measurements from different techniques to measure the same theoretical construct (e.g., SO_2) using naive regression analysis is known to lead to severe distortions in characterizing analytical bias yet is still routinely used in place of more appropriate calibration methods. Attempts at source apportionment (needed for separating coal-fired emissions impacts from other sources) have often involved the use of exploratory factor analysis even though this method cannot identify sources appropriately and does not, by itself, account for autocorrelation in the time series. Finally, the use of generalized additive models and generalized linear models for times series data do not necessarily account appropriately for the autocorrelation in the response time series resulting in parameters with inflated statistical significance and confidence intervals that are too narrow. Appropriate procedures for diagnosing this problem are rarely employed so that the resulting models do not account for all the systematic effects.

Even in a designed prospective study, not all statistical analysis decisions can (or should) be made before data collection begins - although the major analysis questions should be delineated. Given that this is a retrospective study, without the actual data in hand, not even all study design questions can be addressed definitively - additional analysis will be needed to make final decisions on appropriate models and analysis after the complete data base is assembled.

1.2 Proposed Design

The basic design proposed in the original cooperative agreement was a time series design that would model health outcomes (typically counts) as a function of explanatory exposure variables. A diagram for the proposed study is shown in **Figure 1.** The PITT-PM study would model health impacts for the seven-county region over the time period from 1999 to 2006. However, due to limitations in available data as described below, not all analyses will cover this entire time period. On the other hand, exposure data from the entire 35-county region will be used to inform exposure estimates over the seven-county region. The smallest unit of time available for most health outcome data is a day (24 hours) while the spatial resolution is limited to at best ZIP code area. This requires exposure data from various multiple-monitor networks for each measured parameter to be averaged to create daily time series as discussed below. A time series power analysis was performed and indicated that at least three years of daily data would be needed in order to reasonably characterize the relationship between $PM_{2.5}$ and health outcomes. The design for the

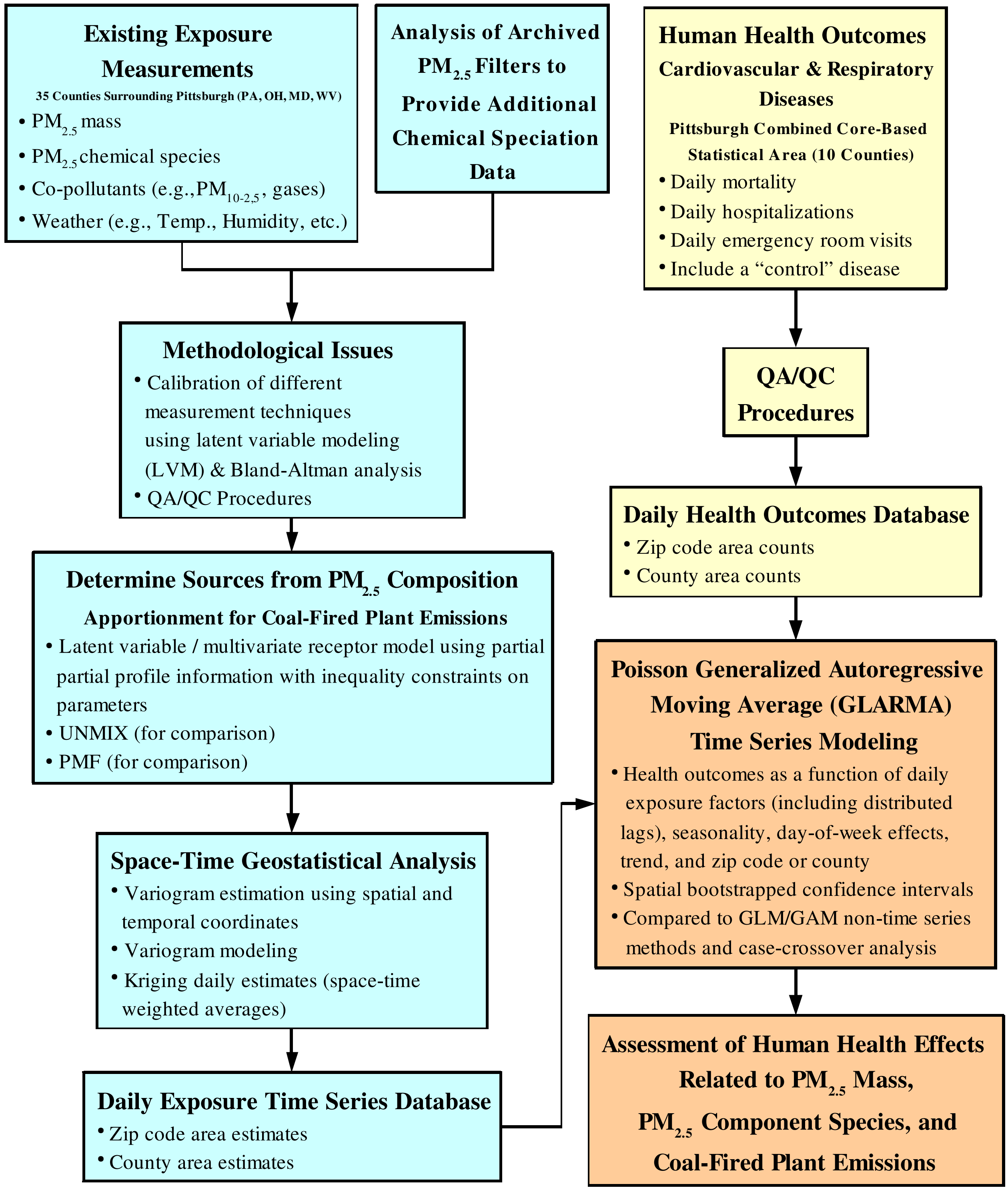

Existing Exposure Measurements
35 Counties Surrounding Pittsburgh (PA, OH, MD, WV)
• $PM_{2.5}$ mass
• $PM_{2.5}$ chemical species
• Co-pollutants (e.g.,$PM_{10\text{-}2,5}$, gases)
• Weather (e.g., Temp., Humidity, etc.)
Analysis of Archived $PM_{2.5}$ Filters to Provide Additional Chemical Speciation Data
Human Health Outcomes
Cardiovascular & Respiratory Diseases
Pittsburgh Combined Core-Based Statistical Area (10 Counties)
• Daily mortality
• Daily hospitalizations
• Daily emergency room visits
• Include a "control" disease
Methodological Issues
• Calibration of different measurement techniques using latent variable modeling (LVM) & Bland-Altman analysis
• QA/QC Procedures
QA/QC Procedures
Daily Health Outcomes Database
• Zip code area counts
• County area counts
Determine Sources from $PM_{2.5}$ Composition
Apportionment for Coal-Fired Plant Emissions
• Latent variable / multivariate receptor model using partial partial profile information with inequality constraints on parameters
• UNMIX (for comparison)
• PMF (for comparison)
Poisson Generalized Autoregressive Moving Average (GLARMA) Time Series Modeling
• Health outcomes as a function of daily exposure factors (including distributed lags), seasonality, day-of-week effects, trend, and zip code or county
• Spatial bootstrapped confidence intervals
• Compared to GLM/GAM non-time series methods and case-crossover analysis
Space-Time Geostatistical Analysis
• Variogram estimation using spatial and temporal coordinates
• Variogram modeling
• Kriging daily estimates (space-time weighted averages)
Assessment of Human Health Effects Related to $PM_{2.5}$ Mass, $PM_{2.5}$ Component Species, and Coal-Fired Plant Emissions
Daily Exposure Time Series Database
• Zip code area estimates
• County area estimates

Figure 1: Overview of the proposed PITT-PM retrospective epidemiological study design for the Pittsburgh region from 1999 to 2006.

study is described in the following subsections. The technical details and supporting information for the proposed design are shown below in **Sections 2–4**. An overall work plan which includes a list of tasks to be completed and an estimated budget by task and year is shown in **Section 5**.

1.2.1 Existing Exposure Measurements

There are nine significant sources of $PM_{2.5}$ mass, $PM_{2.5}$ speciation, co-pollutant, and meteorological data for the 35-county (PA, OH, WV, MD) region surrounding Pittsburgh during the period 1999 to 2005 :

- U.S. Environmental Protection Agency's Air Quality System (AQS)
- National Energy Technology Laboratory Office of Science and Technology (NETL/OST)
- Pittsburgh Air Quality Study (PAQS)
- Upper Ohio River Valley Project (UORVP)
- Steubenville Comprehensive Air Monitoring Program (SCAMP)
- Clean Air Status and Trends Network (CASTNet)
- Interagency Monitoring of Protected Visual Environments (IMPROVE)
- Federal Aviation Administration Automated Surface Observing System (ASOS) / Automated Weather Observing System (AWOS)
- Roadway Weather Information System (RWIS)

Exposure data from 2006 was not generally available at the time of investigation but most of the findings would likely extend directly to data from 2006. Data was obtained from each of these sources and inventories were performed for each exposure parameter of interest. Because of the large quantity of $PM_{2.5}$ total mass, PM_{10} total mass, gaseous pollutant (i.e., SO_2, NO_2, CO, and O_3), and meteorological data that were collected in the 35-county region between 1999 and 2005, it was not practical to perform a day-by-day inventory of these data and instead the data were reviewed to determine the time period, frequency, time resolution, and method of collection for each parameter at each monitoring site. There were 47 monitoring sites that measured ambient $PM_{2.5}$ mass concentrations during some or all of the period from 2000 to 2005. (Sixteen sites in Allegheny County, six additional sites in the Pittsburgh Metropolitan Statistical Area (MSA) comprising Allegheny, Beaver, Butler, Fayette, Washington, and Westmoreland counties, and one additional site in Armstrong County in western Pennsylvania collected $PM_{2.5}$ data during 1999 but a site-by-site inventory was not made for this year.)

There were fifteen sites that monitored for a complete suite of $PM_{2.5}$ chemical components, including ions, carbon, and trace and crustal elements, during some or all of the period from 1999 to 2005. The site with the greatest number of days of existing, complete $PM_{2.5}$ speciation data is the Lawrenceville site in the City of Pittsburgh.

Co-pollutant data of interest include daily ambient concentrations of SO_2, $O_{3,}$ CO, and NO_3. A total of sixty-four sites measured some or all of these gaseous species between 1999 and 2005. A site-by-site inventory was performed for the years from 2000 to 2005. All of the gaseous pollutant measurements were made using continuous monitors resulting in data with an hourly or finer resolution that can be used to compute daily averages or other statistics such as the maximum 1-hour average concentration.

Although numerous sites in the 35-county region collected temperature data, relative humidity or dew point data, and wind speed and direction data between 1999 and 2005, the thirteen ASOS/AWOS sites located at airports throughout the region are probably the best source of meteorological data.

1.2.2 Analysis of Archived $PM_{2.5}$ Filters

Many of the sites that determined ambient air concentrations of $PM_{2.5}$ chemical species between 1999 and 2005 also collected additional filter-based $PM_{2.5}$ samples that were not analyzed for chemical composition. Chemical analysis of these filters, where feasible, would substantially augment the existing speciated $PM_{2.5}$ data record. The feasibility of obtaining $PM_{2.5}$ chemical composition data from these archived filters depends on the method originally used to sample the particles, the type of filter, and the manner in which the samples were stored, among other things. Although there are important limitations (such as the underestimation of semi-volatile components and the inability to assess elemental and organic carbon from Teflon-filter-based samples using thermal/optical techniques), this data represents a unique and extremely valuable resource that will disappear if not taken advantage of during the next few years. Eight monitoring sites located in Allegheny, Greene, Washington, and Westmoreland counties have substantial available inventories of archived filters consisting of more than 8,400 samples.

1.2.3 Calibration and QA/QC for Exposure Data

In order to assemble the prerequisite data bases of daily time series for the exposure factors some methodological issues need to be addressed. The networks used to measure these factors consist of individual monitoring sites at specific locations and operating during certain periods of time. Before the geostatistical analysis and modeling can begin, it will be necessary to equilibrate measurements made by equipment using different techniques to measure the same parameters. Thus it will be necessary to determine the amount of agreement between sets of measurements so that they can be adjusted for any bias (systematic error) and that the relative imprecision (random error)of each technique can be assessed. Many researchers have, unfortunately, attempted to use linear regression and correlation analysis naively to address this problem. The error in this approach has been pointed out by numerous researchers including Bland and Altman, and Ripley. [need references here] We propose using latent variable modeling to examine the nature of any relative bias and imprecision and determine the appropriate adjustment to place all measurements on the same footing.

1.2.4 Source Apportionment

In principle, the $PM_{2.5}$ chemical species data at the monitoring sites indirectly provide information about the sources of the ambient fine particles. Theoretically, the relationship between the sources (particle emitters) and receptors (the monitoring sites) could be represented by a chemical mass balance. If enough source profile information were available, the mass balance equations relating the observed species concentrations to the emitting sources could be solved using regression analysis. Typically, the required extensive source profile information is not readily available so that the source profiles also need to be estimated. Using a very limited number of source profiles, a latent variable multivariate receptor model can be constructed and the sources identified. Block bootstrap sampling then can be used to account for the time dependency in the component specie time series to provide realistic confidence intervals for the parameters as an aid the identification and characterization of sources. The resulting particulate emissions source contributions can be used to assess the relationship between a coal-fired power generation plant source and health outcomes. This methodology was applied to 23 $PM_{2.5}$ components from the Lawrenceville (Pittsburgh) site to demonstrate the approach for a five source model based on previous research by the Allegheny County Health Department.

1.2.5 Space-Time Geostatistical Analysis and Modeling

To produce optimal daily averages requires the appropriate weighting which can be determined by a space-time geostatistical analysis for each parameter. Each daily time series value for a given parameter would be an optimally weighted average of measured values distributed in both space and time. Typically, measured values closer in space and time to the area being estimated should have larger weights than values farther away. The arrangement of the measurements and the shape and size of the region being estimated will also affect the weights. For example, two measurements close together in space, or time, or both will tend to have less total weight than two measurements that are far apart. Computationally, areas are represented numerically by sets of points on a regular grid with each grid point being individually estimated and the average of the grid points taken as the the value for the region. The optimal weights would be determined by using the method of kriging. Kriging requires information about the correlation among the monitoring site locations and between the monitoring sites and the area being estimated. The correlation information comes from a model fitted to the estimated space-time variogram for a given parameter.

1.2.6 Daily Exposure Time Series Database

To facilitate the time series and other modeling, the raw exposure data must be assembled into a daily exposure data base. To minimize exposure misclassification and to match up with the health outcome data, daily values will be estimated by ZIP code area. County estimates can then be constructed from the ZIP code area estimates.

1.2.7 Human Health Outcomes

Based on the completed comprehensive inventory and assessment of available mortality and morbidity datasets, the identified sources are:

- National Center for Health Statisics (NCHS) Division of Vital Statistics
- Pennsylvania Department of Health Bureau of Health Statistics and Research
- Allegheny County (PA) Health Department
- Pennsylvania Health Care Cost Containment Council Hospital Discharge Data Sets (1999-2006)
- Ohio Department of Health
- West Virginia Healthcare Authority Hospital Discharge Datasets
- Emergency Department Visit Data (from individual hospitals)
- UPMC Medical Archival Retrieval System (MARS)
- Real-Time Outbreak and Desease Surveillance (RODS) Data

1.2.8 QA/QC for Health Data

1.2.8.1 Accuracy and Verification of Health Data

As previously described, this proposal for the retrospective epidemiological assessment of the health effects related to $PM_{2.5}$ and its component species will make use of existing secondary data on mortality, hospitalizations and ED visits within the ten county area (or larger) of study. All health datasets to be used will be obtained primarily at the onset of the project period.

Mortality data will be obtained from the Pennsylvania Department of Health Bureau of Health Statistics and Research and verified using National Center for Health Statistics (NCHS) Division of Vital Statistics. Recent quality analysis comparing these electronic datasets to death certificates suggests that the error rate is 2% or less. Hospitalization data is collected by the Pennsylvania Health Care Cost Containment Council (PHC4). The data are processed using a series of validation rules before being finalized and made available for further analysis and public release. PHC4 edits the data and provides error reports to each data source. The health care facility will make error corrections and provide PHC4 with corrected information. Compliance across health care institutions in Pennsylvania approaches 100% (99% in recently released 2006 reports). Emergency department (ED) data will be acquired from individual hospitals/hospital systems through directed agreements. If necessary, the investigators will utilize an "honest broker" system to acquire identified ED data from hospitals for use in the study. Verification of the accuracy and integrity of the ED and other data will be conducted by the data research associate and will include ID verification, ICD 9/10 verification and reconciliation, data range, and type verification, and duplicate entry checks. Additional data editing and report generation will be performed quarterly to assure data integrity and completeness. Any data and data collection ambiguities will be brought to the attention of the study principal investigator by the data research associate for immediate resolution.

1.2.8.2 Privacy and Confidentiality of Health Data

These health-based datasets will likely contain identifiable subject information and the release of said information is governed by the privacy and confidentiality regulations of Health Insurance Portability and Accountability Act (HIPAA) of 1996. The release of datasets containing health outcomes information at the individual level to external agencies or researchers is governed by the protected access procedures of the participating agencies. HIPAA's Privacy Rule makes provisions for a "limited data set," authorized only for public health, research, and health care operations purposes. A limited data set must have all direct identifiers removed, including:

- Name and social security number;
- Street address, e-mail address, telephone and fax numbers;
- Certificate/license numbers;
- Vehicle identifiers and serial numbers;
- URLs and IP addresses;
- Full face photos and any other comparable images;
- Medical record numbers, health plan beneficiary numbers, and other account numbers;
- Device identifiers and serial numbers; and
- Biometric identifiers, including finger and voice prints.

- A limited data set could include the following (potentially identifying) information:
- Admission, discharge, and service dates;
- Dates of birth and, if applicable, death;
- Age (including age 90 or over); and
- Five-digit ZIP code or any other geographic subdivision, such as state, county, city, precinct and their equivalent geocodes (except street address).

Covered entities such as the Pennsylvania Health Care Cost Containment Council (PHC4), and individual hospitals must condition the disclosure of the limited data set on execution of a "data use agreement.” The Pennsylvania Department of Health also requires the execution of a data use agreement for access to mortality datasets. This agreement: 1) Establishes the permitted uses and disclosures of such information by the recipient, consistent with the purposes of research, public health, or health care operations; limits who can use or receive the data; and 2) Requires the recipient to agree not to re-identify the data or contact the individuals. In addition, the data use agreement must contain adequate assurances that the recipient will use appropriate physical, technical and administrative safeguards to prevent use or disclosure of the limited data set other than as permitted by HIPAA and the data use agreement, or as required by law. Alternatively, if a covered entity becomes aware of a violation of the data use agreement, it must take reasonable steps to remedy the problem or, if unsuccessful, discontinue disclosure of PHI to the recipient and report the problem to DHHS.

The minimum necessary standard governs covered entities' disclosures, and recipients' uses, of limited data sets. The covered entity may place reasonable reliance that a requested disclosure is indeed the minimum necessary for the stated purposes, or make its own determination that a lesser amount of information would be sufficient.

All health related records and information pertaining to the involvement of human subjects in the proposed retrospective research study will be kept strictly confidential and housed on password secured computers and/or in locked file cabinets at the University of Pittsburgh School of Medicine and Graduate School of Public Health. Only the study investigators and designated staff will have access to the subject records and data. Any subject names and/or other personal identifiers will be removed from data forms or electronic files prior to record filing, database preparation and analysis. Individual subject records will be identified only by a unique study code to ensure subject confidentiality.

1.2.9 Daily Health Outcomes Database

Individual level health outcomes data for the period from 1999-2006 will obtained from the various health data collection entities/agencies and assembled initially into a series of separate datasets based on the outcome of interest, broadly classified as deaths, hospitalizations, ED visits and potentially physicians' office visits. Daily counts of the health outcomes of interest (deaths, hospitalizations, ED visits and potentially physicians' office visits) will be determined from the individual health records and aggregated at the county and ZIP Code levels. International Classification of Diseases, Revision 9 [ICD-9] and Revision 10 (ICD-10) codes will be available in the datasets and will be used to categorize cardiovascular and respiratory/pulmonary outcomes for sub-analyses of specific diseases. Also available for possible examination and validation is a "composite daily health effects" count variable that would be used to quantify, for any given day, all cardiovascular and/or respiratory deaths, hospitalizations, ED visits and potentially physicians' office visits for time series analyses.

1.2.10 Poisson Generalized Autoregressive Moving Average Time Series Modeling

To model the relationships between $PM_{2.5}$ mass, constituent species, and latent source factors on the one hand and health outcomes on the other, the use of generalized autoregressive moving average (GLARMA) time series models are proposed. This type of model, as with any "generalized" approach allows the modeling of the observed daily counts as the discrete outcome from a Poission random variable. The GLARMA model, however, goes further and allows modeling of the autocorrelation in the error that tends to exist even after accounting for all measured covariates – something which is ignored by the use of generalized linear models (GLMs) and generalized additive models (GAMs) and its omission tends to substantially overestimate the statistical significance of the estimated parameters.

The region to be studied should include Allegheny County and the surrounding nine counties although data from the larger 35-county region would be used to improve the daily kriged estimates of exposure. The power analyses indicate that the minimum time period would be three years of daily observed counts and corresponding exposure factors. Rather than trying to produce one daily time series (for each health outcome and exposure factor) for a single large region, it is proposed to create separate time series for each ZIP code area and to treat in a conceptual sense the ZIP code areas as if they were "subjects" in a longitudinal mixed effects analysis. In this approach, each ZIP code area has its own trajectory over time and the goal is to include all ZIP code areas in the estimation of the overall relationship between $PM_{2.5}$ and each particular health outcome count. The benefits of this approach are that there is less opportunity to average out the relationship between $PM_{2.5}$ and health outcome count due to spatial heterogeneity in the timing of changes in exposure and health outcome among different areas and at the same time the nature of any spatial heterogeneity in the relationship between $PM_{2.5}$ and health outcome could be examined. The power to detect the overall (fixed effect) relationship will be enhanced to the extent there is less heterogeneity in the time series regression (random effect) coefficients for each ZIP code area.

The overall design is conceptually similar to a longitudinal repeated measures design where each "subject" or experimental unit is a ZIP code area and the repeated measures are a time series of health outcome counts with the inclusion of a number of measured covariates to account for confounding factors.1 The greater the heterogeneity of the ZIP code areas, the wide the confidence intervals on the model parameters describing the relationship between the explanatory factors and the health outcome counts.

In a typical longitudinal study, the subjects are typically assumed to be statistically independent. The analysis using ZIP code areas as the experimental units, however, is complicated by the spatial correlation likely to exist among the ZIP code areas. The daily health outcome counts of two ZIP code areas are likely to be correlated (contradicting the usual assumption of statistical independence among the experimental units) and this correlation is likely to be a function of the average distance between the areas – closer ZIP code areas are likely to be more highly positively correlated than areas far apart. Thus there is no easy way to account for this spatial correlation which will impact the statistical significance of estimated parameters and the widths of the corresponding confidence intervals. To address this problem we propose to use a spatial bootstrap sampling procedure to incorporate the spatial autocorrelation and make it possible to assess statistical significance and to produce realistic confidence intervals. The spatial bootstrap is similar in concept to the block bootstrap which handles temporal correlation and is proposed for use with the latent variable multivariate receptor modeling.

The spatial bootstrapped GLARMA analysis can be compared to generalized additive models (GAMs) and generalized linear models (GLMs) which are non-time-series methods and to a case-crossover analysis. In the GAM/GLM approach, the autocorrelated structure is not necessarily accommodated unless the explanatory factors completely account for all the effects and the resulting residuals are statistically independent – a highly unusual and unlikely outcome. Typically, the model consisting of all available measured exposure factors does not produce temporally uncorrelated residuals. Spatial bootstrapping could be applied to GAM/GLM models but that would not correct for the temporal correlation. Case-crossover designs have the potential for better control of subject heterogeneity given that each subject is his/her own control. Each event (e.g., an outcome that results in hospitalization of the subject) is a “case” and the exposure factors at the time of the event are matched to a “control” for the same subject which occurs at a time point either before, or after, or time points both before and after, the event. Spatial bootstrap sampling can be applied to the case-crossover analysis to account for spatial dependency between ZIP code areas.

1.2.11 Assessment of Human Health Effects

The analyses described above allow complex models to be fitted to the observed data that characterize the relationship between the explanatory exposure factors and the various human health outcome counts. Having controlled for many if not all plausible confounding factors, the resulting model parameter coefficients for $PM_{2.5}$ mass, $PM_{2.5}$ components, and latent coal-fired particulate emission factors should reflect the impact of fine particulates on human health. The spatial bootstrap sampling procedure should produce realistic uncertainty bounds for each estimated parameter in the fitted models.

2 Exposure Data Assessment

2.1 Introduction

The design and feasibility of a retrospective epidemiological study of $PM_{2.5}$ emitted by coal-fired power plants in the Pittsburgh, Pennsylvania, region depend strongly upon the quantity and quality of air monitoring data that were collected in the region during the time period of interest.

In order to ascertain the health effects of ambient fine particles originating from coal-fired power plants, several types of air monitoring data are required. Most importantly, detailed $PM_{2.5}$ speciation data must be available, as these data provide a means for differentiating between $PM_{2.5}$ derived from coal-fired power plants and $PM_{2.5}$ derived from other types of sources. In addition to $PM_{2.5}$ mass concentrations and basic compositional data, which include concentrations of major ionic and carbonaceous $PM_{2.5}$ components such as sulfate (SO_4^{2-}), nitrate (NO_3^-), elemental carbon (EC), and organic carbon (OC), the concentrations of trace element species must be known. These elements are essential for use as tracers in source apportionment; for example, Se is commonly used as a marker of primary emissions from coal combustion (Suarez and Ondov, 2002). Moreover, in spite of their very small ambient air concentrations, trace metals may have implications for public health. A number of trace metal species (i.e., Sb, As, Be, Cd, Cr, Co, Pb, Mn, Hg, Ni, and Se) are classified as Hazardous Air Pollutants (HAPs) in the 1990 Clean Air Act Amendments, and toxicological evidence suggests that certain transition metals (e.g., Fe, Zn, Cu, V) in particulate matter that are not listed as HAPs may nevertheless elicit adverse health responses (e.g., Carter et al., 1997; Zelikoff et al., 2002; and Adamson et al., 2000).

Concentrations of gaseous pollutants, including carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ozone (O_3), must also be available. The U.S. Environmental Protection Agency (EPA) has set primary National Ambient Air Quality Standards (NAAQS) for all of these gases because of their potential to adversely affect public health. Moolgavkar and Luebeck (1996) and Lipfert and Wyzga (1997) criticized several early particulate matter epidemiology studies (e.g., Schwartz and Dockery, 1992; Schwartz et al., 1996) for not adequately considering the potential confounding effects of gases such as CO and NO_2. Several more recent studies (Moolgavkar, 2003; Villeneuve et al., 2003) that considered both particulate matter and gaseous pollutants found that gases were more strongly associated with mortality than was particulate matter. In Pittsburgh, Chock et al. (2000) reported that although PM_{10} was significantly associated with daily mortality among those less than 75 years of age in non-seasonal single- and multi-pollutant models, the use of seasonal models revealed collinearity problems among concentrations of PM_{10}, CO, NO_2, and O_3 (spring and summer), casting doubt upon the findings of the non-seasonal models. Hence, a $PM_{2.5}$ epidemiology study in the Pittsburgh region must consider the potential for the confounding effects of gaseous pollutants and seasonality. Potential health effects of $PM_{10-2.5}$, the coarse particle fraction that along with $PM_{2.5}$ constitutes PM_{10}, should be considered as well, because coarse particles have also been epidemiologically associated with mortality (e.g., Ostro et al., 2000; Mar et al., 2000).

Finally, meteorological data must be available. The epidemiological models must account for the effects of variables such as temperature and relative humidity. Knowledge of wind speed and direction may also aid source apportionment or geostatistical modeling.

Thus, the feasibility of performing a retrospective epidemiology study of $PM_{2.5}$ and its components in the Pittsburgh region depends on the ability, at a minimum, to assemble a nearly continuous stream of daily average values of all of these variables such that: (1) the values provide a representative estimate of the exposure of the population being considered, and (2) the data stream is contiguous enough to provide, in conjunction with the size of the population being considered, sufficient results so that the advanced statistical techniques employed can produce reliable and valid conclusions.

A number of air monitoring campaigns collected $PM_{2.5}$ mass, $PM_{2.5}$ speciation, co-pollutant, and meteorological data in Pittsburgh and surrounding areas between 1999 and 2005, and several of these campaigns have succeeded in applying source apportionment methodologies to $PM_{2.5}$ speciation data in order to resolve time series representing the probable contribution of coal-fired power plants to ambient $PM_{2.5}$ in the region (e.g., Pekney et al., 2005; Maranche, 2006; Connell et al., 2006; Martello et al., 2006). However, these air monitoring campaigns generally were not designed with the intent of providing data for an epidemiology study. As a result, it is unlikely that any single air monitoring site in the Pittsburgh region collected a sufficient quantity of data to permit such a study to be performed. It is likely, though, that information from the numerous monitoring sites that operated in the Pittsburgh region during 1999-2005 can be combined to provide daily estimates of the region's exposure to $PM_{2.5}$, $PM_{2.5}$ components, co-pollutants, and pertinent meteorological parameters over a sufficiently long period for a time series epidemiology study focusing on the effects of these variables. Compared with exposure monitoring for a prospective epidemiology study, in which the monitoring site locations, sampling schedule, sampling and analytical methodologies, and quality control procedures are designed specifically to meet the study's needs, the exposure data for the proposed retrospective study in Pittsburgh must be derived from monitoring activities that have already been conducted. As such, there are a number of challenges associated with merging these data into a coherent exposure database for use in the study, which arise because the data from the various sites were collected during different time periods, at different frequencies and time resolutions, and using different measurement techniques. If these challenges can be overcome, however, the cost and time required for performing a retrospective study using the Pittsburgh region's valuable, expansive set of existing air monitoring data is expected to be substantially less than the cost and time required for performing a prospective study with its associated sampling and analytical requirements.

Hence, a major goal of the current study was to determine whether there is a sufficient quantity and quality of air monitoring data available for the Pittsburgh region from 1999-2005 to permit a retrospective epidemiologic study of $PM_{2.5}$ resulting from coal-fired power plant emissions, and if so, to develop a plan for using these data in such a study. This goal was accomplished by completing a series of four subtasks, as follows:

1. Inventory existing $PM_{2.5}$ (mass and speciation), co-pollutant, and meteorological data that are available from the Pittsburgh region during the time period of interest

2. Inventory archived filters that might be analyzed to augment the speciated $PM_{2.5}$ data record for the Pittsburgh region during the time period of interest

3. Assess the quality and comparability of the available air monitoring data

4. Develop a plan for the construction of an air monitoring database for use in a retrospective epidemiologic study of $PM_{2.5}$ and its components

The results of these subtasks are discussed in the subsections below.

2.2 Inventory of Existing Air Monitoring Data

As discussed above, the design of a retrospective time series epidemiologic study of ambient $PM_{2.5}$ and its components in the Pittsburgh region is constrained by the availability of air monitoring data from the region, which are needed in the time series model to serve as surrogates for the daily exposures of the region's population to $PM_{2.5}$ from coal-fired power plants, $PM_{2.5}$ from other sources, co-pollutants, and various other potential confounding factors (e.g., temperature and humidity). Hence, prior to designing the study, a comprehensive inventory of existing air monitoring data available from the Pittsburgh region during the time period of interest was completed.

Because of the retrospective nature of the proposed epidemiology study, the study region is defined largely by the availability of existing air monitoring and health outcomes data. (This is in contrast to a prospective study, in which the data collection strategy would likely be tailored to a pre-defined region of interest). Hence, all monitoring sites located in a relatively large 35-county region surrounding Pittsburgh were considered as part of the air monitoring data inventory. The counties constituting this region are listed in **Table 1**. Although the final study design may focus on a smaller area, monitoring data from this larger region will nevertheless be useful for assessing the spatial variability of pollutants and informing geostatistical models used to compute exposure estimates.

Table 1: Counties considered in air monitoring data inventory.

State	County	State	County
MD	Garrett	PA	Greene
OH	Belmont	PA	Indiana
OH	Carroll	PA	Jefferson

State	County
OH	Columbiana
OH	Guernsey
OH	Harrison
OH	Jefferson
OH	Mahoning
OH	Monroe
OH	Noble
OH	Trumbull
PA	Allegheny
PA	Armstrong
PA	Beaver
PA	Butler
PA	Cambria
PA	Clarion
PA	Fayette
PA	Lawrence
PA	Mercer
PA	Somerset
PA	Venango
PA	Washington
PA	Westmoreland
WV	Brooke
WV	Hancock
WV	Marion
WV	Marshall
WV	Monongalia
WV	Ohio
WV	Preston
WV	Wetzel

There are a number of sources of $PM_{2.5}$ mass, $PM_{2.5}$ speciation, co-pollutant, and meteorological data from this 35-county region between 1999 and 2005, as follows:

- **The U.S. Environmental Protection Agency's Air Quality System (AQS)**
  AQS includes numerous monitoring sites located throughout the region that sampled for some or all of the parameters of interest between 1999 and 2005. The AQS sites in Allegheny County, Pennsylvania are operated by the Allegheny County Health Department (ACHD); those in the western Pennsylvania counties other than Allegheny are operated by the Pennsylvania Department of Environmental Protection (PA DEP); those in eastern Ohio are operated by the Ohio Environmental Protection Agency (Ohio EPA) or Mahoning-Trumbull Air Pollution Control Agency, and those in northwestern West Virginia are operated by the West Virginia Department of Environmental Protection (WV DEP). AQS data were obtained from the U.S. EPA's Technology Transfer Network (http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm) for purposes of this inventory.

- **The National Energy Technology Laboratory Office of Science and Technology (NETL/OST) Monitoring Site**
  The NETL/OST site was situated on the U.S. Department of Energy (DOE) National Energy Technology Laboratory (NETL) campus in Bruceton, which is located in a suburban area of southern Allegheny County, Pennsylvania. The site operated from July 1999 through September 2004, and it measured $PM_{2.5}$ mass concentrations, gaseous pollutant concentrations, and meteorological conditions during much or all of this period, while also including various measurements of $PM_{2.5}$ chemical components and intermittent measurements of PM_{10} mass concentrations. The site was operated by

DOE-NETL/OST; data collected there were obtained from Don Martello of DOE-NETL for use in the inventory presented here.

- **The Pittsburgh Air Quality Study (PAQS)**
  PAQS included extensive air monitoring between May 2001 and September 2002 at a "supersite" in Schenley Park, which is located in the Oakland section of the City of Pittsburgh. $PM_{2.5}$ mass concentrations, $PM_{2.5}$ chemical composition, gaseous pollutant concentrations, PM_{10} mass concentrations, and meteorological conditions were monitored at the site during most or all of this period. PAQS was conducted by Carnegie Mellon University, with funding provided by the U.S. DOE and U.S. EPA. Data collected as part of PAQS were obtained from the DOE-NETL Air Quality Database project (http://www.pmdata.org) or from Allen Robinson, one of the PAQS program's principal investigators.

- **The Upper Ohio River Valley Project (UORVP)**
  UORVP included three monitoring sites that were collocated with AQS sites in the Lawrenceville section of Pittsburgh, in Holbrook, Greene County, Pennsylvania, and in Morgantown, Monongalia County, West Virginia. The Lawrenceville and Holbrook sites included intermittent filter-based measurements of $PM_{2.5}$ mass, $PM_{2.5}$ chemical composition, and PM_{10} mass between February 1999 and January 2002, and the Lawrenceville site also featured daily $PM_{2.5}$ mass and speciation sampling from October 2002 through February 2003. $PM_{2.5}$, PM_{10} (Lawrenceville only), gaseous pollutants, and meteorological conditions were also monitored continuously at the sites during the study period. The Morgantown site included a limited amount of $PM_{2.5}$ mass sampling between 1999 and 2001 to supplement the sampling being conducted there by the WV DEP. UORVP was conducted by Advanced Technology Systems, Inc. (ATS) under an award from the U.S. DOE; data from the program were obtained from the DOE-NETL Air Quality Database project (http://www.pmdata.org).

- **The Steubenville Comprehensive Air Monitoring Program (SCAMP)**
  SCAMP included five monitoring sites located in Ohio, West Virginia, and Pennsylvania that operated between May 2000 and May 2002. The central monitoring site on the campus of Franciscan University of Steubenville in Steubenville, Ohio, measured $PM_{2.5}$ mass, $PM_{2.5}$ composition, PM_{10} and gaseous pollutant concentrations, and meteorological conditions. The four satellite sites, which were located in Wheeling, West Virginia, Tomlinson Run State Park, West Virginia, Hopedale, Ohio, and Latrobe, Pennsylvania, measured $PM_{2.5}$ mass and certain $PM_{2.5}$ species. The SCAMP ambient air monitoring program was conducted by CONSOL Energy Inc. Research & Development under a cooperative agreement with the U.S. DOE; the SCAMP data used in this inventory were obtained from CONSOL's databases from the project.

- **The Clean Air Status and Trends Network (CASTNet)**
  Two sites from the U.S. EPA's CASTNet that are located in the 35-county greater Pittsburgh region collected $PM_{2.5}$ mass and chemical speciation data between March 1999 and May 2001. These are the

M.K. Goddard site in Mercer County, Pennsylvania, and the Quaker City site in Noble County, Ohio. Both of these sites, as well as the CASTNet's Laurel Hill site, also collected O_3 and meteorological data throughout the time period of interest. CASTNet data were obtained from the Interagency Monitoring of Protected Visual Environments (IMPROVE) online database (http://vista.cira.colostate.edu/improve) and from the CASTNet online database (http://epa.gov/castnet/).

- **Interagency Monitoring of Protected Visual Environments (IMPROVE)**
  Four IMPROVE sites located in the 35-county greater Pittsburgh region collected $PM_{2.5}$ mass concentration, $PM_{2.5}$ chemical composition, and PM_{10} mass concentration data during the time period of interest. These are the M.K. Goddard and Quaker City sites, which began sampling in the spring of 2001 when CASTNet $PM_{2.5}$ sampling was discontinued, the Pittsburgh site, which is collocated with the AQS site at Lawrenceville and began sampling in April 2004, and the Frostburg site, which is located in Garrett County, Maryland, and likewise began sampling in April 2004. IMPROVE data were obtained from the IMPROVE on-line database (http://vista.cira.colostate.edu/improve).

- **Federal Aviation Administration Automated Surface Observing System (ASOS) / Automated Weather Observing System (AWOS)**
  Meteorological data from the time period of interest are available from ASOS/AWOS stations located at airports throughout the 35-county greater Pittsburgh region. These data can be obtained from the National Climatic Data Center (NCDC) (http://www.ncdc.noaa.gov/oa/climate/climatedata.html) or from the Pennsylvania MESONET (http://pasc.met.psu.edu/MESONET/archive/alldatainv.html).

- **Roadway Weather Information System (RWIS)**
  Meteorological data were also collected by the Pennsylvania Department of Transportation's (PennDOT's) RWIS. Data from late 2001 to the present are available from the Pennsylvania MESONET (http://pasc.met.psu.edu/MESONET/archive/alldatainv.html).

Data from each of these sources were obtained as indicated above, and inventories were performed for each parameter of interest. The inventories were generally conducted in accordance with the checklist that is included in **Appendix A** to this report. Because of the large quantity of $PM_{2.5}$ total mass, PM_{10} total mass, gaseous pollutant (i.e., SO_2, NO_2, CO, O_3), and meteorological data that were collected in the 35-county region between 1999 and 2005, it was not practical to perform a day-by-day inventory of these data. Rather, the data were reviewed to determine the time period, frequency, time resolution, and method of collection for each parameter at each monitoring site. Any prolonged periods of missing or invalid data were also noted.

$PM_{2.5}$ chemical speciation data from the greater Pittsburgh region are much less abundant than $PM_{2.5}$ and PM_{10} total mass dta and gaseous pollutant data during the 1999-2005 time period because of the cost and level of effort associated with determining $PM_{2.5}$ speciation, and because collection of these data is not required to assess compliance with a NAAQS. Hence, because the design and feasibility of the proposed epidemiology study depend strongly upon the availability of these speciation data, a day-by-day inventory was performed for each $PM_{2.5}$ chemical constituent at each monitoring site in the 35-county greater Pittsburgh region in order to ensure an accurate assessment of the quantity of existing data. The inventory results are stored in the "AvailableData" database that is included on the CD accompanying this report. The database, which was developed to be consistent with the checklist provided in **Appendix A**, uses codes of "1" (data available) and "0" (no data available) to indicate with a daily resolution whether $PM_{2.5}$ mass data (daily and hourly), $PM_{2.5}$ ion data (SO_4^{2-}, NO_3^-, Cl^-, NH_4^+, K^+, Na^+, continuous SO_4^{2-}, continuous NO_3^-), $PM_{2.5}$ carbon data (elemental carbon, organic carbon, continuous elemental and organic carbon), $PM_{2.5}$ elemental composition data (40 elements), $PM_{2.5}$ water-soluble elemental composition data, and PM_{10} data (daily and hourly) are available. Fields and sub-tables are also included to house information about the sampling and analytical methods used to produce the data. A diagram of the database design is provided as **Appendix B**.

Certain measurements were made with a finer-than-daily time resolution. These include all continuous or semi-continuous measurements, as well as certain filter-based measurements that involved collection of multiple filters throughout the course of a day. For these measurements, data were considered to be available for a given day (i.e., a "1" was assigned) only if valid observations covering at least 19 hours (i.e., 79%) of the day were available. Similarly, measurements in which a single filter was exposed for greater than 29 hours were considered to be invalid. Otherwise, a measurement was only considered to be invalid if it was qualified as such according to the quality assurance / quality control (QA/QC) procedures followed by the group responsible for collecting and reporting the data, if no value was reported, or if the reported value was physically unreasonable (in cases where the data had not yet undergone stringent QA/QC). Data that were "flagged" but not marked as invalid were considered to be valid for purposes of this inventory. QA/QC procedures followed by the various monitoring programs identified above are discussed in **Section 2.4.3** of this report. Finally, for cases in which collocated measurements of a parameter were made using different methods on a given day at a given site, only the preferred method is cited in the AvailableData database.

Inventory results for all of the parameters of interest are summarized below.

2.2.1 $PM_{2.5}$ Mass Concentration Data

Table 2 summarizes inventory results for $PM_{2.5}$ mass concentration data collected by monitoring sites in the 35-county greater Pittsburgh region between 2000 and 2005. There were 47 monitoring sites that measured ambient $PM_{2.5}$ mass concentrations during some or all of this period. (A number of these sites also collected $PM_{2.5}$ data in 1999, but a site-by-site inventory was not performed for that year). Sixteen of these sites were located in Allegheny County, and 23 were located in the seven-county Pittsburgh Metropolitan Statistical Area (MSA) comprising Allegheny, Armstrong, Beaver, Butler, Fayette, Washington, and Westmoreland counties in western Pennsylvania. **Figure 2** presents a map showing the locations of the $PM_{2.5}$ monitoring sites. To provide some indication of the value of the sites for characterizing the exposure of the region's population, the site locations are layered over a plot of population density.

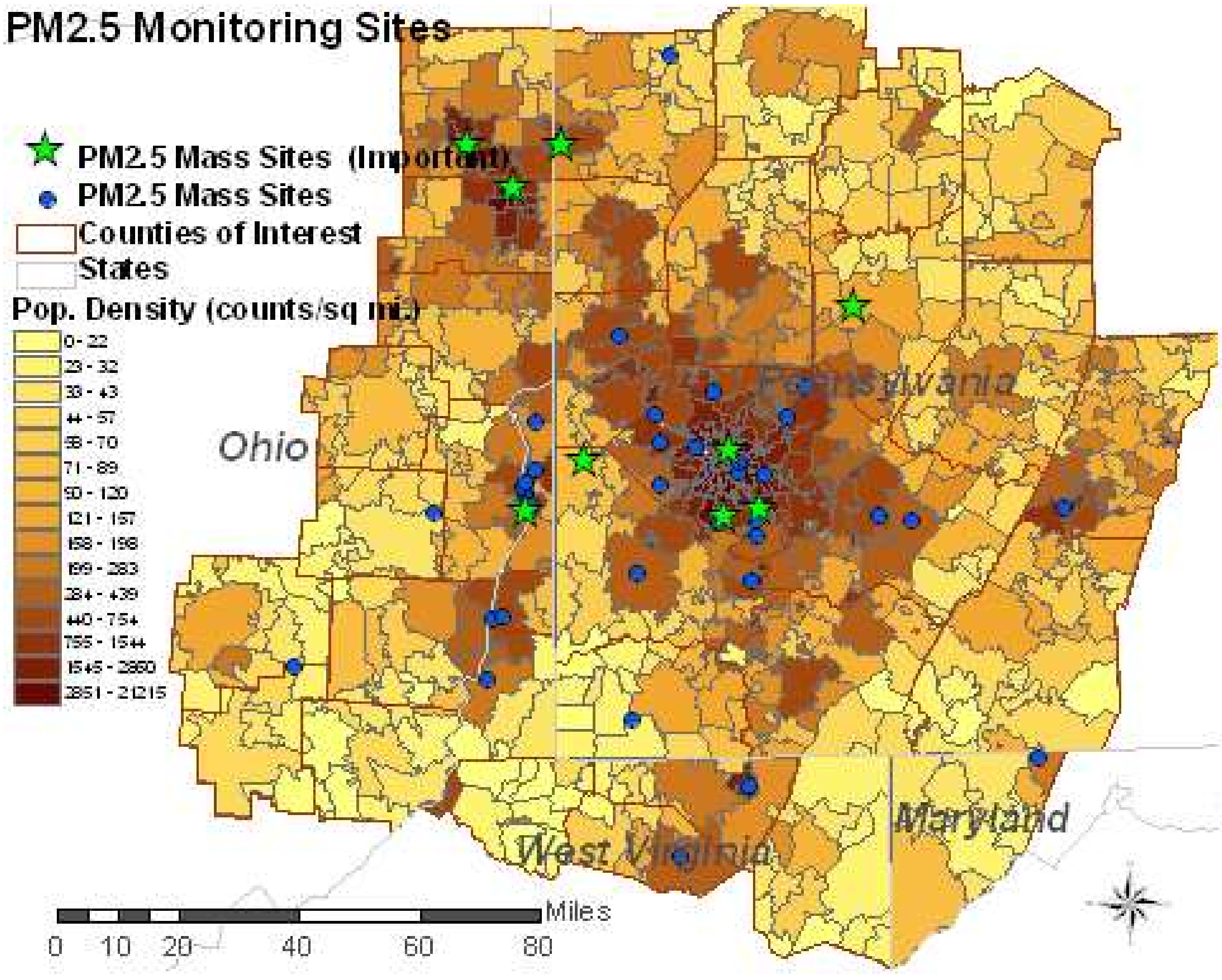


Figure 2: $PM_{2.5}$ monitoring sites in the 35-county greater Pittsburgh region, 2000-2005.

Table 2: Summary of $PM_{2.5}$ total mass concentration data collected by monitoring sites in the 35-county greater Pittsburgh region between 2000 and 2005.

Site Name / ID	State	County	Latitude, Longitude	Program	Sampling Method	Approximate Sampling Period[a]	Approximate Sampling Schedule
FRRE1	MD	Garrett	39.7058N 79.0122W	IMPROVE	IMPROVE	Apr 2004 – 2005	1 in 3 days
Hopedale	OH	Harrison	40.32N 80.90W	SCAMP	FRM	May 2000 – May 2002	Daily
390810016	OH	Jefferson	40.3628N 80.6156W	AQS	FRM	Jan 2000 – Oct 2003	1 in 3 days
390810017	OH	Jefferson	40.3661N 80.6150W	AQS	FRM[b] TEOM	Nov 2003 – 2005 Apr 2004 – 2005	1 in 3 days Continuous
390811001	OH	Jefferson	40.3219N 80.6064W	AQS	FRM	Jan 2000 – 2005	Daily (until 1/31/04) 1 in 6 days (2/3/04 – 2005)
Franciscan U. of Steubenville	OH	Jefferson	40.38N 80.62W	SCAMP	FRM[b] TEOM	May 2000 – May 2002 June 2000 – May 2002	Daily Continuous
390990005	OH	Mahoning	41.1111N 80.6453W	AQS	FRM	Jan 2000 – 2005	Daily (through 2004) 1 in 6 days (2005)
390990014	OH	Mahoning	41.0959N 80.6584W	AQS	FRM[b] TEOM	Oct 2002 – 2005 Oct 2002 – 2005	Daily (through 2004) 1 in 3 days (2005) Continuous
QAK272/572 QUCI1	OH	Noble	39.9428N 81.3378W	CASTNet IMPROVE	CASTNet IMPROVE	Jan 2000 – Apr 2001 May 2001 – 2005	1 in 6 days 1 in 3 days
391550007	OH	Trumbull	41.2142N 80.7875W	AQS	FRM	Jan 2000 – 2005	Daily (through 2004) 1 in 3 days (2005)
420030008	PA	Allegheny	40.4656N 79.9611W	AQS[c]	FRM[b] TEOM	Jan 2000 – 2005 May 2000 – 2005	Daily Continuous
420030021	PA	Allegheny	40.4136N 79.9414W	AQS	FRM[b]	Jan 2000 – 2005	1 in 3 days
420030064	PA	Allegheny	40.3236N 79.8683W	AQS	FRM[b] TEOM	Jan 2000 – 2005 Jan 2000 – 2005	Daily Continuous
420030067	PA	Allegheny	40.3819N 80.1856W	AQS	FRM	Jan 2000 – 2005	1 in 3 days
420030093	PA	Allegheny	40.6072N 80.0208W	AQS	FRM	Jan 2000 – 2005	1 in 6 days
420030095	PA	Allegheny	40.4869N 80.1881W	AQS	FRM	Jan 2000 – 2005	1 in 6 days
420030097	PA	Allegheny	40.5531N 80.2033W	AQS	FRM	Jan 2000 – Dec 2000	1 in 6 days
420030116	PA	Allegheny	40.4736N 80.0772W	AQS	FRM	Jan 2000 – 2005	1 in 3 days
420030131	PA	Allegheny	40.2894N 80.0050W	AQS	FRM	Jan 2000 – Feb 2003	1 in 6 days
420030133	PA	Allegheny	40.2601N 79.8865W	AQS	FRM	Feb 2003 – 2005	1 in 6 days
420031008	PA	Allegheny	40.6186N 79.7272W	AQS	FRM	Jan 2000 – 2005	1 in 3 days
420031301	PA	Allegheny	40.4025N 79.8603W	AQS	FRM	Jan 2000 – 2005	1 in 3 days
420033007	PA	Allegheny	40.2944N 79.8867W	AQS	FRM	Jan 2001 – 2005	1 in 6 days
420039002	PA	Allegheny	40.5469N 79.7839W	AQS	FRM	Jan 2000 – 2005	1 in 6 days
Bruceton	PA	Allegheny	40.3065N 79.9794W	NETL/OST	FRM[b] TEOM	Jan 2000 – Jun 2004 Jan 2000 – Sep 2004	Daily Continuous

Site Name / ID	State	County	Latitude, Longitude	Program	Sampling Method	Approximate Sampling Period[a]	Approximate Sampling Schedule
Schenley Park	PA	Allegheny	40.4395N 79.9405W	PAQS	FRM[b] TEOM	May 2001 – Jun 2002 Jul 2001 – Aug 2002	Daily Continuous
420050001	PA	Armstrong	40.8142N 79.5650W	AQS	TEOM	Jan 2000 – 2005	Continuous
420070014	PA	Beaver	40.7478N 80.3167W	AQS	FRM TEOM	Jan 2000 – 2005 Jul 2004 – 2005	1 in 3 days Continuous
420210011	PA	Cambria	40.3097N 78.9150W	AQS	FRM TEOM	Jan 2000 – 2005 Aug 2004 – 2005	1 in 3 days Continuous
Holbrook	PA	Greene	39.8162N 80.2846W	UORVP	FRM/SFS TEOM	Jan 2000 – Jan 2002 Jan 2000 – Jul 2002	Intermittent Continuous
420850100	PA	Mercer	41.2150N 80.4850W	AQS	FRM	Apr 2000 – 2005	Daily
MKG513 MKG01	PA	Mercer	41.4269N 80.1453W	CASTNet IMPROVE	CASTNet IMPROVE	Jan 2000 – May 2001 Apr 2001 – 2005	1 in 6 days 1 in 3 days
421250005	PA	Washington	40.1467N 79.9022W	AQS	FRM	Jan 2000 – 2005	1 in 3 days
421250200	PA	Washington	40.1706N 80.2614W	AQS	FRM	Jan 2000 – 2005	1 in 3 days
421255001	PA	Washington	40.4453N 80.4208W	AQS	FRM/FEM[b]	Jan 2000 – 2005	Daily
421290008	PA	Westmoreland	40.3047N 79.5057W	AQS	FRM/FEM[b]	Jan 2000 – 2005	1 in 3 days
St. Vincent College	PA	Westmoreland	40.29N 79.40W	SCAMP	FRM	May 2000 – May 2002	Daily
540090005	WV	Brooke	40.3381N 80.5972W	AQS	FRM	Jan 2000 – 2005	1 in 3 days
540290011	WV	Hancock	40.3945N 80.6120W	AQS	FRM	Jan 2000 – 2005	1 in 3 days
540291004	WV	Hancock	40.4215N 80.5809W	AQS	FRM	Jan 2000 – 2005	1 in 3 days
Tomlinson Run State Park	WV	Hancock	40.54N 80.58W	SCAMP	FRM	May 2000 – May 2002	Daily
540490006	WV	Marion	39.4808N 80.1353W	AQS	FRM	Jan 2000 – 2005	1 in 3 days
540511002	WV	Marshall	39.9160N 80.7341W	AQS	FRM[b]	Jan 2000 – 2005	1 in 3 days
540610003	WV	Monongalia	39.6494N 79.9211W	AQS[d]	FRM[b]	Jan 2000 – 2005	1 in 3 days
540690008	WV	Ohio	40.0638N 80.7205W	AQS	FRM	Jan 2000 – Dec 2004	1 in 3 days
540690010	WV	Ohio	NA NA	AQS	FRM	2005	1 in 3 days
Wheeling Jesuit University	WV	Ohio	40.07N 80.69W	SCAMP	FRM	May 2000 – May 2002	Daily

Notes: FRM = Federal Reference Method. FEM = Federal Equivalent Method. TEOM = tapered element oscillating microbalance. SFS = sequential filter sampler. [a]Some sites had prolonged periods of missing data within the listed time frame. [b]Twenty-four hour average $PM_{2.5}$ mass concentrations are also available from speciation or other sampler data for days on which speciation or other sampling was performed at this site. These data fall within the date range and frequency listed for the FRM sampler and are therefore not listed separately in the table. [c]$PM_{2.5}$ mass data for this site also available from the UORVP and IMPROVE monitoring programs. [d]$PM_{2.5}$ mass data for this site also available from the UORVP monitoring program.

As shown in **Figure 2** the $PM_{2.5}$ monitoring sites are spatially well distributed throughout the region. Monitoring sites were located in or near most areas of high population density during at least part of the time period of interest, and several monitoring sites were located in less densely populated areas to provide an indication of ambient $PM_{2.5}$ concentrations in more rural portions of the region. Nine of the 47 sites measured $PM_{2.5}$ mass concentrations on a 1-in-1 day frequency for at least a four-year period during 2000-2005. These sites are denoted with green stars in **Figure 2**, because they are the sites that are capable of providing the time series of daily $PM_{2.5}$ mass concentration data that would be required for a retrospective time series epidemiologic study. The 38 remaining sites, which are indicated with blue dots, measured $PM_{2.5}$ on a less-than-daily frequency or during a shorter period of time than the "important" sites; however, these sites may nevertheless be useful for developing a spatial model of $PM_{2.5}$ concentrations in the region that can be used to improve exposure estimates.

Three of the nine "important" sites are located in Allegheny County: the Lawrenceville site (420030008), which is situated in an urban area of the City of Pittsburgh, the Liberty Borough site (420030064), which is situated in the Monongahela River Valley near a major coke production facility, and the Bruceton site, which is situated in a suburban area of southern Allegheny County. Two more are located in comparatively remote areas in Florence, Washington County (421255001) to the west of Pittsburgh, and in Kittanning, Armstrong County (420050001) to the northeast of Pittsburgh. The remaining four are located in or near Mingo Junction, Ohio (390811001), Youngstown, Ohio (390990005), Warren, Ohio (391550007), and Sharon, Pennsylvania (420850100).

2.2.2 $PM_{2.5}$ Chemical Speciation Data

As discussed above, a day-by-day inventory of $PM_{2.5}$ chemical speciation data was performed for each monitoring site in the 35-county greater Pittsburgh region that collected these data between 1999 and 2005. **Table 3** provides an overview of these inventory results for the 15 sites in the region that monitored for a complete suite of $PM_{2.5}$ chemical components, including ions, carbon, and trace and crustal elements, during some or all of the time period of interest. For each monitoring site, the "number of days with complete $PM_{2.5}$ speciation" is the number of days for which fine particulate SO_4^{2-}, NO_3^-, EC, OC, and elemental (for at least 15 elements) mass concentration data were all determined and valid during the same 24-hour period. (Ammonium, which constitutes a substantial portion of the total mass of ambient $PM_{2.5}$ in the Pittsburgh region, is not included in the definition of "complete $PM_{2.5}$ speciation," because fine particulate NH_4^+ is almost entirely associated with SO_4^{2-} and NO_3^-, and its concentration can be estimated from concentrations of these species). **Figure 3** shows the locations of the $PM_{2.5}$ speciation monitoring sites that are listed in **Table 3** **Figure 4** presents a time line showing the days on which $PM_{2.5}$ speciation sampling occurred at the various monitoring sites.

As shown in **Table 3**, the site with the greatest number of days of existing, complete $PM_{2.5}$ speciation data is the Lawrenceville site in the City of Pittsburgh. This site is an important source of exposure information for the proposed epidemiology study because of its abundance of existing data and its central location within the region's most densely populated area. Three monitoring campaigns collected $PM_{2.5}$ speciation data at the Lawrenceville site between 1999 and 2005: AQS, UORVP, and IMPROVE. AQS monitoring, conducted by ACHD, produced 422 days with complete $PM_{2.5}$ speciation as a result of predominantly 1-in-3 day sampling between June 30, 2001, and April 10, 2005. If the UORVP data are merged with the AQS data, the total number of days with complete speciation data increases to 587 between February 17, 1999, and April 10, 2005), and if the IMPROVE data are merged with the AQS and UORVP data, the total number of days increases to 603. Also, for purposes of conducting a $PM_{2.5}$ time-series epidemiology study, it is ultimately necessary to assemble an exposure database containing data for each day of the study (as opposed to data for every third or sixth day). The UORVP monitoring activities provided a 5-month stream of 1-in-1 day $PM_{2.5}$ speciation data for the period between October 1, 2002, and February 27, 2003, which will be useful for assembling such an exposure database for the Pittsburgh region.

The M.K. Goddard site in Mercer County, Pennsylvania, and the Quaker City site in Noble County, Ohio, have the second and third largest numbers of days with complete $PM_{2.5}$ speciation of the 15 sites listed in **Table 3**. Each of these sites performed $PM_{2.5}$ speciation measurements on a 1-in-6 day frequency between March 1999 and May 2001 as part of the CASTNet program, and on a 1-in-3 day frequency between spring 2001 and the present as part of the IMPROVE program. However, because of their less-than-daily sampling frequencies and their locations in remote areas more than 100 km from the City of Pittsburgh, these sites are of less importance than the Lawrenceville site for representing the exposure of the region's population to chemical components of $PM_{2.5}$.

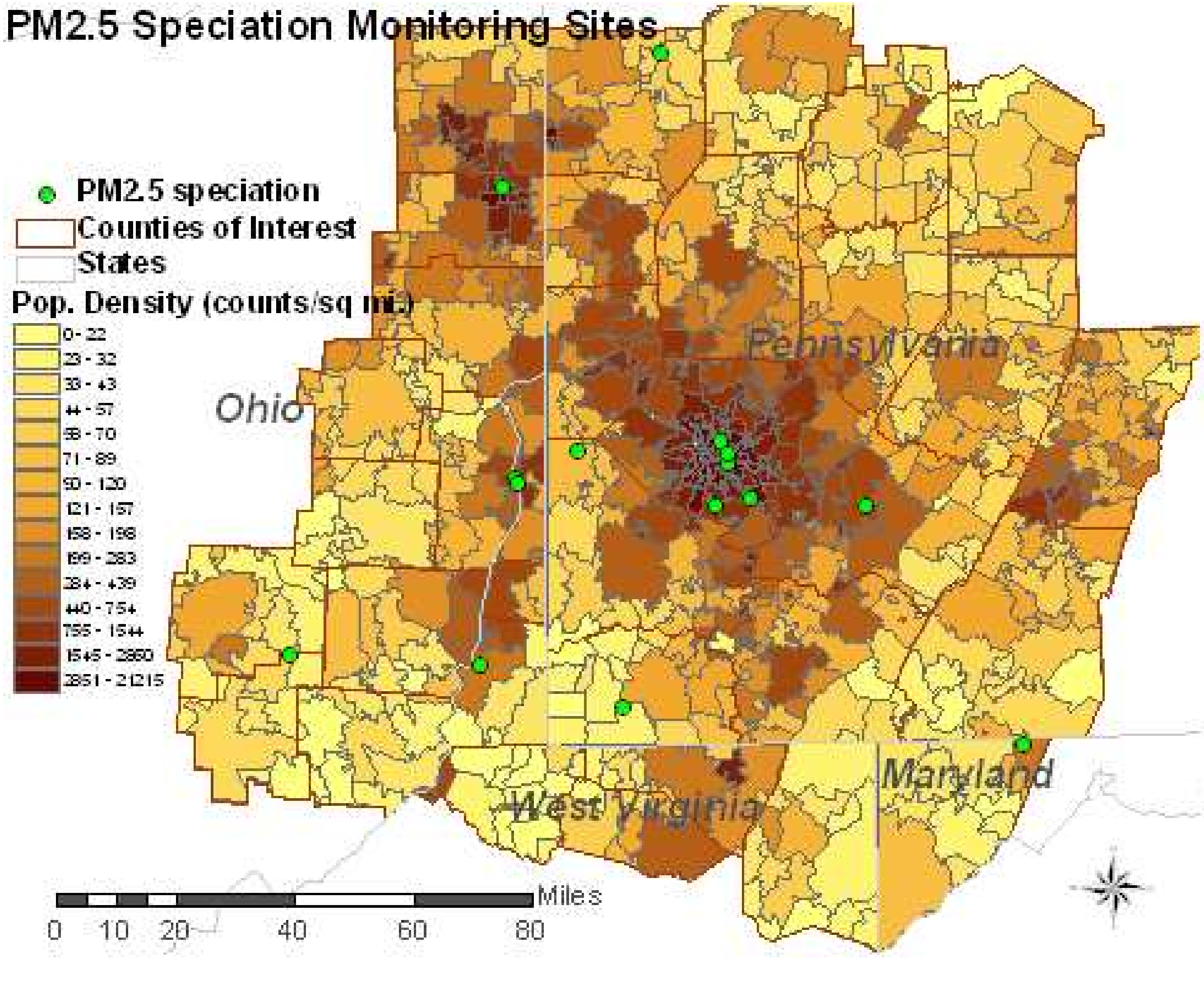


Figure 3: PM$_{2.5}$ speciation monitoring sites in the 35-county greater Pittsburgh region, 1999-2005.

Table 3: Overview of data inventory results for sites in the 35-county greater Pittsburgh region that collected $PM_{2.5}$ speciation data between 1999 and 2005.

Site	Site Network Code(s)	County, State	Latitude, Longitude	Program	Approximate Period of $PM_{2.5}$ Speciation Sampling[a]	Approximate Frequency of $PM_{2.5}$ Speciation Sampling	Number of Days with Complete $PM_{2.5}$ Speciation[b]
Bruceton (BRU)	NA	Allegheny, PA	40.3065N 79.9794W	NETL/OST	10/28/99 – 09/30/01	Intermittent	171
Florence (FLO)	421255001	Washington, PA	40.4453N 80.4208W	AQS	06/30/01 – 04/10/05	1 in 6 days[c]	254
Franciscan U. (FRA)	NA	Jefferson, OH	40.38N 80.62W	SCAMP	08/16/00 – 05/06/02	1 in 4 days	104
Frostburg (FRO)	FRRE1	Garrett, MD	39.7058N 79.0122W	IMPROVE	04/18/04 – 12/29/04	1 in 3 days	83
Greensburg (GRE)	421290008	Westmoreland, PA	40.3047N 79.5057W	AQS	06/30/01 – 04/10/05	1 in 6 days[c]	256
Hazelwood (HAZ)	420030021	Allegheny, PA	40.4136N 79.9414W	AQS	06/30/01 – 09/30/03	1 in 6 days[c]	145
Holbrook (HOL)	NA	Greene, PA	39.8162N 80.2846W	UORVP	02/17/99 – 08/8/01	Intermittent	97
Lawrenceville (LAW)	420030008, PITT1	Allegheny, PA	40.4656N 79.9611W	AQS, UORVP, IMPROVE	02/17/99 – 04/10/05	Intermittent[d]	603
Liberty (LIB)	420030064	Allegheny, PA	40.3236N 79.8683W	AQS	10/6/03 – 04/10/05	1 in 6 days	74
M. K. Goddard (MKG)	MKG01, MKG513	Mercer, PA	41.4269N 80.1453W	IMPROVE, CASTNet	03/01/99 – 12/29/04	Intermittent[e]	561
Moundsville (MOU)	540511002	Marshall, WV	39.9178N 80.7342W	AQS	06/02/04 – 04/10/05	1 in 6 days	53
Quaker City (QUA)	QUCI1, QAK272, QAK572	Noble, OH	39.9428N 81.3378W	IMPROVE, CASTNet	03/01/99 – 12/29/04	Intermittent[f]	559
Schenley Park	NA	Allegheny, PA	40.4395N	PAQS	07/01/01 – 07/20/02	Daily	333

Site	Site Network Code(s)	County, State	Latitude, Longitude	Program	Approximate Period of $PM_{2.5}$ Speciation Sampling[a]	Approximate Frequency of $PM_{2.5}$ Speciation Sampling	Number of Days with Complete $PM_{2.5}$ Speciation[b]
(SCH)			79.9405W				
Steubenville (STE)	390810017	Jefferson, OH	40.3661N 80.6150W	AQS	08/01/04 – 4/10/05	1 in 6 days	33
Youngstown (YOU)	390990014	Mahoning, OH	41.0958N 80.6584W	AQS	02/13/02 – 04/10/05	1 in 6 days	183

[a]At the time of the inventory, data were available for the AQS sites through 4/10/05 and for the IMPROVE sites through 12/29/04. [b]"Complete $PM_{2.5}$ speciation" defined as including SO_4^{2-}, $NO3^-$, elemental carbon, organic carbon, and elemental (for at least 15 elements) mass concentration data. [c]$PM_{2.5}$ speciation was determined at a higher frequency during several monitoring intensives, but in some cases was not determined at all for prolonged periods following the intensives. [d]AQS speciation monitoring occurred on an approximately 1-in-3 day frequency from 6/30/01-4/10/05, although speciation was determined at a higher frequency during several monitoring intensives, and in some cases was not determined at all for prolonged periods following the intensives. UORVP speciation monitoring occurred between 2/17/99 and 2/27/03, and included daily speciation monitoring from 10/1/02-2/27/03. IMPROVE speciation monitoring occurred on a 1-in-3 day frequency from 4/18/04 - 12/29/04. [e]CASTNet speciation monitoring occurred on a 1-in-6 day frequency from 3/1/99 - 5/31/01; IMPROVE speciation monitoring occurred on a 1-in-3 day frequency from 4/19/01 - 12/29/04. [f]CASTNet speciation monitoring occurred on a 1-in-6 day frequency from 3/1/99 - 5/1/01; IMPROVE speciation monitoring occurred on a 1-in-3 day frequency from 5/4/01 - 12/29/04.

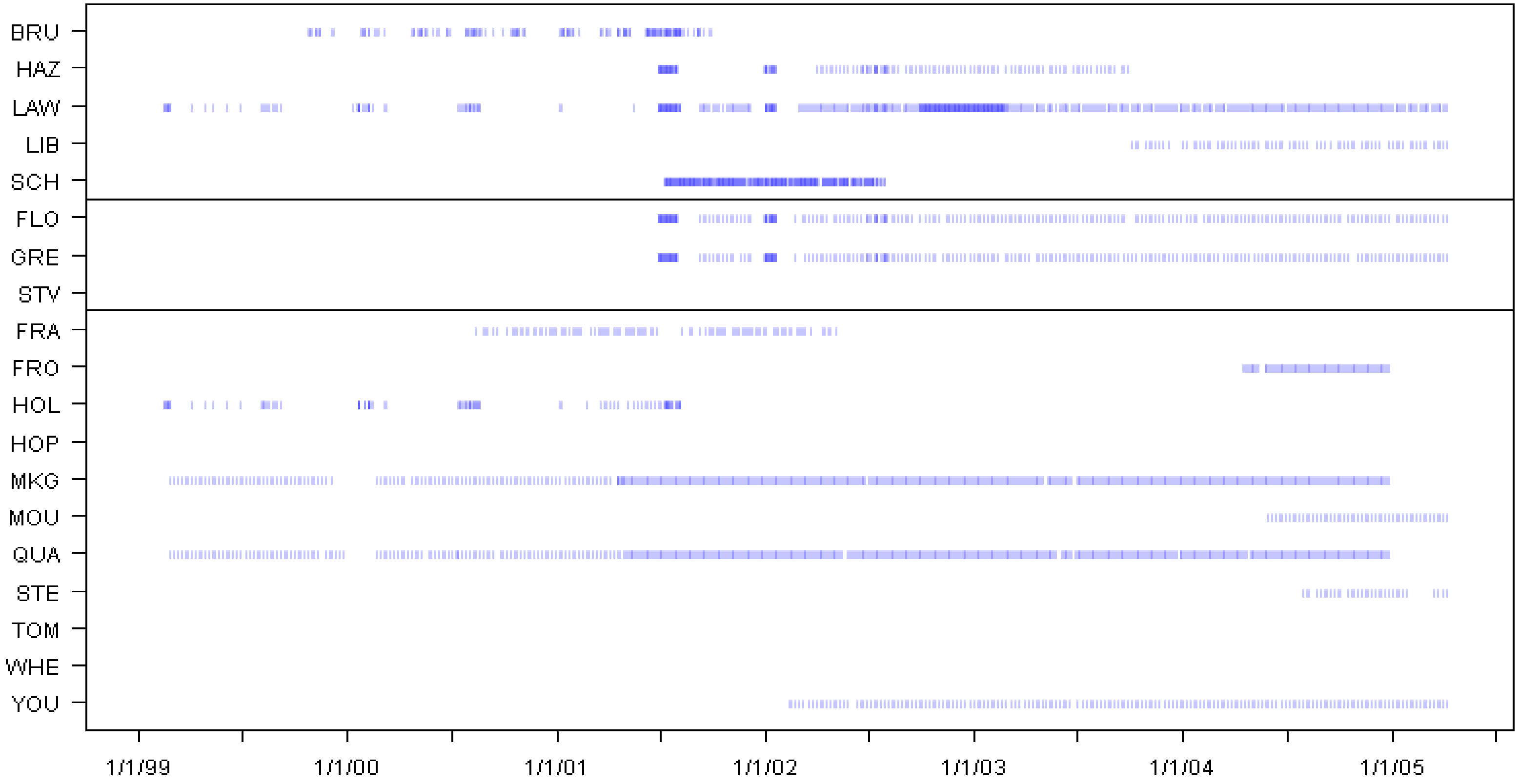


Figure 4: Time line showing the days for which a complete set of $PM_{2.5}$ speciation data (as defined in the text) are available from the sites in the 35-county greater Pittsburgh region that monitored for $PM_{2.5}$ speciation between 1999 and 2005. Sites in the top portion of the plot are located in Allegheny County; sites in the middle portion are located in the Pittsburgh MSA, and sites in the lower portion are located outside of the Pittsburgh MSA.

Table 4: Detailed summary of $PM_{2.5}$ speciation data availability by species and monitoring site for the 35-county greater Pittsburgh region between 1999 and 2005. Inventory for 2005 does not include all data collected in that year. At the time of the inventory, data were available for the AQS sites through 4/10/05 and for the IMPROVE sites through 12/29/04.

	Number of Days With:				
Site	**Sulfate**	**Nitrate**	**EC/OC**	**Elements**	**Complete Speciation**
Bruceton	996	1062	1078	206	171
Florence	255	255	255	256	254
Franciscan University of Steubenville	151	151	142	127	104
Frostburg	83	83	83	83	83
Greensburg	259	259	259	260	256
Hazelwood	146	146	146	146	145
Holbrook	97	97	97	97	97
Hopedale	129	129	0	0	0
Lawrenceville	606	606	605	604	603
Liberty	75	75	75	74	74
M.K. Goddard	569	569	566	564	561
Moundsville	53	53	53	53	53
Quaker City	563	563	564	562	559
Schenley Park	399	374	398	375	333
Steubenville	33	33	33	34	33
St. Vincent College	155	155	0	0	0
Tomlinson Run State Park	161	161	0	0	0
Wheeling Jesuit University	96	96	0	0	0
Youngstown	187	187	185	184	183

The PAQS monitoring site in Pittsburgh's Schenley Park collected a complete set of $PM_{2.5}$ speciation data on 333 (86%) of the days between July 1, 2001, and July 20, 2002. Although this site only operated for approximately one year, it is an important source of $PM_{2.5}$ speciation information for the proposed epidemiology study during that period because of its 1-in-1 day sampling frequency and its location in central Allegheny County. The Schenley site is located only about 3 km from the Lawrenceville site; the feasibility of using data from these sites interchangeably is explored in **Section 2.4** of this report.

The remaining 11 monitoring sites listed in **Table 3** each collected a complete set of $PM_{2.5}$ speciation data on less than 300 days during the inventoried period. Among these sites, the AQS sites in Florence and Greensburg had the greatest number of days (254 and 256, respectively) with complete $PM_{2.5}$ speciation. Although $PM_{2.5}$ speciation was only determined every sixth day for these sites, these data are nevertheless useful for assessing the spatial variability of $PM_{2.5}$ in the immediate Pittsburgh vicinity. As shown in **Figure 3** the $PM_{2.5}$ speciation monitoring sites in Allegheny County cover only a narrow region extending approximately due south from the City of Pittsburgh in the center of the county to Bruceton and Liberty in the southern part of the county. Speciation data collected at the Florence and Greensburg sites could be utilized to model exposures in the western and eastern portions of the county, respectively.

Although the quantification of $PM_{2.5}$ speciation data presented in **Table 3** accurately reflects the amount of speciation monitoring conducted at most of the sites in the 35-county greater Pittsburgh region, its exclusion of days for which some but not all of the desired $PM_{2.5}$ components were measured understates the importance of several monitoring sites in the region. Thus, **Table 4** summarizes the $PM_{2.5}$ speciation data inventory results for each monitoring site by individual $PM_{2.5}$ component (components that were generally determined from a common sample, including carbonaceous species and elemental species, are grouped together in the table). **Appendix C** presents time lines, similar to the one presented in **Figure 4**, for these individual components.

As shown in **Table 4,** $PM_{2.5}$ speciation data, including fine particulate SO_4^{2-} and NO_3^- mass concentrations (as well as mass concentrations of water-soluble elemental components of $PM_{2.5}$, which are not indicated in the table), were measured at several monitoring sites that did not sample for a complete suite of $PM_{2.5}$ components, and hence are not included in **Table 3** or in **Figure 3.** These include the SCAMP sites at Hopedale, Ohio (HOP); St. Vincent College in Latrobe, Pennsylvania (STV); Tomlinson Run State Park, West Virginia (TOM); and Wheeling Jesuit University in Wheeling, West Virginia (WHE). Moreover, for a few sites, including the SCAMP site at Franciscan University of Steubenville and the PAQS site at Schenley Park, the number of days having a complete set of $PM_{2.5}$ speciation data is substantially less than the number of days having data for individual $PM_{2.5}$ components, reflecting the effects of scattered cases of missing or invalid data for individual $PM_{2.5}$ components on the inventory results for “complete speciation.”

The results presented in **Table 3** particularly understate the amount of $PM_{2.5}$ speciation data collected at the NETL/OST Bruceton monitoring site. As indicated in **Table 3,** complete sets of $PM_{2.5}$ speciation data are available for only 171 days at the Bruceton site between October 28, 1999, and September 30, 2001. However, this low count results from the fact that only a small portion of the $PM_{2.5}$ samples that were collected at the site have been submitted for elemental analysis. (Per **Section 2.3** of this report, these samples, which are still being archived, may be analyzed as part of the proposed epidemiology study to appreciably enhance amount of $PM_{2.5}$ speciation data available from the Bruceton site.) As shown in **Table 4,** SO_4^{2-}, NO_3^-, and EC and OC mass concentrations were each determined at the Bruceton site on approximately 1000 days during the period of interest. Sulfate data were collected between October 18, 1999, and May 4, 2004; nitrate data were collected between October 18, 1999, and March 20, 2004, and carbon data were collected between August 20, 1999, and June 1, 2003. It is noteworthy that many of these data were measured using semi-continuous monitors that may exhibit appreciable bias or imprecision relative to the filter-based techniques that were commonly employed by other monitoring sites in the region. The implications of using these semi-continuous speciation data are explored in **Section 2.4** of this report. Nevertheless, the Bruceton monitoring site, like the Lawrenceville and Schenley Park sites, is an important source of $PM_{2.5}$ speciation information because of its long period of daily monitoring and its location in Allegheny County. The utility of the Bruceton site may be even greater than indicated in **Table 4**, because data collected at the site using a PC-BOSS sampler were not included in the data inventory that is summarized here. Based on a review of the logbook from the NETL/OST sampling site, PC-BOSS samples were collected on about 550 days between November 1999 and February 2002. It is known that many of the samples that were collected between November 1999 and December 2000 have been analyzed to determine concentrations of sulfate, nitrate, elemental carbon, organic carbon (both non-volatile and semi-volatile), and in some cases elements (Modey and Eatough, 2004). Hence, if obtained, these data could supplement the already extensive database of ambient $PM_{2.5}$ component concentrations available from the Bruceton monitoring site.

Thus, none of the individual monitoring sites in the 35-county greater Pittsburgh region that measured $PM_{2.5}$ speciation between 1999 and 2005 are ideally suited for providing exposure estimates for a time series epidemiologic study of the health effects of $PM_{2.5}$ components. The sites that determined a full suite of $PM_{2.5}$ components on a daily basis (e.g., the Schenley Park site) did not operate for the multiple-year period likely required by the study, and the sites that operated for several years (e.g., the Lawrenceville site and the Bruceton site) either determined $PM_{2.5}$ composition on a less-than-daily frequency or did not routinely determine all of the components of interest. However, the inadequacies of individual sites do not necessarily preclude a feasible study. **Sections 2.3, 2.4, and 2.5** of this report examine ways in which the existing $PM_{2.5}$ speciation data from these individual sites can be combined and supplemented with new data obtained by analyzing archived filter-based $PM_{2.5}$ samples in order to allow the construction of time series of daily exposure estimates suitable for use in an epidemiology study.

2.2.3 Co-Pollutant Data

As discussed above, co-pollutant data of interest for a retrospective epidemiologic study of $PM_{2.5}$ include daily ambient concentrations of SO_2, O_3, CO, and NO_2. **Table 5** lists the monitoring sites in the 35-county greater Pittsburgh region that measured these gaseous species between 2000 and 2005, and indicates the time periods during which measurements were made at each site. (As with the $PM_{2.5}$ mass concentration data presented in **Section 2.2.1**, a number of these sites also collected gaseous pollutant data in 1999, but a site-by-site inventory was not performed for that year). All of the gaseous pollutant measurements represented in **Table 5** were made using continuous monitors, resulting in data with an hourly or finer resolution that can be used to compute daily averages (or other metrics appropriate for quantifying exposure, such as maximum 1-hour average concentration, maximum 8-hour average concentration, etc.). **Figures 5 through 8** show the locations of the sites that measured each species. Again, if a site collected data year-round for at least four years during the inventoried period, it is denoted with a star as an “important” site.

There were 49 sites that monitored SO_2, 34 sites that monitored O_3, 20 sites that monitored CO, and 16 sites that monitored NO_2 concentrations in the 35-county region between 2000 and mid-2005. The maps presented in **Figures 5 through 8** suggest that monitoring sites for gaseous pollutants were generally well-positioned to characterize exposures for the region’s most populated areas, although coverage is generally poor for rural parts of the region and for the northeastern portion of the Pittsburgh MSA. Also, with the exception of one monitor in Steubenville, Ohio, that operated between May 2000 and May 2002, no NO_2 concentrations were measured in the region’s non-Pennsylvania counties during the time period of interest.

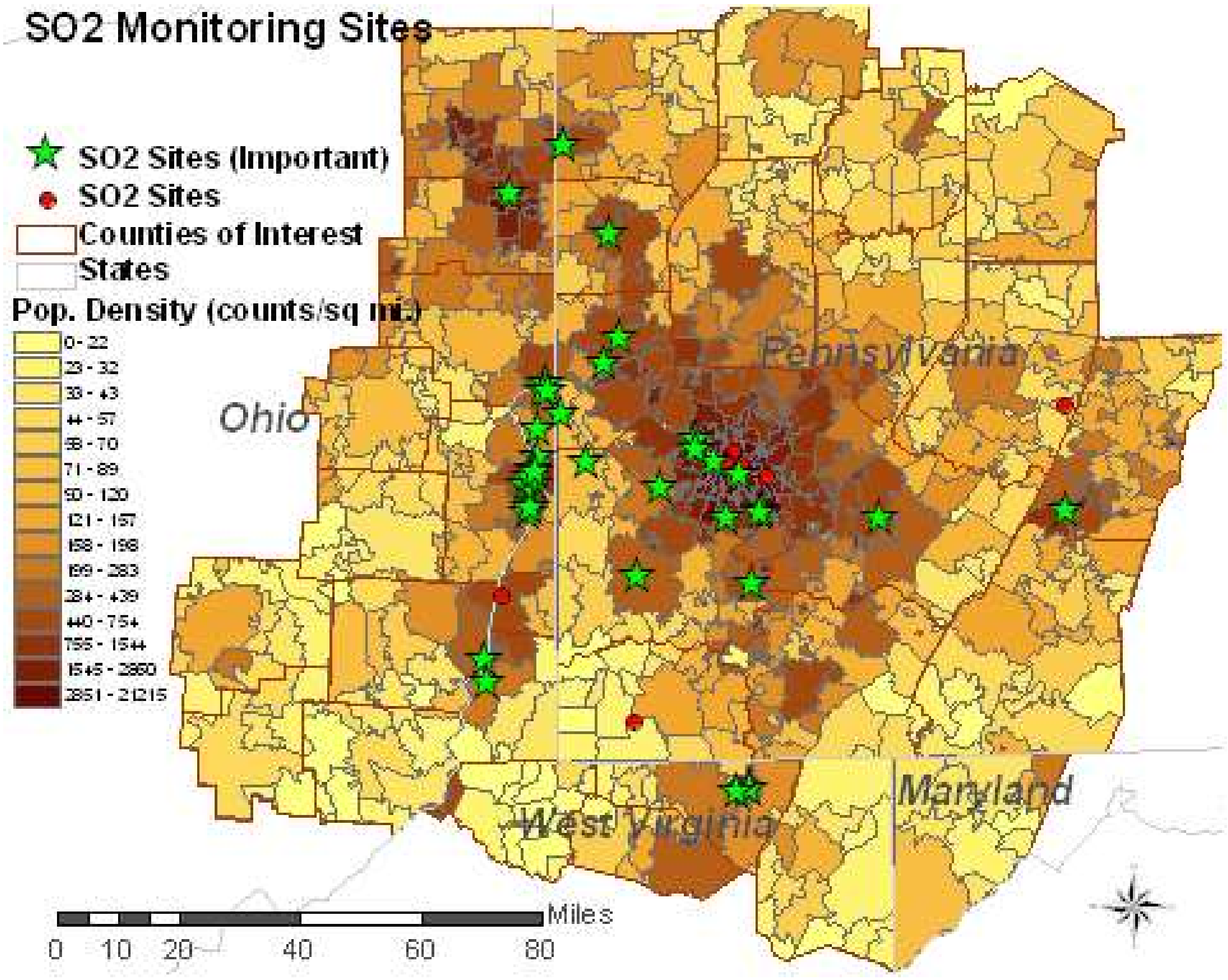


Figure 5: SO_2 monitoring sites in the 35-county greater Pittsburgh region, 2000-2005.

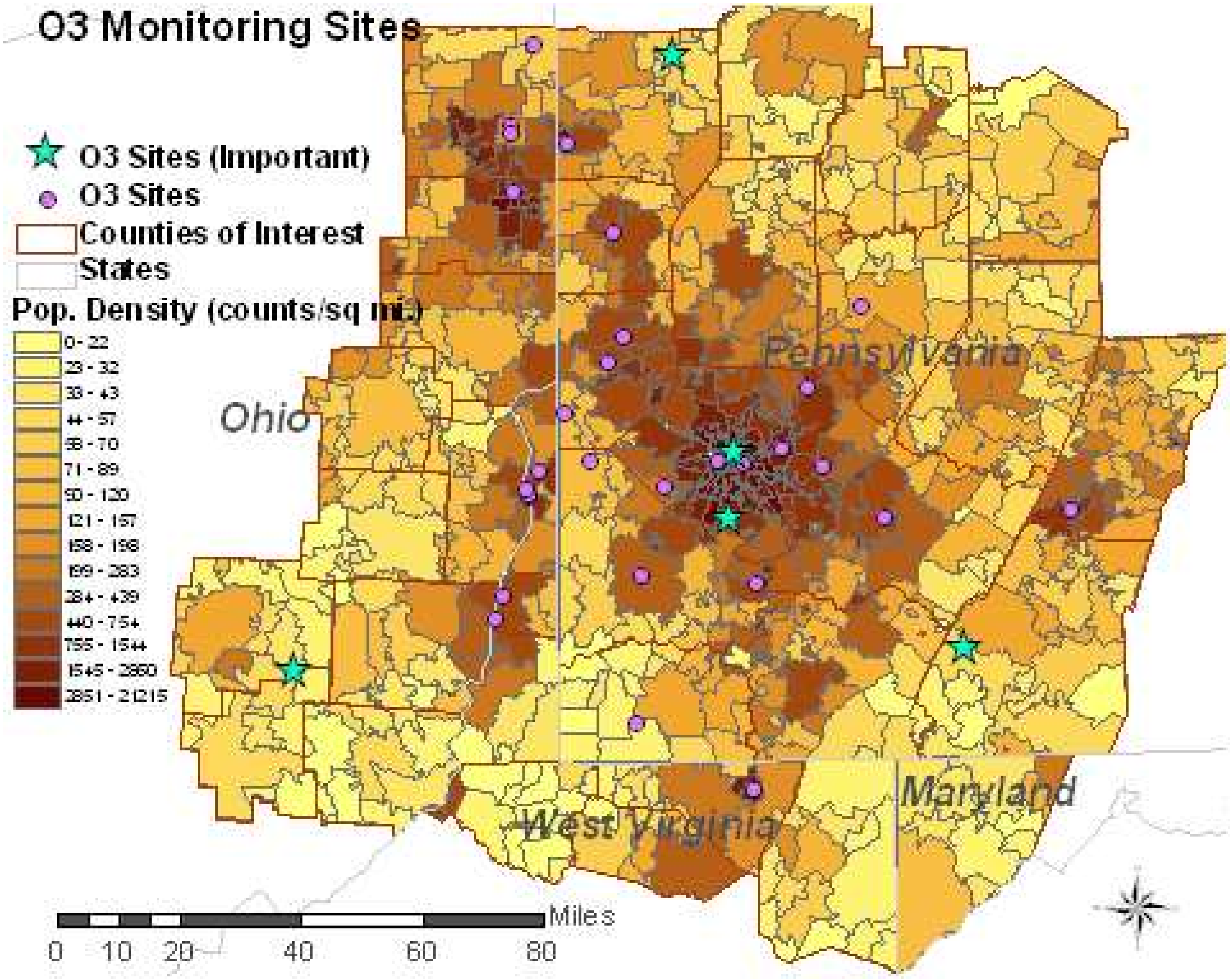


Figure 6: O_3 monitoring sites in the 35-county greater Pittsburgh region, 2000-2005.

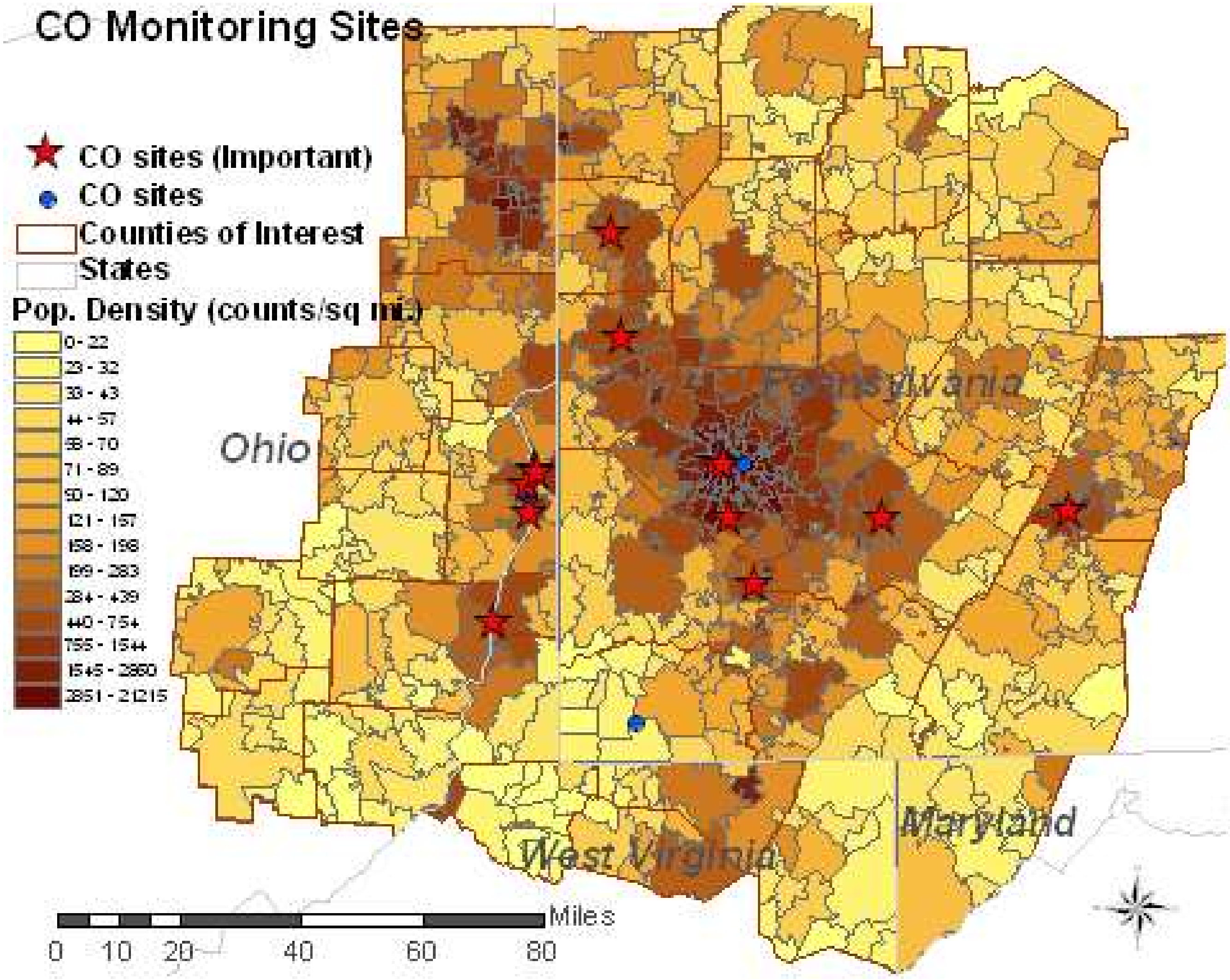


Figure 7: CO monitoring sites in the 35-county greater Pittsburgh region, 2000-2005.

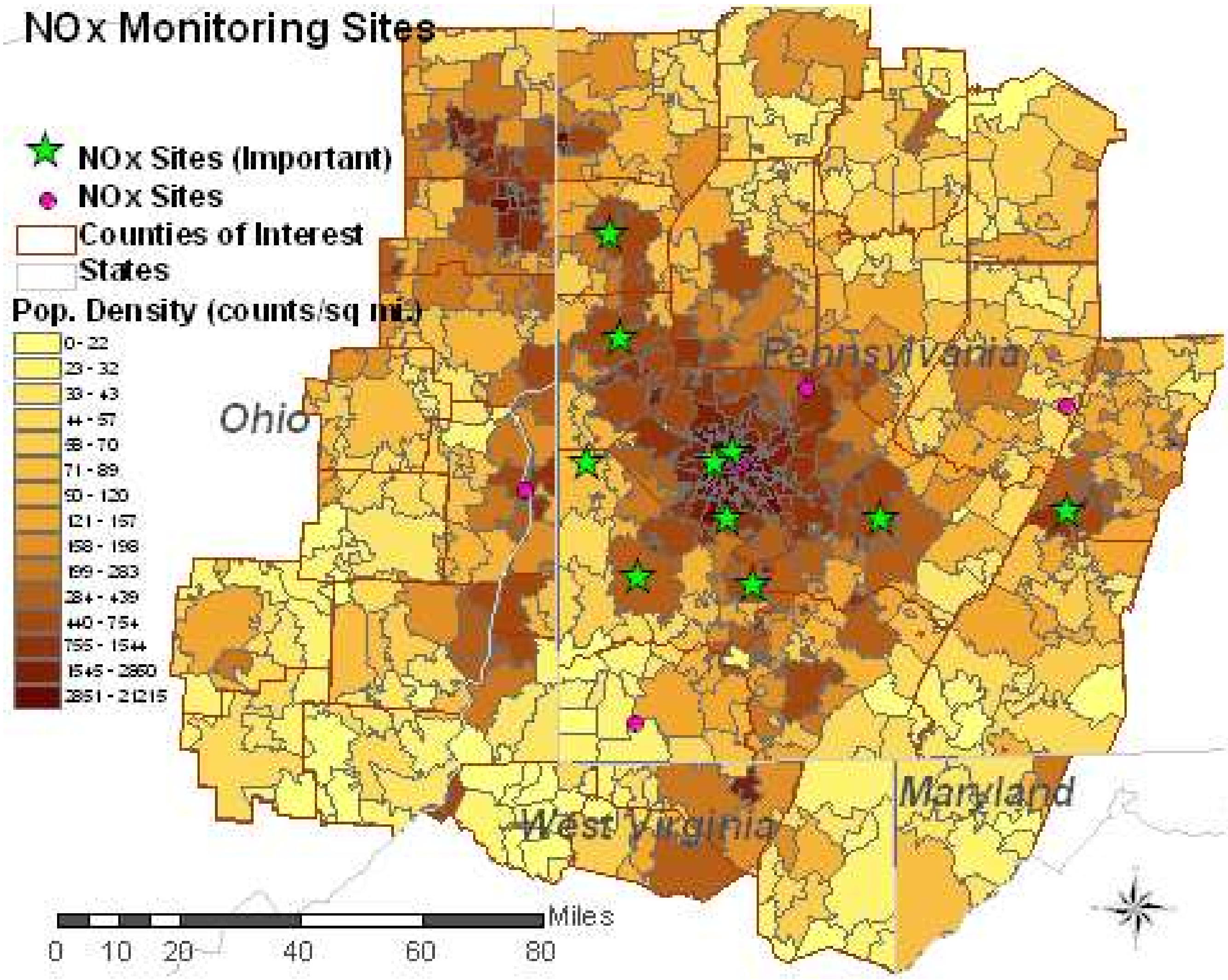


Figure 8: NO2 monitoring sites in the 35-county greater Pittsburgh region, 2000-2005.

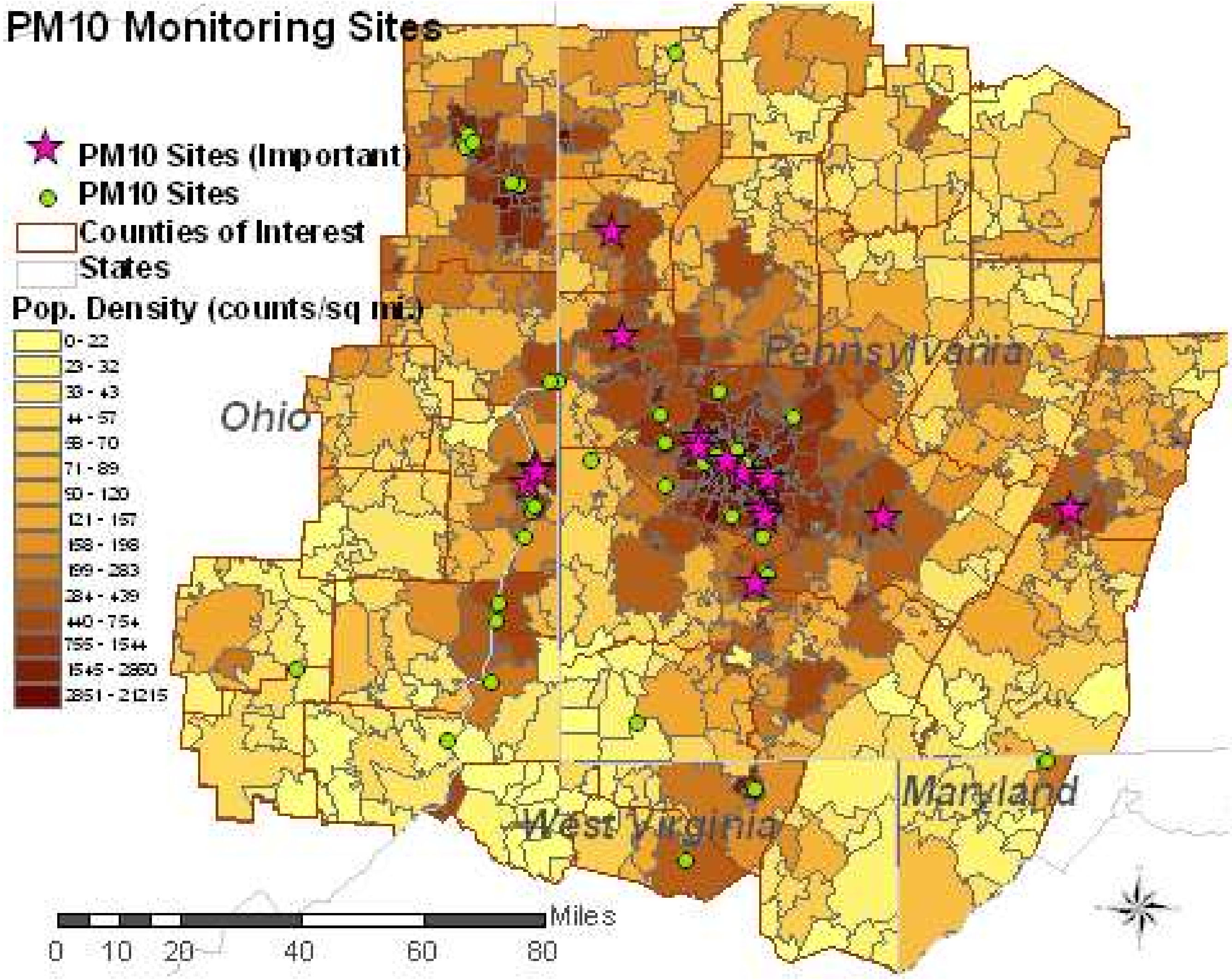


Figure 9: PM$_{10}$ monitoring sites in the 35-county greater Pittsburgh region, 2000-2005.

Table 5: Summary of continuous gaseous pollutant data collected by monitoring sites in the 35-county greater Pittsburgh region between 2000 and 2005.[a]

Site Name / ID	State	County	Latitude / Longitude	Program	SO_2	O_3	CO	NO_2
390133002	OH	Belmont	39.9681N, 80.7475W	AQS	1/00 - 5/05			
390290016	OH	Columbiana	40.6347N, 80.5464W	AQS	1/00 - 4/00			
390290022	OH	Columbiana	40.6350N, 80.5467W	AQS	1/01 - 5/05			
390810016	OH	Jefferson	40.3628N, 80.6156W	AQS	1/00 - 11/03	4/00 - 10/03[b]	1/00 - 11/03	
390810017	OH	Jefferson	40.3661N, 80.6150W	AQS	11/03 - 5/05	4/04 - 5/05[b]	11/03 - 1/04	
390811001	OH	Jefferson	40.3219N, 80.6064W	AQS	1/00 - 1/04		1/00 - 5/05	
Franciscan U.	OH	Jefferson	40.38N,80.62W	SCAMP	5/00 - 5/02	5/00 - 5/02	5/00 - 5/02	5/00 - 5/02
390990013	OH	Mahoning	41.0961N, 80.6586W	AQS	1/00 - 5/05	4/00 - 5/05[b]		
QAK172	OH	Noble	39.9428N, 81.3373W	CASTNet		1/00 - 12/05		
391550008	OH	Trumbull	41.2589N, 80.6661W	AQS		4/00 - 10/01[b]		
391550009	OH	Trumbull	41.4539N, 80.5917W	AQS		4/00 - 5/05[b]		
391550011	OH	Trumbull	41.2401N, 80.6631W	AQS		4/02 - 5/05[b]		
420030002	PA	Allegheny	40.5006N, 80.0719W	AQS	1/00 - 6/05			
420030008	PA	Allegheny	40.4656N, 79.9611W	AQS/UORVP	1/00 - 7/02	1/00 - 6/05		1/00 - 6/05
420030010	PA	Allegheny	40.4456N, 80.0164W	AQS	1/00 - 6/05	4/00 - 6/05[b]	1/00 - 6/05	1/00 - 6/05
420030021	PA	Allegheny	40.4136N, 79.9414W	AQS	1/00 - 6/05			
420030031	PA	Allegheny	40.4433N, 79.9906W	AQS	1/00 - 12/00		5/03 - 6/05	1/00 - 6/01
420030038	PA	Allegheny	40.4389N, 79.9972W	AQS			1/00 - 6/05	
420030052	PA	Allegheny	40.4414N, 80.0033W	AQS			1/00 - 4/00	
420030064	PA	Allegheny	40.3236N, 79.8683W	AQS	1/00 - 6/05			
420030067	PA	Allegheny	40.3819N, 80.1856W	AQS	1/00 - 6/05	4/00 - 6/05[b]		
420030088	PA	Allegheny	40.4722N, 79.8200W	AQS		4/00 - 7/01[b]		
420030116	PA	Allegheny	40.4736N, 80.0772W	AQS	1/00 - 6/05			
420031005	PA	Allegheny	40.6172N, 79.7322W	AQS		1/00 - 6/05[c]		7/01 - 6/05
420031301	PA	Allegheny	40.4025N, 79.8603W	AQS	1/00 - 12/00			
420033003	PA	Allegheny	40.3181N, 79.8811W	AQS	1/00 - 6/05			
420033004	PA	Allegheny	40.3050N, 79.8889W	AQS	1/00 - 12/00			
Bruceton	PA	Allegheny	40.3065N, 79.9794W	NETL/OST	3/00 - 6/04	3/00 - 6/04	3/00 - 6/04	3/00 - 6/04
Schenley Park	PA	Allegheny	40.4395N, 79.9405W	PAQS	7/01 - 8/02	7/01 - 8/02	7/01 - 8/02	7/01 - 8/02
420050001	PA	Armstrong	40.8142N, 79.5650W	AQS		4/00 - 6/05[b]		
420070002	PA	Beaver	40.5625N, 80.5042W	AQS	1/00 - 6/05	4/00 - 6/05[b]		
420070005	PA	Beaver	40.6847N, 80.3597W	AQS	1/00 - 6/05	4/00 - 6/05[b]		
420070014	PA	Beaver	40.7478N, 80.3167W	AQS	1/00 - 6/05	4/00 - 6/05[b]	1/00 - 6/05	1/00 - 6/05

Site Name / ID	State	County	Latitude / Longitude	Program	SO_2	O_3	CO	NO_2
420210011	PA	Cambria	40.3097N, 78.9150W	AQS	1/00 - 6/05	4/00 - 6/05[b]	1/00 - 6/05	1/00 - 6/05
420590002	PA	Greene	39.8162N, 80.2849W	AQS/UORVP	4/00 - 6/05[d]	4/00 - 6/05[d]	4/00 - 6/05[b]	1/00 – 11/01
420630004	PA	Indiana	40.5633N, 78.9200W	AQS	11/04 - 6/05	4/05 - 6/06		11/04 - 6/05
420730015	PA	Lawrence	40.9958N, 80.3467W	AQS	1/00 - 6/05	4/00 - 6/05[b]	1/00 - 6/05	1/00 - 6/05
420850100	PA	Mercer	41.2150N, 80.4850W	AQS	1/00 - 6/05	4/00 - 6/05[b]		
MKG113	PA	Mercer	41.4271N, 80.1451W	CASTNet		1/00 - 12/05		
LRL117	PA	Somerset	39.9878N, 79.2515W	CASTNet		1/00 - 12/05		
421250005	PA	Washington	40.1467N, 79.9022W	AQS	1/00 - 6/05	4/00 - 6/05[b]	1/00 - 6/05	1/00 - 6/05
421250200	PA	Washington	40.1706N, 80.2614W	AQS	1/00 - 6/05	4/00 - 6/05[b]		1/00 - 6/05
421255001	PA	Washington	40.4453N, 80.4208W	AQS	1/00 - 6/05	4/00 - 6/05[b]		1/00 - 6/05
421290006	PA	Westmoreland	40.4281N, 79.6931W	AQS		4/00 - 6/05[b]		
421290008	PA	Westmoreland	40.3047N, 79.5057W	AQS	1/00 - 6/05	4/00 - 6/05[b]	1/00 - 6/05	1/00 - 6/05
540090005	WV	Brooke	40.3381N, 80.5972W	AQS	1/00 - 6/05			
540090007	WV	Brooke	40.3901N, 80.5857W	AQS	1/00 - 6/05			
540290005	WV	Hancock	40.5291N, 80.5762W	AQS	1/00 - 6/05			
540290007	WV	Hancock	40.4602N, 80.5768W	AQS	1/00 - 6/05			
540290008	WV	Hancock	40.6157N, 80.5601W	AQS	1/00 - 6/05			
540290009	WV	Hancock	40.4274N, 80.5925W	AQS	1/00 - 6/05		1/00 - 6/05	
540290011	WV	Hancock	40.3945N, 80.6120W	AQS	1/00 - 6/05		1/00 - 6/05	
540290014	WV	Hancock	40.4355N, 80.6006W	AQS	1/00 - 12/03			
540290015	WV	Hancock	40.6183N, 80.5408W	AQS	1/00 - 6/05			
540290016	WV	Hancock	40.4119N, 80.6017W	AQS	1/00 - 7/04			
540291004	WV	Hancock	40.4215N, 80.5809W	AQS	1/00 - 6/05	4/00 - 6/05[b]	1/00 - 6/05	
540511002	WV	Marshall	39.9160N, 80.7341W	AQS	1/00 - 6/05			
540610003	WV	Monongalia	39.6494N, 79.9211W	AQS	1/00 - 6/05	4/00 - 6/05[b]		
540610004	WV	Monongalia	39.6331N, 79.9572W	AQS	1/00 - 12/01			
540610005	WV	Monongalia	39.6483N, 79.9578W	AQS	1/00 - 6/05			
540690007	WV	Ohio	40.1204N, 80.6993W	AQS	1/00 - 11/03	4/00 - 10/03[b]		
540690008	WV	Ohio	40.0638N, 80.7205W	AQS			1/00 - 12/04	
540690009	WV	Ohio	40.0688N, 80.7211W	AQS		4/04 - 10/04		
540690010	WV	Ohio	N/A	AQS		4/05 - 6/05		

[a]Dates shown indicate the approximate period of data collection (m/yy – m/yy); at the time of inventory, data had been reported for AQS sites through 5/05 or 6/05. [b]Data were collected only during ozone season (April – October). [c]No data reported 11/00 – 3/01, 11/01 – 4/02. [d]AQS data only reported during ozone season (April – October); additional non-ozone season data reported by UORVP in 2000-2001.

Nevertheless, there were 35 SO_2 monitors, 5 O_3 monitors, 13 CO monitors, and 10 NO_2 monitors in the 35-county region that collected data year-round for at least four years. (The low number of "important" O_3 monitoring sites results from the fact that many sites only measured ambient O_3 concentrations during ozone season, which runs from April through October). Among these, 15 of the SO_2 monitors, 2 of the O_3 monitors, 6 of the CO monitors, and 8 of the NO_2 monitors were sited in the 7-county Pittsburgh MSA, where much of the region's population is concentrated. Hence, data from these numerous "important" monitoring sites could be used (possibly in combination with spatial information derived from sites that generated less data) to estimate ambient gaseous pollutant concentrations for purposes of an epidemiology study.

$PM_{10\text{-}2.5}$ mass concentration data are also desired for inclusion in the proposed epidemiology study. In most cases, $PM_{10\text{-}2.5}$ concentrations were not measured directly, but must be estimated by differencing measured concentrations of PM_{10} and $PM_{2.5}$. The inventory of $PM_{2.5}$ mass concentration data collected in the 35-county greater Pittsburgh region between 2000 and 2005 was summarized in **Section 2.2.1**. **Table 6** and **Figure 9** indicate the locations of the monitoring sites that measured PM_{10} mass concentrations in the region during that time period. Ideally, spatially resolved $PM_{10\text{-}2.5}$ mass concentrations would be estimated from collocated PM_{10} and $PM_{2.5}$ measurements made at monitoring sites throughout the region. However, as shown in **Table 7**, which lists monitoring sites that simultaneously measured PM_{10} and $PM_{2.5}$ during 2000-2005, only one site in the region performed daily, collocated PM_{10} and $PM_{2.5}$ measurements for a period of four years or more. Moreover, this site, the AQS Liberty Borough monitoring station (420030064), is probably not well suited for representing the exposures of the larger region's population, because it is strongly affected by emissions from a large nearby coke production facility.

Table 6: Summary of PM_{10} total mass concentration data collected by monitoring sites in the 35-county greater Pittsburgh region between 2000 and 2005.

Site Name / ID	State	County	Latitude, Longitude	Program	Sampling Method	Approximate Sampling Period	Approximate Sampling Schedule
FRRE1	MD	Garrett	39.7058N 79.0122W	IMPROVE	IMPROVE	Apr 2004 - 2005	1 in 3 days
390131003	OH	Belmont	40.1064N 80.7097W	AQS	FRM Hi-Vol	Jan 2000 - Jan 2004	1 in 6 days
390290020	OH	Columbiana	40.6397N 80.5239W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 6 days
390290022	OH	Columbiana	40.6350N 80.5467W	AQS	FRM Hi-Vol	Jan 2001 - 2005	1 in 6 days
390810001	OH	Jefferson	40.2614N 80.6336W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 6 days
390810016	OH	Jefferson	40.3628N 80.6156W	AQS	FRM Hi-Vol	Jan 2000 - Oct 2003	1 in 6 days
390810017	OH	Jefferson	40.3661N 80.6150W	AQS	FRM Hi-Vol	Nov 2003 - 2005	1 in 6 days
390811001	OH	Jefferson	40.3219N 80.6064W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 3 days
Franciscan U. of Steubenville	OH	Jefferson	40.38N 80.62W	SCAMP	FRM	May 2000 - May 2002	Daily
390990005	OH	Mahoning	41.1111N 80.6453W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 6 days
390990006	OH	Mahoning	41.1167N 80.6697W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 6 days
391110001	OH	Monroe	39.7706N 80.8686W	AQS	FRM Hi-Vol	Jan 2000 - Jan 2004	1 in 6 days
QUCI1	OH	Noble	39.9428N 81.3378W	IMPROVE	IMPROVE	May 2001 - 2005	1 in 3 days
391550005	OH	Trumbull	41.2308N 80.8019W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 6 days
391550006	OH	Trumbull	41.2019N 80.8106W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 6 days
391550007	OH	Trumbull	41.2142N 80.7875W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 6 days
420030002	PA	Allegheny	40.5006N 80.0719W	AQS	FRM Hi-Vol TEOM	Jan 2000 - 2005 Jan 2000 - 2005	1 in 6 days Continuous
420030021	PA	Allegheny	40.4136N 79.9414W	AQS	TEOM	Jan 2000 - 2005	Continuous
420030027	PA	Allegheny	40.4383N 80.0689W	AQS	TEOM	Jan 2000 - Jul 2001	Continuous
420030031	PA	Allegheny	40.4433N	AQS	TEOM	Jan 2000 - 2005	Continuous

Site Name / ID	State	County	Latitude, Longitude	Program	Sampling Method	Approximate Sampling Period	Approximate Sampling Schedule
			79.9906W				
420030064	PA	Allegheny	40.3236N 79.8683W	AQS	FRM Hi-Vol TEOM	Jan 2000 - 2005 Jan 2000 - 2005	Daily Continuous
420030067	PA	Allegheny	40.3819N 80.1856W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 6 days
420030092	PA	Allegheny	40.4561N 80.0261W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 6 days
420030093	PA	Allegheny	40.6072N 80.0208W	AQS	FRM Hi-Vol	Jan 2000 - Dec 2000	1 in 6 days
420030095	PA	Allegheny	40.4869N 80.1881W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 6 days
420030097	PA	Allegheny	40.5531N 80.2033W	AQS	FRM Hi-Vol	Jan 2000 - Dec 2000	1 in 6 days
420030116	PA	Allegheny	40.4736N 80.0772W	AQS	FRM Hi-Vol TEOM	Jan 2000 - Dec 2000 Jan 2000 - 2005	1 in 6 days Continuous
420030133	PA	Allegheny	40.2601N 79.8865W	AQS	FRM Hi-Vol	Apr 2003 - Jun 2004	1 in 6 days
420031301	PA	Allegheny	40.4025N 79.8603W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 6 days
420032001	PA	Allegheny	40.3967N 79.8636W	AQS	FRM Hi-Vol TEOM	Jan 2000 - 2005 Jan 2000 - 2005	1 in 6 days Continuous
420033004	PA	Allegheny	40.3050N 79.8889W	AQS	TEOM	Jan 2000 - Jul 2001	Continuous
420033006	PA	Allegheny	40.3261N 79.8806W	AQS	FRM Hi-Vol TEOM	Jan 2000 - Dec 2000 Jan 2000 - 2005	1 in 6 days Continuous
420033007	PA	Allegheny	40.2944N 79.8867W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 6 days
420037004	PA	Allegheny	40.3081N 79.8703W	AQS	TEOM	Jan 2000 - 2005	Continuous
420039002	PA	Allegheny	40.5469N 79.7839W	AQS	FRM Hi-Vol	Jan 2000 - Dec 2000	1 in 6 days
Bruceton	PA	Allegheny	40.3065N 79.9794W	NETL/OST	Dichotomous	Aug 2002 - Sep 2004	Intermittent
Lawrenceville / PITT1	PA	Allegheny	40.4656N 79.9611W	UORVP UORVP IMPROVE	TEOM DRI SFS IMPROVE	Jun 1999 - Jul 2002 Feb 1999 - Aug 2001 Apr 2004 - 2005	Continuous Intermittent 1 in 3 days
Schenley Park	PA	Allegheny	40.4395N 79.9405W	PAQS	Dichotomous	Jun 2001 - Jul 2002	Daily
420070014	PA	Beaver	40.7478N 80.3167W	AQS	TEOM	Aug 2000 - 2005	Continuous
420210011	PA	Cambria	40.3097N	AQS	TEOM	Jul 2000 - 2005	Continuous

Site Name / ID	State	County	Latitude, Longitude	Program	Sampling Method	Approximate Sampling Period	Approximate Sampling Schedule
			78.9150W				
Holbrook	PA	Greene	39.8162N 80.2846W	UORVP	DRI SFS	Feb 1999 - Aug 2001	Intermittent
420730015	PA	Lawrence	40.9958N 80.3467W	AQS	TEOM	Aug 2000 - 2005	Continuous
MKGO1	PA	Mercer	41.4269N 80.1453W	IMPROVE	IMPROVE	Apr 2001 - 2005	1 in 3 days
421250005	PA	Washington	40.1467N 79.9022W	AQS	TEOM	Aug 2000 - 2005	Continuous
421255001	PA	Washington	40.4453N 80.4208W	AQS	FRM Hi-Vol	Aug 2000 - 2005	1 in 6 days
421290007	PA	Westmoreland	40.1667N 79.8750W	AQS	FRM Hi-Vol	Aug 2000 - 2005	1 in 6 days
421290008	PA	Westmoreland	40.3047N 79.5057W	AQS	TEOM	Aug 2000 - 2005	Continuous
540090005	WV	Brooke	40.3381N 80.5972W	AQS	FRM Hi-Vol	Jan 2000 - 2005	1 in 3 days
540290009	WV	Hancock	40.4274N 80.5925W	AQS	TEOM	Jan 2000 - 2005	Continuous
540290011	WV	Hancock	40.3945N 80.6120W	AQS	TEOM	Jan 2000 - 2005	Continuous
540290014	WV	Hancock	40.4355N 80.6006W	AQS	TEOM	Jan 2000 - Dec 2003	Continuous
540291004	WV	Hancock	40.4215N 80.5809W	AQS	FRM Hi-Vol TEOM	Jan 2000 - 2005 Jan 2000 - 2005	1 in 6 days Continuous
540490006	WV	Marion	39.4808N 80.1353W	AQS	FRM Hi-Vol	Jan 2000 - Dec 2000	1 in 6 days
540511002	WV	Marshall	39.9160N 80.7341W	AQS	FRM Hi-Vol	Jan 2000 - Mar 2004	1 in 6 days
540610003	WV	Monongalia	39.6494N 79.9211W	AQS	FRM Hi-Vol	Jan 2000 - Dec 2004	1 in 6 days
540690008	WV	Ohio	40.0638N 80.7205W	AQS	FRM Hi-Vol	Jan 2000 - Dec 2004	1 in 6 days

Table 7: Summary of monitoring sites in the 35-county greater Pittsburgh region from which PM$_{10-2.5}$ data could be obtained by differencing PM$_{10}$ and PM$_{2.5}$ mass concentrations measured between 2000 and 2005.

Site Name / ID	State	County	Latitude, Longitude	Program	Sampling Method (PM$_{10}$ / PM$_{2.5}$)	Approximate Period	Approximate Frequency
FRRE1	MD	Garrett	39.7058N 79.0122W	IMPROVE	IMPROVE / IMPROVE	Apr 2004 - 2005	1 in 3 days
390810016	OH	Jefferson	40.3628N 80.6156W	AQS	FRM Hi-Vol / FRM	Jan 2000 - Oct 2003	1 in 6 days
390810017	OH	Jefferson	40.3661N 80.6150W	AQS	FRM Hi-Vol / FRM	Nov 2003 - 2005	1 in 6 days
390811001	OH	Jefferson	40.3219N 80.6064W	AQS	FRM Hi-Vol / FRM	Jan 2000 - 2005	1 in 3 days (until 2/3/04), 1 in 6 days (2/3/04 - 2005)
Franciscan U. of Steubenville	OH	Jefferson	40.38N 80.62W	SCAMP	FRM / FRM	May 2000 - May 2002	Daily
390990005	OH	Mahoning	41.1111N 80.6453W	AQS	FRM Hi-Vol / FRM	Jan 2000 - 2005	1 in 6 days
QUCI1	OH	Noble	39.9428N 81.3378W	IMPROVE	IMPROVE / IMPROVE	May 2001 - 2005	1 in 3 days
391550007	OH	Trumbull	41.2142N 80.7875W	AQS	FRM Hi-Vol / FRM	Jan 2000 - 2005	1 in 6 days
420030021	PA	Allegheny	40.4136N 79.9414W	AQS	TEOM / FRM	Jan 2000 - 2005	1 in 3 days
420030064	PA	Allegheny	40.3236N 79.8683W	AQS	FRM Hi-Vol, TEOM / FRM, TEOM	Jan 2000 - 2005	Daily, Continuous
420030067	PA	Allegheny	40.3819N 80.1856W	AQS	FRM Hi-Vol / FRM	Jan 2000 - 2005	1 in 6 days
420030093	PA	Allegheny	40.6072N 80.0208W	AQS	FRM Hi-Vol / FRM	Jan 2000 - Dec 2000	1 in 6 days
420030095	PA	Allegheny	40.4869N 80.1881W	AQS	FRM Hi-Vol / FRM	Jan 2000 - 2005	1 in 6 days
420030097	PA	Allegheny	40.5531N 80.2033W	AQS	FRM Hi-Vol / FRM	Jan 2000 - Dec 2000	1 in 6 days
420030116	PA	Allegheny	40.4736N 80.0772W	AQS	TEOM / FRM	Jan 2000 - 2005	1 in 3 days
420030133	PA	Allegheny	40.2601N 79.8865W	AQS	FRM Hi-Vol / FRM	Apr 2003 - Jun 2004	1 in 6 days
420031301	PA	Allegheny	40.4025N 79.8603W	AQS	FRM Hi-Vol / FRM	Jan 2000 - 2005	1 in 6 days
420033007	PA	Allegheny	40.2944N 79.8867W	AQS	FRM Hi-Vol / FRM	Jan 2001 - 2005	1 in 6 days
420039002	PA	Allegheny	40.5469N 79.7839W	AQS	FRM Hi-Vol / FRM	Jan 2000 - Dec 2000	1 in 6 days

Site Name / ID	State	County	Latitude, Longitude	Program	Sampling Method (PM_{10} / $PM_{2.5}$)	Approximate Period	Approximate Frequency
Bruceton	PA	Allegheny	40.3065N 79.9794W	NETL/OST	Dichotomous / Dichotomous	Aug 2002 - Sep 2004	Intermittent
Lawrenceville, PITT1	PA	Allegheny	40.4656N 79.9611W	UORVP, IMPROVE	TEOM, SFS, IMPROVE / TEOM, SFS, IMPROVE	Jan 2000 - 2005	Intermittent (Continuous Jan 2000 – Jul 2002)
Schenley Park	PA	Allegheny	40.4395N 79.9405W	PAQS	Dichotomous / Dichotomous	Jun 2001 - Jul 2002	Daily
420070014	PA	Beaver	40.7478N 80.3167W	AQS	TEOM / FRM, TEOM	Aug 2000 - 2005	1 in 3 days (until Jun 2004), Continuous (Jul 2004 - 2005)
420210011	PA	Cambria	40.3097N 78.9150W	AQS	TEOM / FRM, TEOM	Jul 2000 - 2005	1 in 3 days (until Jul 2004), Continuous (Aug 2004 - 2005)
Holbrook	PA	Greene	39.8162N 80.2846W	UORVP	DRI SFS / FRM, SFS, TEOM	Jan 2000 - Aug 2001	Intermittent
MKGO1	PA	Mercer	41.4269N 80.1453W	IMPROVE	IMPROVE / IMPROVE	Apr 2001 - 2005	1 in 3 days
421250005	PA	Washington	40.1467N 79.9022W	AQS	TEOM / FRM	Aug 2000 - 2005	1 in 3 days
421255001	PA	Washington	40.4453N 80.4208W	AQS	FRM Hi-Vol / FRM, FEM	Aug 2000 - 2005	1 in 6 days
421290008	PA	Westmoreland	40.3047N 79.5057W	AQS	TEOM / FRM, FEM	Aug 2000 - 2005	1 in 3 days
540090005	WV	Brooke	40.3381N 80.5972W	AQS	FRM Hi-Vol / FRM	Jan 2000 - 2005	1 in 3 days
540290011	WV	Hancock	40.3945N 80.6120W	AQS	TEOM / FRM	Jan 2000 - 2005	1 in 3 days
540291004	WV	Hancock	40.4215N 80.5809W	AQS	FRM Hi-Vol, TEOM / FRM	Jan 2000 - 2005	1 in 3 days
540490006	WV	Marion	39.4808N 80.1353W	AQS	FRM Hi-Vol / FRM	Jan 2000 - Dec 2000	1 in 6 days
540511002	WV	Marshall	39.9160N 80.7341W	AQS	FRM Hi-Vol / FRM	Jan 2000 - Mar 2004	1 in 6 days
540610003	WV	Monongalia	39.6494N 79.9211W	AQS	FRM Hi-Vol / FRM	Jan 2000 - Dec 2004	1 in 6 days
540690008	WV	Ohio	40.0638N 80.7205W	AQS	FRM Hi-Vol / FRM	Jan 2000 - Dec 2004	1 in 6 days

Hence, better exposure estimates may be derived by first individually modeling the spatially resolved daily $PM_{2.5}$ and PM_{10} concentrations measured throughout the region in order to produce regional estimates for the daily concentrations of each of these species, and then differencing the results to arrive at regional daily $PM_{10-2.5}$ concentration estimates. This approach allows the region's available air monitoring information to be more fully utilized, as it does not exclude data from sites that measured only PM_{10} or $PM_{2.5}$ but not both. Estimates computed in this way could then be validated against estimates derived from collocated PM_{10} and $PM_{2.5}$ measurements using days for which data are available. As discussed in **Section 2.2.1** and shown in **Table 6** there were nine sites in the 35-county region that measured $PM_{2.5}$ mass concentrations on a daily basis for at least four years between 2000 and 2005, and there were 16 sites in the region that measured PM_{10} mass concentrations for at least four years during that period. These sites would be of primary importance for estimating $PM_{10-2.5}$ concentrations according to the procedure above.

2.2.4 Meteorological Data

Numerous sites in the 35-county greater Pittsburgh region collected temperature data, relative humidity or dew point data, and wind speed and direction data between 1999 and 2005. Per the discussion below, all of these data are available for use in the proposed epidemiology study.

The 13 ASOS/AWOS sites located at airports throughout the region are probably the best source of meteorological data during 1999-2005, as hourly weather observations including temperature, dew point, and wind speed and direction were routinely collected by all of these sites during the entire period. These observations, which are made according to standard protocols for use by the National Weather Service and Federal Aviation Administration, are available from the NCDC or Pennsylvania MESONET, as discussed earlier. **Table 8** indicates the locations of the ASOS / AWOS sites in the region.

Table 8: ASOS / AWOS weather stations in the 35-county greater Pittsburgh region from which hourly data are available during 1999-2005.

Call Sign	Name	State	County	Latitude	Longitude
KYNG	YOUNGSTOWN REGIONAL AIRPORT	OH	Trumbull	41.25	-80.667
KPIT	PITTSBURGH INTERNATIONAL AP	PA	Allegheny	40.5	-80.233
KAGC	PITTSBURGH ALLEGHENY CO AP	PA	Allegheny	40.35	-79.917
KBVI	BEAVER FALLS ARPT	PA	Beaver	40.767	-80.4
KBTP	BUTLER CO. (AWOS)	PA	Butler	40.783	-79.95
KJST	JOHNSTOWN CAMBRIA COUNTY AP	PA	Cambria	40.317	-78.833
KIDI	INDIANA/STEWART FLD	PA	Indiana	40.633	-79.1
KDUJ	DUBOIS FAA AP	PA	Jefferson	41.183	-78.9
KFKL	FRANKLIN	PA	Venango	41.383	-79.867
KAFJ	WASHINGTON (AWOS)	PA	Washington	40.133	-80.283
KLBE	ARNOLD PALMER RGNL	PA	Westmoreland	40.267	-79.4
KMGW	MORGANTOWN HART FIELD	WV	Monongalia	39.65	-79.917

Call Sign	Name	State	County	Latitude	Longitude
KHLG	WHEELING OHIO COUNTY AP	WV	Brooke	40.183	-80.65

Thirty-five RWIS sites in western Pennsylvania also monitored meteorological conditions, including temperature, wind speed and direction, relative humidity, and dew point, between 1999 and 2005. These sites, which are operated by PennDOT, are identified in **Table 9**. As discussed above, RWIS data collected between late 2001 and the present are available from the Pennsylvania MESONET.

In addition to the ASOS / AWOS and RWIS weather stations, many of the sites that measured $PM_{2.5}$ or co-pollutant concentrations in the 35-county region also continuously monitored surface meteorological conditions. The Allegheny County Health Department monitored temperature and wind speed at its Avalon (420030002), Hazelwood (420030021), Liberty (420030064), and South Fayette (420030067) AQS monitoring sites throughout the time period of interest. There are, however, large periods of missing data at Avalon during August 2000 – September 2000, November 2002 – December 2002, and January 2003 – May 2003; at Hazelwood during August 1999 – December 2002 and January 2003 – May 2003; at Liberty during January – September 1999; and at South Fayette during 2003 – 2004 (for temperature). ACHD also monitored temperature and wind speed at Glassport during 1999 – 2001, and at Clairton, North Braddock, and a second site in Hazelwood during 1999 – 2000.

Like ACHD, the Pennsylvania DEP measured temperature and wind speed at its ambient air monitoring sites in western Pennsylvania during the time period of interest. These sites include Pittsburgh in Allegheny County; Kittanning in Armstrong County; Beaver Falls, Brighton Township, and Hookstown in Beaver County; Johnstown in Cambria County; Holbrook in Greene County; New Castle in Lawrence County; Farrell in Mercer County; Charleroi, Florence, and Washington in Washington County; and Greensburg and Murrysville in Westmoreland County. Data collected at these sites from June 27, 2001, through the present are available from the Pennsylvania MESONET.

The NETL/OST Bruceton site, PAQS Schenley site, UORVP Lawrenceville site, and SCAMP Franciscan University of Steubenville site each included surface weather stations that continuously monitored a suite of meteorological conditions. Temperature, relative humidity, and wind speed and direction data were collected at the Bruceton site between April 2000 and June 2004, at the Schenley site between July 2001 and September 2002, at the Franciscan University site between May 2000 and May 2002 (wind direction data are invalid during entire period), and at the Lawrenceville site between January 2000 and December 2002 (relative humidity monitoring began in August 2000).

Table 9: RWIS weather stations in western Pennsylvania for which hourly data are available from the Pennsylvania MESONET.

Site	County	Lat.	Lon.	Start Date[a]
I-79 EXIT 060	ALLEGHENY	40.447	-80.11	12/1/2001
SR 0060 @ BEAVER - ALLEGHENY CO LINE HOPEWELL	ALLEGHENY	40.55	-80.276	12/1/2001

Site	County	Lat.	Lon.	Start Date[a]
I-376 EXIT 10A @ CHURCHILL EXIT	ALLEGHENY	40.442	-79.827	12/1/2001
SR0028 @ TARENTUM	ALLEGHENY	40.653	-79.725	12/1/2001
SR 0028 @ SOUTH OF DISTANT	ARMSTRONG	40.933	-79.362	6/27/2001
SR 0060 @ SR 0051 CHIPPEWA	BEAVER	40.741	-80.371	12/1/2001
I-79 EXIT 088	BUTLER	40.792	-80.125	12/1/2001
SR 0022 @ CRESSON MOUNTAIN	CAMBRIA	40.461	-78.565	12/1/2001
SR 0022 @ CHICKORY MOUNTAIN	CAMBRIA	40.437	-78.906	12/1/2001
I-80 EXIT 053 @ MP 55	CLARION	41.191	-79.514	12/1/2001
SR 0119 @ UNIONTOWN	FAYETTE	39.949	-79.653	12/1/2001
SR 0043 @ SMITHFIELD	FAYETTE	39.821	-79.774	12/1/2001
SR 0040 @ SUMMIT MT.	FAYETTE	39.85	-79.659	12/1/2001
SR 0653 @ LAURAL HILL	FAYETTE	39.956	-79.364	12/1/2001[b]
I-79 EXIT 002 @ WELCOME CENTER	GREEN	39.795	-80.076	12/1/2001
SR 0018 @ NETTLE HILL	GREEN	39.798	-80.38	12/1/2001
SR 0022 @ EAST BLAIRSVILLE	INDIANA	40.45	-79.157	12/1/2001
SR 0422 @ PENN RUN	INDIANA	40.606	-79.047	12/1/2001
I-80 EXIT 097 @ ROADSIDE REST	JEFFERSON	41.152	-78.91	12/1/2001
SR 0060 @ SR 0224 UNION TWP	LAWRENCE	41.011	-80.398	12/1/2001
I-80 EXIT 015 @ MP 11	MERCER	41.194	-80.306	9/1/2000
I-80 EXIT 019 @ I-79 JUNCT.	MERCER	41.197	-80.161	12/3/2001
I-79 EXIT 130	MERCER	41.481	-80.166	12/1/2001
SR 0031 @ LARUAL RIDGE	SOMERSET	40.067	-79.266	12/1/2001[c]
SR 0056 @ BABCOCK MT.	SOMERSET	40.194	-78.685	7/22/2001
SR 0219 @ JEROME	SOMERSET	40.196	-78.976	12/1/2001
SR 0219 @ MEYERSDALE BYPASS	SOMERSET	39.816	-79.038	12/1/2001
SR 0008 @ SR 0308	VENANGO	41.267	-79.924	9/26/2001
I-80 EXIT 035 @ MP 37.5	VENANGO	41.196	-79.82	12/3/2001
SR 0027 @ PLEASANTVILLE	VENANGO	41.602	-79.609	12/3/2001
SR0062 @ PRESIDENT	VENANGO	41.449	-79.577	11/14/2001
SR0322 @ VENANGO-MERCER CO LINE	VENANGO	41.481	-79.994	12/1/2001
I-70 EXIT 002 @ WELCOME CENTER	WASHINGTON	40.116	-80.442	12/1/2001
SR 0022 @ STAR LAKE	WASHINGTON	40.426	-80.429	12/1/2001
SR 0030 @ JACKTOWN HILL	WESTMORELAND	40.329	-79.734	11/28/2001

[a]Data available through the end of 2005 except where indicated; [b]Data available through 8/24/02; [c]Data available through 11/12/04.

Finally, temperature, relative humidity, and wind speed and direction were also continuously monitored at the CastNET's Quaker City (QAK172), M.K. Goddard (MKG113), and Laurel Hill (LRL117) sites throughout the 1999 – 2005 time period. These data are available from the CASTNet website.

2.3 Inventory of Archived Filter-Based $PM_{2.5}$ Samples

Many of the monitoring sites in the Pittsburgh region that determined ambient air concentrations of $PM_{2.5}$ chemical components between 1999 and 2005 also collected additional filter-based $PM_{2.5}$ samples that were not analyzed for chemical composition but have been archived and would be available for analysis. Determination of the chemical composition of these samples, where feasible, would substantially augment the speciated $PM_{2.5}$ data record for the Pittsburgh region during the time period of interest for the proposed epidemiology study. Hence, in addition to the inventory of existing air monitoring data discussed in **Section 2.2**, an inventory of archived $PM_{2.5}$ filter-based samples available from the Pittsburgh region between 1999 and 2005 was completed as part of the current feasibility assessment.

The feasibility of obtaining $PM_{2.5}$ chemical composition data from archived $PM_{2.5}$ samples depends on a number of factors, including the method originally used to sample the particles, the type of filter on which the samples were collected, and the manner in which the samples were stored following collection. Most of the archived $PM_{2.5}$ samples available from monitoring sites in the Pittsburgh region are 24-hour integrated samples that were collected on Teflon filters according to the Federal Reference Method for $PM_{2.5}$. As described by Chow and Watson (1998), archived Teflon-filter-based $PM_{2.5}$ samples can be analyzed to determine trace and crustal elements and inorganic ions via a two-step process. Trace and crustal elements are first determined by X-ray fluorescence spectroscopy (XRF) or proton induced X-ray emission spectroscopy (PIXE), which are nondestructive methods capable of determining elements with atomic numbers between 11 (sodium) and 92 (uranium). Inorganic ions (e.g., SO_4^{2-}, NO_3^-, NH_4^+, Na^+, K^+) are then determined by ion chromatography (IC), a destructive technique that requires extraction of the sample in deionized water (containing a small amount of ethanol as a wetting agent). Hence, this two-step process can determine most of the $PM_{2.5}$ chemical components of interest that were identified in **Section 2.2** above.

There are several important limitations, however. Concentrations of semi-volatile $PM_{2.5}$ components, such as NO_3^- and NH_4^+, may be underestimated as a result of artifacts arising from sampling, storage, and analytical procedures. NO_3^- and NH_4^+ concentrations determined from Teflon-filter-based samples collected using FRM monitors are often biased low relative to concentrations determined using speciation samplers (e.g., employing nylon filters) because of losses of volatile NH_4NO_3 from the FRM samples (Jansen et al., 2002; Connell et al., 2005a; Frank, 2006). Loss of semi-volatile material is also likely if filters are stored at room temperature rather than under refrigeration. Finally, if archived FRM samples are analyzed according to the two-step process described above, volatile compounds can be lost under the vacuum conditions used for XRF or PIXE analysis (Chow and Watson, 1998), resulting in underestimation of their concentrations when these are subsequently determined. Hence, analysis of archived FRM samples for semi-volatile species such as NO_3^- and NH_4^+ is likely to be feasible only if the samples were stored under refrigeration and if some collocated speciation sampling data are available to allow biases to be corrected.

Moreover, because Teflon is itself a carbonaceous material, elemental and organic carbon cannot be determined from Teflon-filter-based samples using the thermal/optical techniques (e.g., Thermal Optical Transmittance [TOT] and Thermal Optical Reflectance [TOR]) commonly applied to samples collected on quartz-fiber filters. Nondestructive light transmission measurements can be used to approximate elemental carbon concentrations in samples collected on Teflon filters, but variations in the filter loading, in the chemical and physical nature of the samples, and in the details of the method being used can introduce substantial error in the results (Chow and Watson, 1998). In addition, transmission of the blank filter is typically measured prior to sampling to enable background correction; this cannot be accomplished with archived filters. Hence, ambient concentrations of fine particulate elemental and organic carbon in the Pittsburgh region likely cannot be derived from archived Teflon-filter-based samples; other sources must be relied upon for these data.

Elemental and organic carbon can be determined from archived samples that were collected on quartz-fiber filters, provided that these samples were stored under refrigeration to prevent losses of semi-volatile organic material. Inorganic ions can also be determined from quartz filters via IC. If both inorganic ions and carbon species are to be determined, a filter punch is first taken for analysis by TOT or TOR, and the remaining area of the filter is extracted for analysis by IC. Trace and crustal elements generally are not determined from quartz-fiber filters, however, because these filters contain high and variable blank concentrations of a number of elements (e.g., Al, Si, S, Cl, K, Ca, Fe, Ni, Cu, Zn, Ba, Pb). Moreover, XRF is subject to biases and decreased sensitivity when applied to quartz-filter-based samples because X-rays are absorbed within the filter fibers and scattered by the relatively thick quartz filter media (Chow and Watson, 1998).

Archived $PM_{2.5}$ samples from sites that already have some preexisting $PM_{2.5}$ chemical speciation data are of most interest for the proposed retrospective epidemiology study, because the existing speciation data can be used to verify the quality of results obtained from the archived filter analyses. For example, several AQS monitoring sites in the Pittsburgh region determined $PM_{2.5}$ speciation from samples collected using a $PM_{2.5}$ speciation sampler on a 1-in-3 or 1-in-6 day frequency and also collected $PM_{2.5}$ samples at a greater frequency (i.e., 1-in-1 or 1-in-3 day) using a Federal Reference Method sampler. Hence, the validity of chemical component concentrations determined from these archived FRM samples, which would be used to provide $PM_{2.5}$ speciation information for days on which the speciation sampler did not operate, could be confirmed by analyzing a subset of FRM samples that were collected on days when the speciation sampler did operate and comparing the results. Calibration curves could also be developed from these comparisons, if necessary, to correct any biases between concentrations determined from the original speciation samples and concentrations determined from the archived FRM samples. **Section 2.4.2** of this report presents such a comparison using pairs of collocated $PM_{2.5}$ samples from the Bruceton monitoring site, each including one sample that was analyzed for ions (SO_4^{2-}, NO_3^-, NH_4^+) soon after collection and one sample that was analyzed for ions after several years of refrigerated storage, and demonstrates how calibration curves can be used to adjust for relative biases among these data prior to utilizing them in epidemiological models.

The inventory of existing $PM_{2.5}$ speciation data presented in Table 4 in **Section 2.2.2** indicates that 71% of the existing fine particulate sulfate, nitrate, and elemental data (quantified according to the number of site-days with available data) and 77% of the existing fine particulate carbonaceous data available from the 35-county greater Pittsburgh region between 1999 and 2005 were collected at monitoring sites located in western Pennsylvania. The 16 western Pennsylvania counties listed in **Table 1** also account for 72% of the population of the 35-county region considered in **Section 2.2**. Hence, because much of the region's population and ambient $PM_{2.5}$ speciation data are concentrated in western Pennsylvania, and because of the desirability of obtaining archived $PM_{2.5}$ samples from sites with preexisting $PM_{2.5}$ speciation data that are likely to be representative of the exposures of the region's population, the inventory of archived filter-based $PM_{2.5}$ samples was performed only for monitoring sites located in western Pennsylvania that performed some $PM_{2.5}$ speciation sampling between 1999 and 2005.

Per the data inventory results presented in **Section 2.2**, the monitoring sites in western Pennsylvania for which archived filter-based $PM_{2.5}$ samples could be used to supplement existing $PM_{2.5}$ speciation data include the ACHD sites at Lawrenceville, Liberty, and Hazelwood, the PA DEP sites at Florence and Greensburg, the NETL/OST site at Bruceton, the UORVP sites at Lawrenceville and Holbrook, and the SCAMP site at St. Vincent College. Archived $PM_{2.5}$ samples are available from the PAQS Schenley Park site as well, although these samples would add little to the already extensive daily record of basic $PM_{2.5}$ speciation data from that site (Allen Robinson, Carnegie Mellon University, personal communication on 1/31/06). For the remaining sites in western Pennsylvania that have preexisting $PM_{2.5}$ speciation data from the time period of interest, which were operated by the CASTNet and IMPROVE networks, all valid $PM_{2.5}$ samples have already been analyzed for chemical composition.

The monitoring groups that operated the candidate sites identified above were contacted to confirm the availability of archived filter-based samples and to determine the feasibility of obtaining these samples if required for use in a future epidemiology study. A day-by-day inventory of archived filter-based samples was assembled for each candidate site, although the meticulousness of the inventorying procedures varied from site-to-site based on the preferences of the group in custody of the samples, the availability of preexisting records regarding the contents of the inventory, and budgetary limitations imposed by the scope of the current feasibility assessment. Inventories for each candidate site were conducted according to one of the following three procedures:

1. Detailed physical inventory including individual identification of each archived filter-based sample.

2. Identification of archived filter-based samples based on a review of database or laboratory records provided by the group in custody of the samples, supplemented by a physical inventory including a total filter count and random spot checks to verify the accuracy of the database / laboratory records.

3. Identification of archived filter-based samples based on a review of database or laboratory records provided by the group in custody of the samples, supplemented by discussions with that group to confirm sample archiving procedures.

The inventories of archived filter-based $PM_{2.5}$ samples were generally conducted in accordance with the checklist that is included in **Appendix A** to this report. As with the results of the inventory of existing

$PM_{2.5}$ speciation data, the results of the inventory of archived $PM_{2.5}$ samples were logged in a database that uses codes of "1" (sample available) and "0" (no sample available) to indicate archived sample availability for each monitoring site with a daily time resolution. In light of the considerations discussed above regarding the effects of sampling and archiving methodologies on the utility of the archived samples, the database also includes fields and sub-tables housing information about the methods used to collect and store each sample. A diagram of the database design is provided as **Appendix D**. The database, named "AvailableFilters," is included on the CD accompanying this report.

On certain days, UORVP sampling at the Lawrenceville site included collection of four 6-hour filter-based $PM_{2.5}$ samples rather than a single 24-hour sample. For these days, a sample was considered to be available (i.e., a "1" was assigned) only if all four filter-based samples were available and valid (as required to satisfy the ≥19-hour data completeness criterion established in **Section 2.2** above). Archived filter-based $PM_{2.5}$ samples were considered to be invalid only if they were qualified as such according to the QA/QC procedures followed by the group responsible for collecting the samples (e.g., for FRM samples, if the reported $PM_{2.5}$ mass measurements were marked as invalid, then the archived samples were likewise considered to be invalid). Samples that were "flagged" but not marked as invalid were considered to be valid for purposes of this inventory. QA/QC procedures followed by the various monitoring programs from which archived filter-based $PM_{2.5}$ samples are available are discussed in **Section 2.4.3** of this report.

Table 10 summarizes the results of the inventory of archived filter-based $PM_{2.5}$ samples that are available from sites in western Pennsylvania that monitored for $PM_{2.5}$ speciation between 1999 and 2005. Results are stratified by site and filter type; the approximate schedule according to which the archived samples were collected and the general method used to store the samples (refrigeration or no refrigeration) are also indicated for each stratum.

Table 10: Estimate of the number of days with archived filter-based $PM_{2.5}$ samples for sites in southwestern Pennsylvania that collected $PM_{2.5}$ speciation data between 1999 and 2005.[a]

Site	County, State	Latitude, Longitude	Program	Filter Type	Refrigerated?[b]	Approximate Period of Archived $PM_{2.5}$ Sample Availability[c]	Approximate Frequency of Archived $PM_{2.5}$ Sample Availability	Number of Days with ≥ 1 Archived $PM_{2.5}$ Sample
Bruceton (BRU)	Allegheny, PA	40.3065N 79.9794W	NETL/OST	Teflon	Yes	07/19/99 – 06/06/04	Daily	1168
Florence (FLO)	Washington, PA	40.4453N 80.4208W	AQS	Teflon	No	01/01/01 – 03/31/05	Daily	1343
Greensburg (GRE)	Westmoreland, PA	40.3047N 79.5057W	AQS	Teflon	No	01/01/01 – 03/29/05	1 in 3 days	490
Hazelwood (HAZ)	Allegheny, PA	40.4136N 79.9414W	AQS	Teflon	No	01/01/00 – 03/26/05	1 in 3 days	566
Holbrook (HOL)	Greene, PA	39.8162N 80.2846W	UORVP UORVP	Teflon Quartz	Yes Yes	02/19/99 – 01/22/02 02/19/99 – 01/22/02	Intermittent Intermittent	240 238
Lawrenceville (LAW)	Allegheny, PA	40.4656N 79.9611W	AQS, UORVP UORVP	Teflon Quartz	Some Yes	02/19/99 – 03/31/05 02/19/99 – 01/22/02	Intermittent[d] Intermittent	1825 227
Liberty (LIB)	Allegheny, PA	40.3236N 79.8683W	AQS	Teflon	No	01/04/00 – 03/31/05	Daily	1817
St. Vincent College (STV)	Westmoreland, PA	40.29N 79.40W	SCAMP	Teflon	Yes	05/13/00 – 05/13/02	Daily[e]	490

[a] Methods for inventorying archived $PM_{2.5}$ samples differed by site; see text for description. [b] See text for explanation. [c] Inventory data for the AQS sites were only available through March 2005; additional filter-based $PM_{2.5}$ samples have been collected at these sites and archived since that time. [d] Teflon filters from intermittent UORVP sampling are available from 2/19/99 – 1/22/02; Teflon filters from daily AQS sampling are available from 1/13/00 – 3/31/05. [e] Filter from every fourth day has already been consumed for ionic and water-soluble elemental analysis.

The Allegheny County Health Department is the largest source of archived filter-based $PM_{2.5}$ samples from the Pittsburgh region during the time period of interest. Teflon-filter-based $PM_{2.5}$ samples are collected using FRM samplers on a daily basis at ACHD's Lawrenceville and Liberty Borough monitoring sites and on a 1-in-3 day basis at ACHD's Hazelwood monitoring site. These samples, which are being stored at the ACHD Air Quality Program offices in Lawrenceville, were inventoried by reviewing AQS data and other database records provided by ACHD (i.e., according to procedure #3 above). In addition, a site visit was conducted to review filter storage procedures. The inventory indicated that there are approximately 1,732 Teflon-filter-based $PM_{2.5}$ samples available from the Lawrenceville site, 1,817 Teflon-filter-based samples available from the Liberty site, and 566 Teflon-filter-based samples available from the Hazelwood site that were collected between January 2000 and March 2005 and could be analyzed to provide some $PM_{2.5}$ compositional data. (The inventory only covered the period through March 2005; however, additional $PM_{2.5}$ samples collected since then would also be available). Because all of the archived $PM_{2.5}$ samples available from ACHD were collected on Teflon filters, ambient elemental and organic carbon concentrations likely cannot be determined from these samples. However, the samples are candidates for ionic and elemental analysis. ACHD stores its filter-based samples in Petri slides that are kept in a freezer for about two years after collection and then transferred to cardboard boxes for long-term storage at ambient temperature. For purposes of the inventory, it was assumed that by the time the proposed retrospective epidemiology study would begin, all $PM_{2.5}$ samples collected by ACHD through March 2005 will have been transferred out of refrigerated storage. Hence, these filters would be subject to losses of semi-volatile material and probably could not be used to obtain reliable nitrate concentration data. The Lawrenceville, Liberty Borough, and Hazelwood monitoring sites each have a number of days from which both existing $PM_{2.5}$ ionic and elemental speciation data (obtained using a speciation sampler) and an archived Teflon-filter-based $PM_{2.5}$ sample are available. There were 552 such days at the Lawrenceville site, 71 such days at the Liberty site, and 82 such days at the Hazelwood site between January 2000 and March 2005. As discussed above, analysis of the chemical composition of the archived $PM_{2.5}$ samples from these days, although not required to fill in gaps in the existing record of $PM_{2.5}$ speciation data, is nevertheless recommended as a means for verifying the accuracy of speciation data determined from the archived $PM_{2.5}$ samples (i.e., by pairwise comparison with the existing data) and for developing calibrations to correct any biases resulting from long-term storage or from differences in sampling and analytical techniques.

Per discussions with ACHD's Air Quality Program personnel, the $PM_{2.5}$ samples being archived at ACHD could be obtained for use in a retrospective epidemiology study, provided that written permission was first obtained from ACHD's Director. ACHD would not release samples until three months after final weighing, and would retain the right to keep selected samples that yielded outlying $PM_{2.5}$ mass results. Data generated from the analysis of archived samples would have to be promptly reported to ACHD in a well-organized format, and details of the methods to be used would have to be reviewed and approved by ACHD prior to analysis. ACHD would not object to the use of destructive methods for sample analysis as long as the results produced by these methods would be of value.

Additional $PM_{2.5}$ samples from the Lawrenceville site were collected as part of the UORVP sampling campaign. All filter-based $PM_{2.5}$ samples collected by UORVP, including those from the Holbrook site in Greene County as well as those from the Lawrenceville site, are being stored under refrigeration at the Desert Research Institute in Reno, Nevada, and would be available for analysis if required for use in a retrospective epidemiology study (Robin Khosah, ATS Chester Engineers, personal communication on 6/19/06). Archived samples from UORVP, which include samples collected on both quartz and Teflon filters using a combination of FRM samplers and Sequential Filter Samplers, were inventoried using records provided by Steven Kohl of Desert Research Institute (i.e., according to procedure #3 above). Results indicate that, between February 1999 and January 2002, there are 230 days for which archived Teflon-filter-based $PM_{2.5}$ samples are available from UORVP sampling at the Lawrenceville site (including 95 days with four 6-hour samples rather than one 24-hour sample and 76 days with duplicate 24-hour samples), 227 days for which archived quartz-filter-based $PM_{2.5}$ samples are available from the Lawrenceville site (including 92 days with four 6-hour samples rather than one 24-hour sample and 62 days with duplicate 24-hour samples), 240 days for which archived Teflon-filter-based $PM_{2.5}$ samples are available from the Holbrook site (including 70 days with duplicate samples), and 238 days for which archived quartz-filter-based $PM_{2.5}$ samples are available from the Holbrook site (including 41 days with duplicate samples). Although the UORVP $PM_{2.5}$ samples were only collected intermittently (either on a 1-in-6 day frequency or on a daily frequency during short sampling intensives), these samples are nevertheless valuable. In particular, the refrigerated quartz-filter-based samples allow ambient elemental and organic carbon concentrations to be determined for about 230 days at an urban site (Lawrenceville) and a rural site (Holbrook) in the Pittsburgh region. Ambient nitrate concentrations can also be determined from the refrigerated quartz or Teflon-filter-based samples. Teflon-filter-based $PM_{2.5}$ samples collected by UORVP at Lawrenceville, when combined with samples collected there by ACHD, increase the number of days with at least one Teflon-filter-based $PM_{2.5}$ sample from that site to 1,825. Again, the validity of results determined from archived UORVP samples collected at Lawrenceville can be verified by comparison with existing $PM_{2.5}$ speciation data from that site. Duplicate archived filter-based $PM_{2.5}$ samples that are available on a number of days from each of the Lawrenceville and Holbrook sites provide further means for quality controlling the results of archived filter analyses.

The NETL/OST Bruceton monitoring site is the largest source of refrigerated archived Teflon-filter-based $PM_{2.5}$ samples from Allegheny County during the time period of interest. Most of the archived $PM_{2.5}$ samples from the Bruceton site were collected using FRM monitors. All samples from the site are being stored in Petri slides under refrigeration at the NETL facility in Bruceton and would be available for destructive or nondestructive analysis if required for use in a retrospective epidemiology study. An inventory of archived Teflon-filter-based $PM_{2.5}$ samples collected at the Bruceton site was performed using logbooks provided by Don Martello of NETL; inventory results were then confirmed by physically counting the total number of Teflon filters in storage at NETL and by spot-checking random batches of filters to confirm identification numbers (i.e., according to procedure #2 above). The inventory based on logbook records agreed with the total filter count to within 3%. As shown in **Table 10**, there are about

1,168 days for which at least one Teflon-filter-based $PM_{2.5}$ sample is available from the Bruceton monitoring site. Per the data presented earlier in Table 4, a major strength of the Bruceton monitoring site was its collection of semi-continuous EC, OC, SO_4^{2-}, and NO_3^- data for multiple years during the time period of interest. Hence, the archived $PM_{2.5}$ samples from the Bruceton monitoring site, if analyzed to provide elemental data and additional SO_4^{2-} and NO_3^- data to supplement the existing semi-continuous $PM_{2.5}$ compositional data, would enable the construction of a multiple-year stream of complete daily $PM_{2.5}$ speciation measurements as required for use in the proposed epidemiology study.

Archived Teflon-filter-based $PM_{2.5}$ samples from Washington and Westmoreland Counties are available from the Pennsylvania Department of Environmental Protection and from the Steubenville Comprehensive Air Monitoring Program. Teflon-filter-based $PM_{2.5}$ samples are collected using FRM samplers on a daily basis at the PA DEP monitoring site in Florence, Washington County, and on a 1-in-3 day basis at the PA DEP monitoring site in Greensburg, Westmoreland County. Samples collected since 2001 are currently being stored at the Pennsylvania State Archives in Harrisburg, PA; however, the oldest of these samples are scheduled to be discarded in April 2007 (George Mentzer, PA DEP, personal communication on 9/19/06). Only samples from the most recent year are kept under refrigeration. Archived $PM_{2.5}$ samples available from the PA DEP were inventoried by reviewing AQS records (procedure #3 above). The inventory indicated that there are approximately 1,343 Teflon-filter-based $PM_{2.5}$ samples available from the Florence site and 490 Teflon-filter-based samples available from the Greensburg site during the January 2001 – March 2005 period that could be analyzed to provide some $PM_{2.5}$ compositional data. (The inventory only covered the period through March 2005; however, additional $PM_{2.5}$ samples collected since then would also be available). Because all of the samples from this period were collected on Teflon filters and are now being stored at room temperature, ambient concentrations of elemental carbon, organic carbon, and nitrate probably cannot be determined from them. However, trace and crustal elements and certain ionic species could likely be determined. As with the ACHD monitoring sites, the PA DEP Florence and Greensburg sites each have a number of days from which both an archived Teflon-filter-based $PM_{2.5}$ sample and existing $PM_{2.5}$ ionic and elemental speciation data are available. There were 210 such days at the Florence site and 195 such days at the Greensburg site between January 2001 and March 2005. Again, analysis of the chemical composition of the archived $PM_{2.5}$ samples from these days for pairwise comparison with the existing speciation data is recommended as a means for verifying the accuracy of speciation data obtained from the archived samples.

Teflon-filter-based $PM_{2.5}$ samples collected between 2000 and 2002 at the SCAMP monitoring site on the campus of St. Vincent College in Latrobe, Westmoreland County, Pennsylvania, are also available. These samples, which were obtained using a $PM_{2.5}$ FRM sampler, are being stored in Petri dishes under refrigeration at the CONSOL Energy Inc. Research and Development facilities in South Park, PA. Filters were inventoried according to procedure #1 above (i.e., physical inventory including individual identification of each archived filter-based sample). As shown in **Table 10**, there are 490 days between May 13, 2000, and May 13, 2002, from which an archived Teflon-filter-based $PM_{2.5}$ sample is available from the St. Vincent College site. $PM_{2.5}$ samples were collected on an approximately daily basis at the site during this two-year period, although samples from every fourth day were consumed for ionic and water-soluble elemental analyses as part of SCAMP and therefore are not available for any further analyses. The St. Vincent College site was equipped with only a single FRM monitor; hence, there are no cases in which both existing $PM_{2.5}$ speciation data and an archived $PM_{2.5}$ sample from the same day are available for pairwise comparison. The accuracy of $PM_{2.5}$ chemical composition results obtained by analyzing archived $PM_{2.5}$ samples from the St. Vincent College site would therefore have to be confirmed by other means (e.g., by pairwise comparison with same-day speciation data from a nearby monitoring site such as the Greensburg site).

Hence, greater than 8,400 archived filter-based $PM_{2.5}$ samples collected by monitoring sites in western Pennsylvania between 1999 and 2005 are available for compositional analysis. (This number does not include field blanks, duplicate samples, and samples collected by monitoring sites that did not have any preexisting $PM_{2.5}$ speciation data). In many cases, these samples provide the only means for obtaining any retrospective $PM_{2.5}$ chemical composition data for a given monitoring site on a given day. **Figure 10** presents a time line showing the days from which archived filter-based $PM_{2.5}$ samples are available for the monitoring sites listed in **Table 10**. For reference, the days with existing, complete $PM_{2.5}$ speciation data (as defined in **Section 2.2.2**) are also shown for each site. As illustrated in **Figure 10**, the analysis of archived filters could substantially augment the record of daily $PM_{2.5}$ chemical composition information available from monitoring sites in western Pennsylvania during the time period of interest for the proposed retrospective epidemiology study. This is especially true for the Bruceton, Florence, Lawrenceville, and Liberty Borough sites, from which $PM_{2.5}$ samples that were collected on an approximately daily basis during a several-year period are available. **Section 2.5** of this report presents a strategy by which $PM_{2.5}$ speciation data determined from these archived samples could be combined with the existing $PM_{2.5}$ speciation data identified in **Section 2.2** to produce time series of daily ambient $PM_{2.5}$ chemical component concentration data suitable for use in a retrospective epidemiology study of $PM_{2.5}$ from coal-fired power plants in the Pittsburgh region. First, however, **Section 2.4** discusses important methodological issues that must be considered prior to using the existing $PM_{2.5}$ speciation data or analyzing archived $PM_{2.5}$ samples.

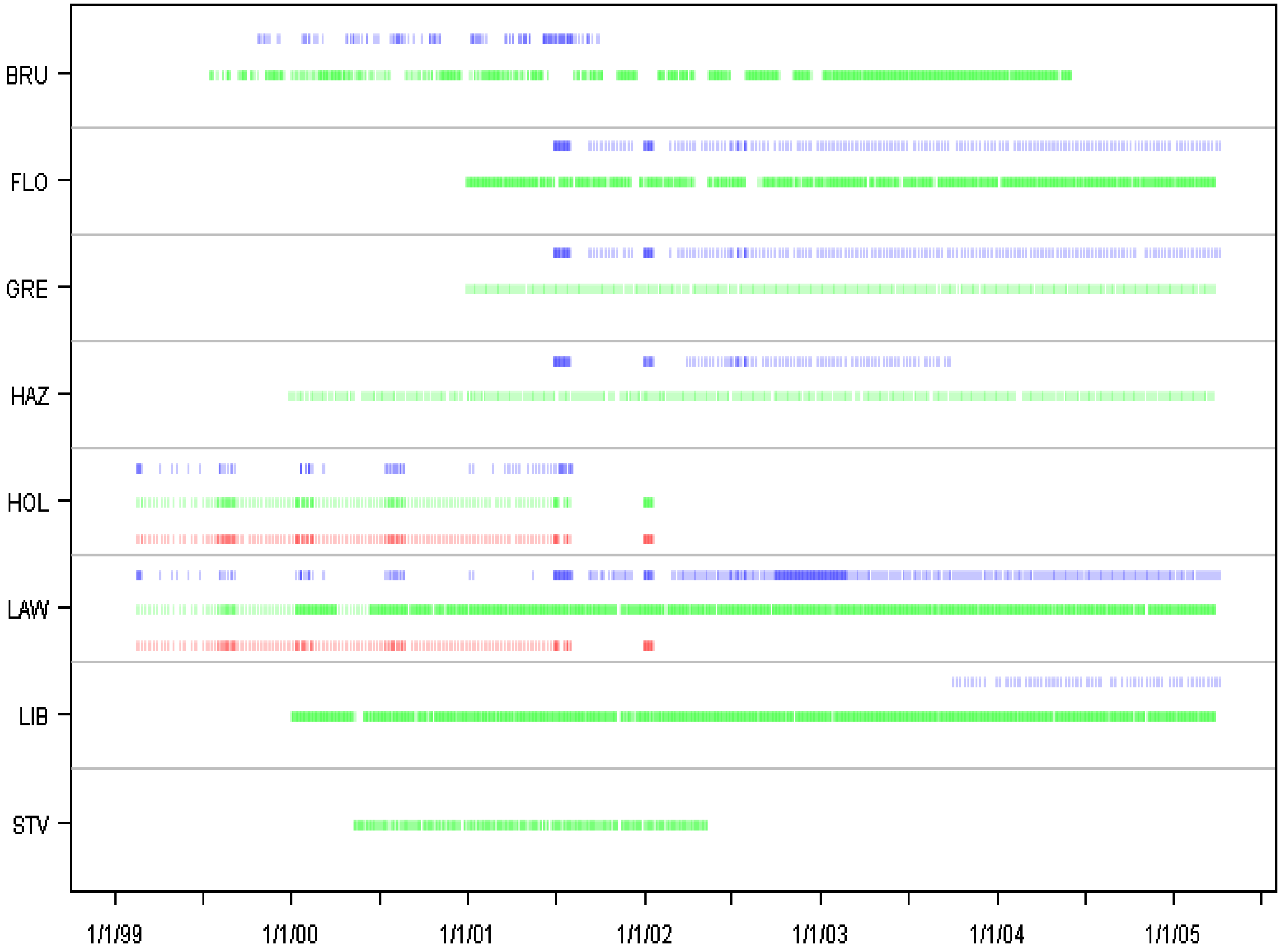


Figure 10: Time line showing the days for which archived Teflon-filter-based $PM_{2.5}$ samples (green) and quartz-filter-based $PM_{2.5}$ samples (red) are available from the sites in western Pennsylvania that monitored for $PM_{2.5}$ speciation between 1999 and 2005. For reference, the time line also indicates the days on which a complete set of $PM_{2.5}$ speciation data has already been determined for each site (blue).

2.4 Quality and Comparability of Available Air Monitoring Data

Ideally, in order to generate a database of ambient pollutant measurements for use in an epidemiology study of the health effects of $PM_{2.5}$ from various sources (including coal-fired power plants), speciated $PM_{2.5}$ data (as well as co-pollutant and meteorological data) having a daily or finer time resolution would be collected simultaneously for a period of several years at multiple monitoring sites uniformly distributed on a regular grid throughout the study region of interest. All of the sites would employ identical sampling methods, analytical (laboratory) methods, data reduction procedures, and QA/QC protocols in order to minimize inter-site biases and imprecisions that could otherwise arise from methodological discrepancies (i.e., such that any differences among the sites would largely reflect true differences in ambient concentrations, rather than measurement artifacts).

$PM_{2.5}$ and other air monitoring data available from the Pittsburgh region between 1999 and 2005 were not collected according to this ideal scenario. As shown by the inventory results presented in **Section 2.2**,

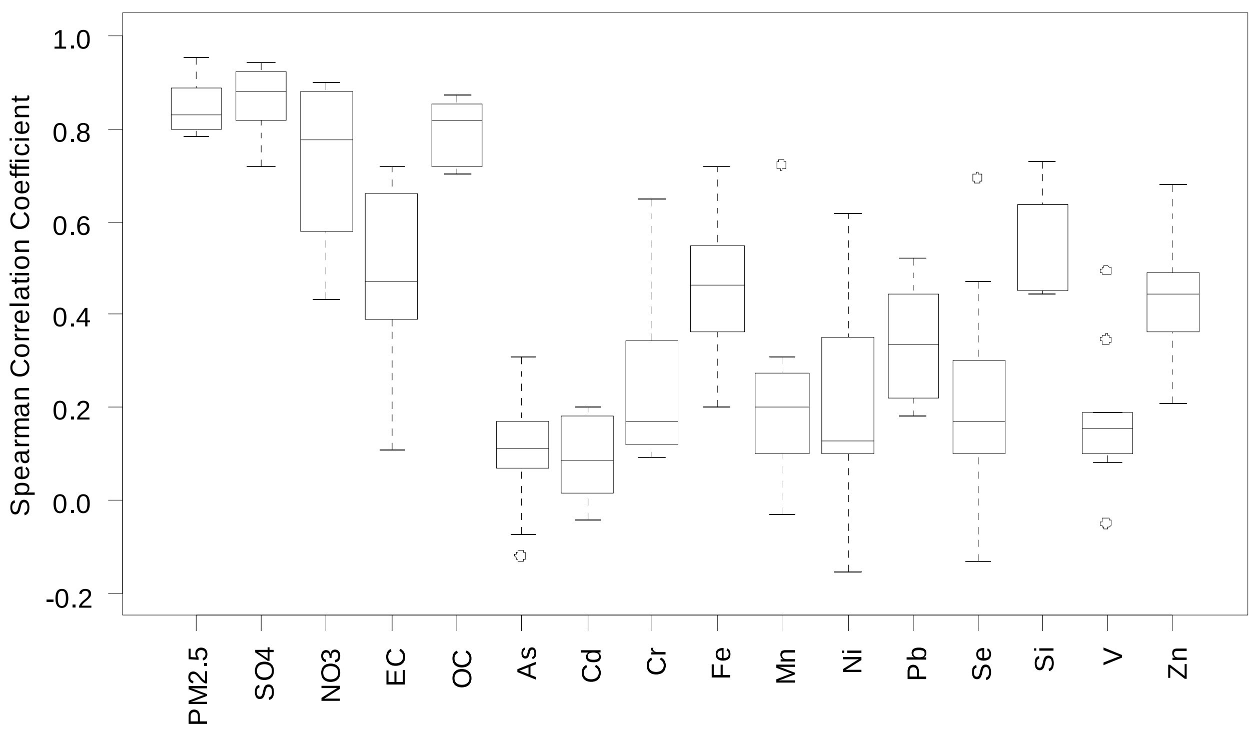


Figure 11: Inter-site Spearman correlation coefficients for $PM_{2.5}$ and select $PM_{2.5}$ components in the Pittsburgh region, based on data collected at the Lawrenceville, Schenley Park, Bruceton, Florence, and Greensburg monitoring sites between 6/30/01 and 7/31/02. Ten correlations are plotted for each variable (with the exception of Cd and Si, for which 6 correlations are plotted), corresponding to the ten possible site pairs.

data from this time period were collected at a variety of different monitoring sites located throughout the region of interest for the proposed retrospective epidemiology study. However, these sites did not feature simultaneous, daily collection of $PM_{2.5}$ speciation data over the multiple-year period required for the retrospective study. Rather, sampling activities at the sites were staggered, such that $PM_{2.5}$ speciation data are available from a number of sites during certain time periods, but only from a single site during others. Moreover, during periods from which data are available only from a single site, the identity of this site is not always the same. Finally, $PM_{2.5}$ mass and speciation data were collected using a number of different sampling and analytical techniques, as summarized in **Table 11**. Hence, the construction of daily time series of $PM_{2.5}$ speciation measurements for use in a retrospective epidemiology study requires that measurements from geographically diverse monitoring sites that were obtained using a wide variety of sampling and analytical methods be combined and used interchangeably to develop daily regional exposure estimates.

Figures 11 and 12 help to illustrate some of the challenges associated with combining data from multiple monitoring sites for use in a $PM_{2.5}$ epidemiology study. **Figure 11** presents boxplots showing inter-site Spearman correlation coefficients computed using pairwise 24-hr average mass concentration data for a subset of $PM_{2.5}$ chemical components that were collected at the Lawrenceville, Schenley Park, Bruceton, Florence, and Greensburg sites. (Hence, with the exceptions of Cd, for which data were not available from the Bruceton site, and Si, for which data were not available from the Schenley Park site, each box in the plot represents the distribution of the 10 inter-site Spearman correlation coefficients computed for the 10 possible site pairs that can be constructed from the list of sites above). **Figure 12** shows, for the same set of components, the ratios of the median concentrations measured at each of the Schenley Park, Bruceton, Florence, and Greensburg sites to the median concentrations measured at the Lawrenceville site. The results in both figures are based on data that were collected during the period from June 30, 2001, through July 31, 2002, during which monitoring activities at the five sites overlapped. Distances between sites ranged from 3 km (between the Lawrenceville and Schenley Park sites) to 79 km (between the Florence and Greensburg sites). As shown in **Figure 11** the strengths of correlations computed for pairs of monitoring sites varied considerably by $PM_{2.5}$ component. Median inter-site Spearman correlation coefficients for $PM_{2.5}$ total mass, fine particulate SO_4^{2-}, and fine particulate OC were greater than 0.8, whereas median inter-site Spearman correlation coefficients for fine particulate As, Cd, Cr, Ni, Se, and V were all less than 0.2. Moreover, median concentrations of $PM_{2.5}$ components measured at monitoring sites throughout the region differed appreciably (e.g., by a factor of two or more) in a number of cases, as shown in **Figure 12**. The appreciable relative bias and lack of correlation observed between sites for a number of $PM_{2.5}$ components are likely attributable both to the geographically diverse locations of the sites, which cause them to be impacted to different extents by various local emission sources of $PM_{2.5}$, and to measurement error, including imprecision that can contribute to the low correlations observed in **Figure 11** or bias that can lead to the disparities in central tendency observed in **Figure 12**.

Table 11: Sampling and analytical methods used by monitoring sites in the 35-county greater Pittsburgh region to determine $PM_{2.5}$ mass and speciation, 1999-2005.

Sampler Type[a]	**Filter Type**	**Start Time – End Time**[b]	**Analytical Method**	**Monitoring Sites**
$PM_{2.5}$ Total Mass				
$PM_{2.5}$ Federal Reference Method	Teflon	12:00 am – 12:00 am	Gravimetry	FLO, GRE, HAZ, HOL, LAW, LIB, MOU, SCH, STE, YOU
$PM_{2.5}$ Federal Reference Method	Teflon	09:00 am – 09:00 am	Gravimetry	FRA, HOP, STV, TOM, WHE
$PM_{2.5}$ Federal Reference Method	Teflon	12:00 pm – 12:00 pm[c]	Gravimetry	BRU
$PM_{2.5}$ Federal Equivalent Method - R&P 2025 Sampler with Very Sharp Cut Cyclone	Teflon	12:00 am – 12:00 am	Gravimetry	FLO, GRE
Met One SASS $PM_{2.5}$ Speciation Sampler	Teflon	12:00 am – 12:00 am	Gravimetry	FLO, GRE, HAZ, LAW, LIB, MOU, STE, YOU
IMPROVE Sampler	Teflon	12:00 am – 12:00 am	Gravimetry	FRO, LAW, MKG, QUA
CASTNet Sampler	Teflon	12:00 am – 12:00 am	Gravimetry	MKG, QUA
Desert Research Institute Sequential Filter Sampler	Teflon	12:00 am – 12:00 am	Gravimetry	HOL, LAW
Andersen Dichotomous Sampler	Teflon	12:00 am – 12:00 am	Gravimetry	SCH
TEOM – 50°C Operation (continuous)	Teflon-Coated Glass Fiber	12:00 am – 12:00 am	Tapered Element Oscillating Microbalance	BRU, FRA, HOL, LAW, LIB, YOU
TEOM – 30°C Operation (continuous)	Teflon-Coated Glass Fiber	12:00 am – 12:00 am	Tapered Element Oscillating Microbalance	BRU, SCH
TEOM with Filter Dynamics Measurement System (FDMS) (continuous)	Teflon-Coated Glass Fiber	12:00 am – 12:00 am	Tapered Element Oscillating Microbalance	STE
Ions				
$PM_{2.5}$ Federal Reference Method	Teflon	09:00 am – 09:00 am	Ion Chromatography	FRA, HOP, STV, TOM, WHE
$PM_{2.5}$ Federal Reference Method (or Andersen RAAS2.5-400 $PM_{2.5}$ Speciation Sampler)	Teflon	12:00 pm – 12:00 pm[c]	Ion Chromatography	BRU
Met One SASS $PM_{2.5}$ Speciation Sampler with MgO Denuder	Nylon	12:00 am – 12:00 am	Ion Chromatography	FLO, GRE, HAZ, LAW, LIB, MOU, STE, YOU
IMPROVE Sampler with Denuder	Nylon	12:00 am – 12:00 am	Ion Chromatography	FRO, LAW, MKG, QUA
CASTNet Sampler with Sodium Carbonate Denuder	Nylon	12:00 am – 12:00 am	Ion Chromatography	MKG, QUA

Sampler Type[a]	Filter Type	Start Time – End Time[b]	Analytical Method	Monitoring Sites
Desert Research Institute Sequential Filter Sampler (or R&P Partisol $PM_{2.5}$ Federal Reference Method Sampler)	Quartz	12:00 am – 12:00 am	Ion Chromatography, Automated Colorimetry, Atomic Absorption	HOL, LAW
CMU $PM_{2.5}$ Speciation Sampler with MgO and Citric Acid Denuders	Teflon, Nylon, Citric-Acid-Impregnated Cellulose Fiber	12:00 am – 12:00 am	Ion Chromatography	SCH
Andersen RAAS2.5-400 Speciation Sampler with MgO Denuder	Nylon	09:00 am – 09:00 am[c]	Ion Chromatography	FRA
PC-BOSS Sampler	Teflon and Quartz	12:00 pm – 12:00 pm	Ion Chromatography	BRU
R&P 8400N Automated Particulate Nitrate Monitor (semi-continuous)	NiChrome Flash Strip	12:00 am – 12:00 am	Flash Volatilization / Chemiluminescence	BRU, SCH
R&P 8400S Automated Particulate Sulfate Monitor (semi-continuous)	Platinum Flash Strip	12:00 am – 12:00 am	Flash Volatilization / Fluorescence	BRU, SCH
Carbon				
Met One SASS $PM_{2.5}$ Speciation Sampler	Quartz	12:00 am – 12:00 am	Thermal Optical Transmittance	FLO, GRE, HAZ, LAW, LIB, MOU, STE, YOU
IMPROVE Sampler	Quartz	12:00 am – 12:00 am	Thermal Optical Reflectance	FRO, LAW, MKG, QUA
CASTNet Sampler	Quartz	12:00 am – 12:00 am	Thermal-Optical Analysis	MKG, QUA
Desert Research Institute Sequential Filter Sampler (or R&P Partisol $PM_{2.5}$ Federal Reference Method Sampler)	Quartz	12:00 am – 12:00 am	Thermal Optical Reflectance	HOL, LAW
CMU TQQQ Sampler	Quartz	12:00 am – 12:00 am	Thermal Optical Transmittance	SCH
Andersen RAAS2.5-400 Speciation Sampler	Quartz	09:00 am – 09:00 am[c]	Thermal Optical Transmittance	FRA
PC-BOSS	Quartz, Charcoal-Impregnated Glass	12:00 pm – 12:00 pm	Temperature-Programmed Volatilization	BRU
R&P 5400 Carbon Monitor (semi-continuous)	Parallel-Plate Impactor	12:00 am – 12:00 am	Thermal-CO_2 Analysis	BRU
Sunset In-Situ Thermal/Optical Carbon Analyzer with Multi-Channel Parallel Plate Diffusion Denuder (semi-continuous)	Quartz	12:00 am – 12:00 am	In-Situ Thermal Optical Transmittance	SCH

Sampler Type[a]	Filter Type	Start Time – End Time[b]	Analytical Method	Monitoring Sites
Elements				
$PM_{2.5}$ Federal Reference Method (or Andersen RAAS2.5-400 $PM_{2.5}$ Speciation Sampler)	Teflon	12:00 pm – 12:00 pm[c]	Proton Induced X-Ray Emission	BRU
Met One SASS $PM_{2.5}$ Speciation Sampler	Teflon	12:00 am – 12:00 am	X-Ray Fluorescence	FLO, GRE, HAZ, LAW, LIB, MOU, STE, YOU
IMPROVE Sampler	Teflon	12:00 am – 12:00 am	X-Ray Fluorescence, Proton Induced X-Ray Emission	FRO, LAW, MKG, QUA
CASTNet Sampler	Teflon	12:00 am – 12:00 am	X-Ray Fluorescence	MKG, QUA
Desert Research Institute Sequential Filter Sampler (or R&P Partisol $PM_{2.5}$ Federal Reference Method Sampler)	Teflon	12:00 am – 12:00 am	X-Ray Fluorescence	HOL, LAW
Thermo-Andersen $PM_{2.5}$ Hi-Vol Sampler	Cellulose	12:00 am – 12:00 am	Inductively Coupled Plasma – Mass Spectrometry	SCH
Andersen RAAS2.5-400 $PM_{2.5}$ Speciation Sampler	Teflon	09:00 am – 09:00 am[c]	Dynamic Reaction Cell Inductively Coupled Plasma – Mass Spectrometry	FRA

[a] Federal Reference Method samplers include the Andersen RAAS2.5-300 and the R&P Partisol-Plus 2025 $PM_{2.5}$ Sequential Air Sampler. [b] The desired period of collection for the proposed epidemiology study is 12:00 am – 12:00 am. Hence, for semi-continuous sampling and filter-based sampling that was conducted with a time resolution finer than 24 hours, the start time and end time were reported as 12:00 am if data could be aggregated in 24-hour periods according to this schedule. [c] Samplers operated from 12:00 am to 12:00 am during EPA sampling intensives.

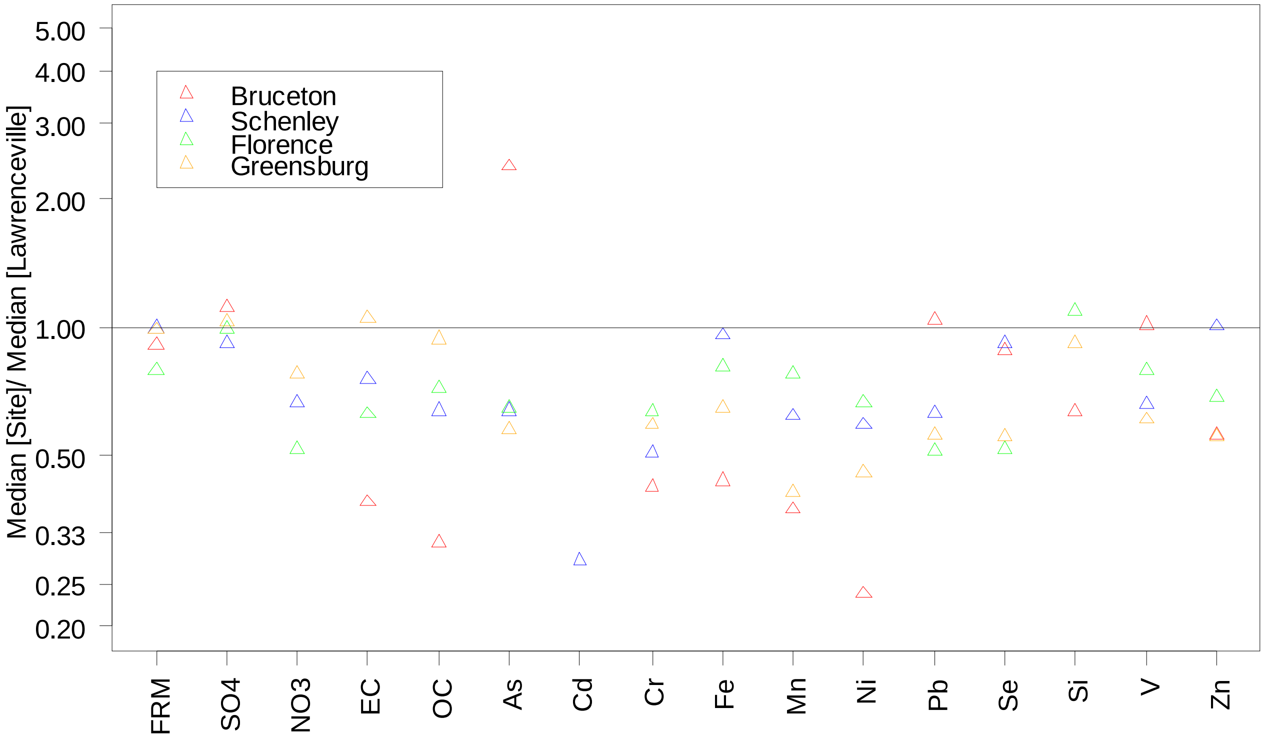


Figure 12: Ratios of median concentrations of PM$_{2.5}$ (FRM) and select PM$_{2.5}$ components measured at the Bruceton, Schenley Park, Florence, and Greensburg sites to median concentrations of these species measured at the Lawrenceville site between 6/30/01 and 7/31/02. (For each comparison, data from a given day were excluded if only one of the two sites under consideration produced a valid measurement that day).

Systematic bias is correctable and therefore does not pose a major problem for the construction of an exposure database for use in a retrospective epidemiology study. It is essential, however, that biased measurements be properly calibrated (e.g., corrected to some reference level so that the relative bias between them is removed) prior to use in the epidemiological models, especially for our proposed study in which measurements made at different sites or using different measurement techniques may need to be used interchangeably to represent exposures on different days, depending on data availability. For example, consider a hypothetical scenario in which ambient concentrations of species X were measured on half of the study days using only method A and on the other half of the study days using only method B, and in which method B is biased 50% low relative to method A. (Methods A and B might be two different sampling/analytical methods used at the same monitoring site, the same sampling/analytical method used at two different monitoring sites, or two different sampling/analytical methods used at two different monitoring sites). If the data obtained using methods A and B are combined to form a single time series of daily concentrations of species X without first accounting for the bias between the methods, then the bias will cause a misrepresentation of the variability in ambient concentrations of species X. (For example, if ambient concentrations measured using method A would have been [1, 2, 1, 1.5] over a four-day period, but measurements made using method B were used to represent ambient concentrations on every second day with no knowledge of the bias between methods A and B, then the time series of ambient concentrations would incorrectly appear to be [1, 1, 1, 0.75]). Hence, it is important that systematic biases among measurement techniques and monitoring sites be identified and corrected (e.g., using a calibration derived from collocated measurements, as described in more detail below) if the data produced by these techniques or sites are to be used interchangeably to represent exposures.

Unlike systematic bias, the instrument imprecision and spatial variability that contribute to the low correlations observed in **Figure 11** for certain $PM_{2.5}$ components cannot be corrected, and are sources of exposure error that must be acknowledged prior to epidemiological modeling. Wade et al. (2006) reported that, based on analyses of $PM_{2.5}$ speciation data (i.e., SO_4^{2-}, NO_3^-, NH_4^+, EC, OC) from Atlanta, Georgia, using modified semivariograms, the population-weighted uncertainty in concentrations of primary pollutants such as EC arising from instrument imprecision and spatial variability tended to account for about 60-70% of the temporal variation in concentrations of these pollutants, whereas the population-weighted uncertainty in concentrations of secondary pollutants due to instrument imprecision and spatial variability was much less (e.g., 25% of the temporal variation in concentrations of SO_4^{2-}). The correlations presented in **Figure 11** for $PM_{2.5}$ components in the Pittsburgh region, while not direct measures of uncertainty, are consistent with the results of Wade et al. The strongest inter-site correlations were observed for SO_4^{2-}, a predominantly regional, secondary pollutant for which measurement techniques tend to be consistent and precise (e.g., 5% imprecision, U.S. EPA, 2001). Inter-site correlations were appreciably weaker for EC, a predominantly locally-emitted, primary pollutant for which measurement techniques have substantially more imprecision (e.g., 5% to 30%, U.S. EPA, 2001). Even lower inter-site correlations were observed for many fine particulate trace element species, which again are locally-emitted, primary pollutants that are subject to large analytical imprecisions (e.g., 20% to >100%, U.S. EPA, 2001) resulting from very low ambient concentrations (i.e., near or below the detection limits of the analytical methods), variable blank concentrations, and other methodological limitations. The poor correlations observed for certain trace element species in **Figure 11** likely further result from the fact that the five sites included in the figure used three different analytical techniques (XRF, PIXE, ICP-MS) for elemental analysis.

Thus, uncertainties arising from sampling and analytical imprecision (i.e., either the imprecision associated with a given sampling or analytical method or the imprecision resulting from the use of multiple methods interchangeably to determine concentrations of a given parameter) and from spatial variability (especially for locally-emitted pollutants, for which the day-to-day variability in concentrations measured at a given monitoring site may not represent the day-to-day variability in concentrations at other locations in the region) contribute to random error (noise) in the exposure estimates used in $PM_{2.5}$ epidemiology studies. Such exposure measurement error can attenuate the effect estimates in population-based time-series studies (Zeger et al., 2000; Wade et al., 2006), decreasing the ability of the epidemiological model to detect an association between a health outcome and an explanatory variable when one truly exists. Because exposure measurement errors resulting from methodological imprecision and spatial variability differ appreciably by $PM_{2.5}$ component, as discussed above, it is important to identify and quantify these errors prior to designing an epidemiological study that is focused on the health effects of $PM_{2.5}$ components.

For the proposed retrospective epidemiology study of $PM_{2.5}$ from coal-fired power plants in the Pittsburgh region, it is particularly important to consider the effects of bias and imprecision in exposure estimates, because these estimates will be derived from a number of different monitoring sites that employed a number of different sampling and analytical techniques. Statistical methods for quantifying bias and imprecision and for calibrating biased measurements are discussed in detail later in this report, as are space-time geostatistical techniques for combining measurements from multiple monitoring sites for purposes of developing time series of exposure estimates. The remainder of this section focuses on assessing the quality, comparability, and limitations of the various sampling and analytical techniques that were used by monitoring sites in the Pittsburgh region to measure ambient mass concentrations of $PM_{2.5}$ and its chemical components between 1999 and 2005 or that would be used to determine these concentrations from archived $PM_{2.5}$ samples. Differences among the QA/QC protocols employed by the various sites are also discussed, as disparities in data validation criteria can affect the comparability of measurements from different monitoring sites.

2.4.1 Comparison of Sampling and Analytical Techniques

Table 11 demonstrates that the $PM_{2.5}$ speciation data available from the Pittsburgh region between 1999 and 2005 were generated using a vast array of sampling and analytical methods and combinations thereof. The methods used to measure a given $PM_{2.5}$ parameter often differed in a number of ways, including some or all of the following:

- Sampler inlet design
- Type of denuder (if used)
- Type of filter (or other sample collection medium)
- Type of backup filter (if used)
- Sampler operating temperature
- Sampling flow rate
- Duration of sample collection
- Sampling start and end times

- Filter / sample handling and storage procedures
- Sample preparation procedures
- Analytical instrument / method
- Data reduction procedures (e.g., use of blank correction)

Each of these factors can affect the quality and comparability of $PM_{2.5}$ speciation results. For example, variations in inlet design can affect particle collection efficiency, causing a relative bias between two samplers. Factors such as filter type, denuder use, backup filter use, and sampler operating temperature influence whether reactive gases are collected with the sample, causing a positive sampling artifact, or whether semi-volatile material is lost from the sample, causing a negative sampling artifact. The sampling flow rate and duration of sample collection (together with the ambient concentration of particles on a particular day) determine the quantity of particles that are collected on the filter, which can affect the ability to determine concentrations of fine particle components that are present in trace amounts (if too little sample is collected, then the analytical methods may not have sufficient sensitivity to determine these components). The sampling schedule (i.e., start and end times) affects whether measured values are directly comparable temporally with other air pollution measurements and with health outcomes data (which typically are tabulated from midnight to midnight). If filters and samples are not handled properly, contamination can cause bias or imprecision in the measured values (depending on whether it is random or systematic), and variations in temperature can affect the extent to which semi-volatile components are retained on the filter. Laboratory procedures, including sample preparation (e.g., extraction, digestion, dilution) and the particular analytical instrument or method being used, also vary considerably in their accuracy, precision, and sensitivity. Finally, differences in data reduction procedures can affect the comparability of reported values. For example, if Group A subtracted average field blank concentrations before reporting $PM_{2.5}$ speciation measurements and Group B did not, then concentrations reported by Group B would be biased high relative to those reported by Group A, all other things being equal.

This section focuses on quantifying differences in the accuracy, precision, and sensitivity of various sampling and analytical techniques that were used to measure ambient concentrations of $PM_{2.5}$ chemical components in the Pittsburgh region between 1999 and 2005. Because our goal is not to improve upon these sampling and analytical methodologies but rather to ascertain the quality and comparability of existing measurements, the individual effects of the various factors identified above are not explored in detail. Rather, a statistical approach based on pairwise comparisons of final reported values, which reflect the combined effects of all of these factors, is employed. Moreover, for purposes of this feasibility assessment, we did not exhaust all of the comparisons that could be specified from the information presented in **Table 11**, but rather focused on those sites and methods that would be of particular importance for purposes of the retrospective epidemiology study. Results are summarized in the subsections below for measurements of $PM_{2.5}$ total mass, $PM_{2.5}$ ionic components, $PM_{2.5}$ carbonaceous components, and $PM_{2.5}$ elemental components.

2.4.1.1 $PM_{2.5}$ Total Mass

Table 11 identifies 12 different methods that were employed by the 19 $PM_{2.5}$ speciation monitoring sites located in the 35-county greater Pittsburgh region to determine ambient $PM_{2.5}$ total mass concentrations between 1999 and 2005. These methods included both integrated methods, which involve sampling to collect fine particles on a filter over a given period of time (e.g., 24 hours) followed by gravimetric analysis in the laboratory to determined the average mass of particles collected per volume of air sampled during that period, and continuous methods, which determine $PM_{2.5}$ concentrations in the field in real time. The integrated methods used to measure $PM_{2.5}$ mass in the Pittsburgh region differed primarily according to the type of sampler used (i.e., Federal Reference Method sampler vs. Federal Equivalent Method sampler vs. dichotomous sampler vs. various speciation samplers) and the sampling schedule (i.e., the NETL/OST Bruceton site generally sampled from 12:00 p.m. to 12:00 p.m.; the five SCAMP monitoring sites generally sampled from 9:00 a.m. to 9:00 a.m., and all of the other sites sampled from 12:00 a.m. to 12:00 a.m.). All of these methods employed Teflon filters for sample collection. All continuous $PM_{2.5}$ mass measurements were made using tapered element oscillating microbalances (TEOMs); however, the sampler configuration and operating temperature varied in some cases.

With the exceptions of the Frostburg, M.K. Goddard, and Quaker City sites, at which $PM_{2.5}$ mass concentrations were measured using only IMPROVE or CASTNet samplers, all of the $PM_{2.5}$ speciation monitoring sites in the greater Pittsburgh region included a $PM_{2.5}$ Federal Reference Method sampler to determine 24-hour average ambient $PM_{2.5}$ mass concentrations. All $PM_{2.5}$ FRM samplers are designed according to consistent specifications (e.g., regarding the size selective inlet, filter, filter cassette, filter holder, flow rate requirements, temperature and pressure monitoring requirements, etc.) established by the U.S. EPA (U.S. EPA, 1997), resulting in reasonable equivalence (i.e., within measurement imprecision) among samplers. Imprecisions estimated from collocated $PM_{2.5}$ FRM samplers are typically on the order of 0.5 to 1.0 $\mu g/m^3$ (Chow and Watson, 1998). It is important to recognize, however, that mass concentrations of $PM_{2.5}$ determined from FRM samplers may deviate from true ambient concentrations. $PM_{2.5}$ FRM samplers do not employ denuders to scrub reactive gases, backup filters to collect revolatilized material, or blank correction methods to adjust for contamination during handling and storage. Hence, biases between $PM_{2.5}$ mass concentrations determined by the FRM and actual ambient $PM_{2.5}$ mass concentrations may result, for example, from losses of semi-volatile species such as NO_3^-, NH_4^+, and OC (Jansen et al., 2002; Frank, 2006). Nevertheless, because the $PM_{2.5}$ NAAQS is based on $PM_{2.5}$ concentrations determined using FRM samplers, and because these samplers provide a methodologically consistent means for comparing $PM_{2.5}$ mass concentrations measured at monitoring sites throughout the Pittsburgh region, FRM measurements will likely be used as the basis for estimating exposures to $PM_{2.5}$ mass for the proposed retrospective epidemiology study.

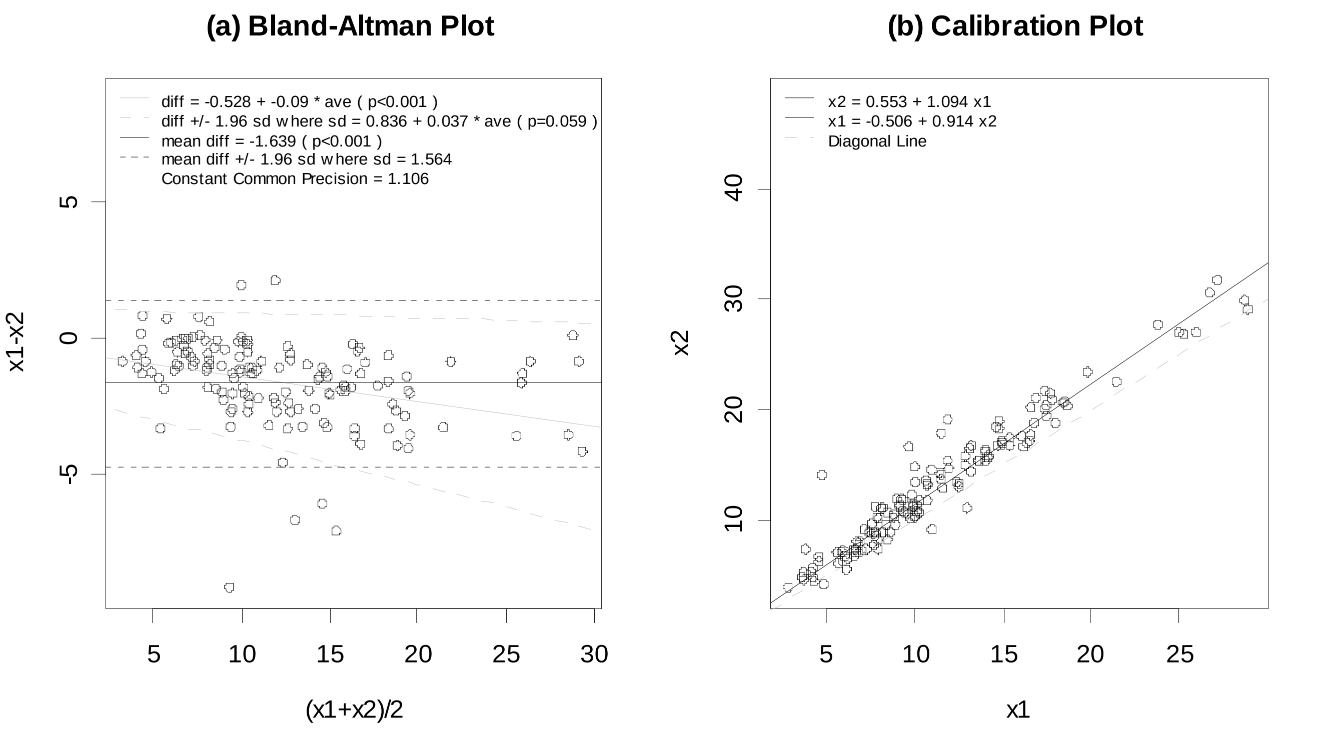


Figure 13: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily PM2.5 concentrations measured at the Lawrenceville site using a Desert Reserach Institute Sequential Filter Sampler (x1) and a PM2.5 Federal Reference Method sampler (x2) between 10/1/02 and 2/27/03.

On certain days at certain sites, valid $PM_{2.5}$ mass concentration data were not determined from FRM sampling (e.g., because of a sampler malfunction or failed QA/QC criterion) but were determined using a speciation, FEM, or dichotomous sampler. Mass concentrations determined by these samplers are generally comparable to those determined by $PM_{2.5}$ FRM samplers, as illustrated in **Figures 13 and 14.** These figures present the results of Bland-Altman analyses comparing 24-hour average $PM_{2.5}$ mass concentrations measured using collocated $PM_{2.5}$ samplers at the Lawrenceville monitoring site. **Figure 13** compares $PM_{2.5}$ concentrations measured by UORVP using a Desert Research Institute Sequential Filter Sampler (SFS) with $PM_{2.5}$ concentrations measured by ACHD using a $PM_{2.5}$ FRM sampler, and **Figure 14** compares $PM_{2.5}$ concentrations measured by ACHD using a Met One SASS $PM_{2.5}$ speciation sampler with $PM_{2.5}$ concentrations measured by ACHD using a $PM_{2.5}$ FRM sampler.

The method of Bland and Altman examines the agreement between two methods by plotting the paired differences between measurements made by the two methods (i.e., x1 – x2) against the corresponding paired average measurements (i.e., [x1 + x2]/2) (Bland and Altman, 1986). A calibration line relating the two methods is then derived by regressing the paired differences on the paired averages. As discussed in a later section of this report, the Bland-Altman technique, when applied to two methods with approximately equal imprecisions, correctly estimates the calibration line describing the relative bias between the two methods, whereas simple linear regression of one method on the other yields a distorted estimate of the calibration line. However, if the imprecisions of the two methods differ or if more than two methods are being compared, then latent variable modeling (LVM) is needed to estimate relative bias. For purposes of this preliminary assessment of method comparability, equal imprecisions are assumed and the relatively simple Bland-Altman technique is used in place of the more complex LVM technique. Prior to calibrating measurements for use in an actual epidemiology study, however, LVM would be applied in all cases to verify the assumption of equal imprecisions.

The Bland-Altman results presented in **Figures 13 and 14** indicate small relative biases between $PM_{2.5}$ concentrations determined by the FRM sampler and those determined by the SFS and SASS samplers. On average, $PM_{2.5}$ concentrations from the SFS sampler were about 1.6 µg/m^3 less than those from the FRM sampler, and $PM_{2.5}$ concentrations from the SASS sampler were about 1.4 µg/m^3 greater than those from the FRM sampler. In both cases, the relative bias between methods varied significantly as a function of

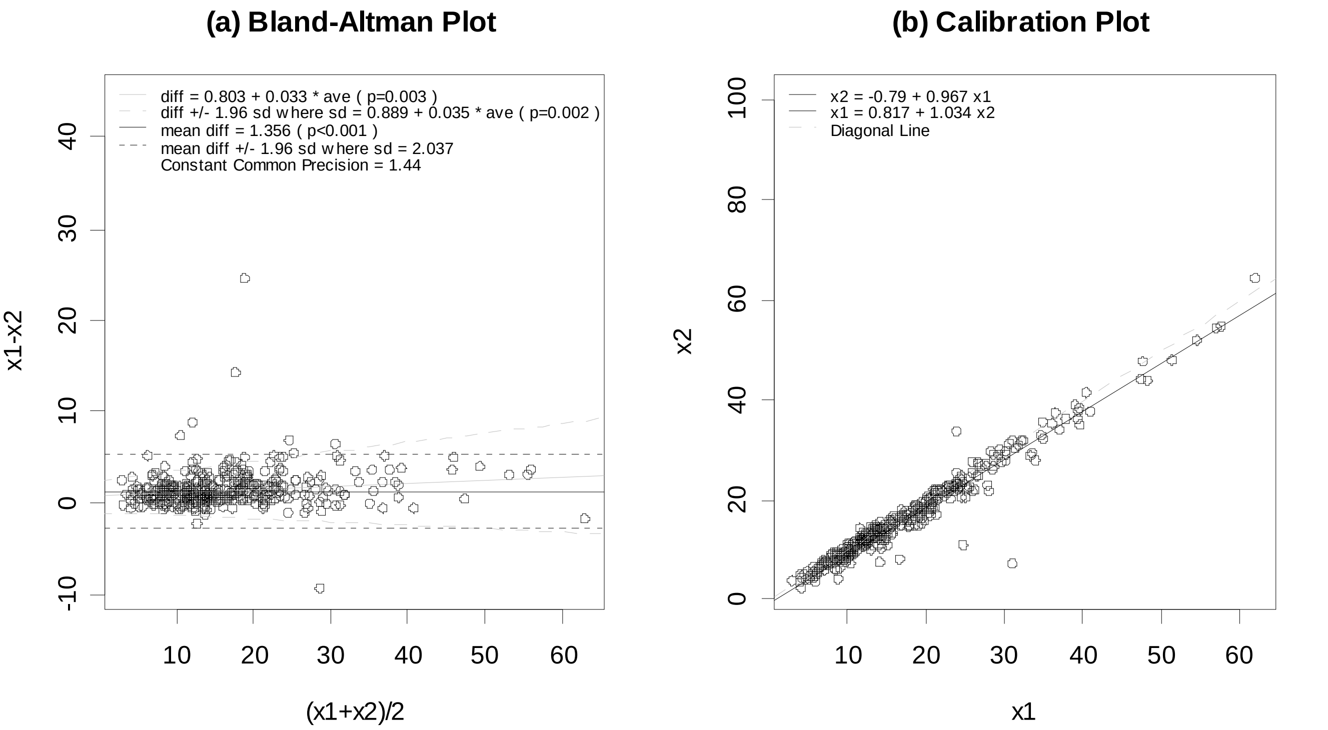


Figure 14: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily $PM_{2.5}$ concentrations measured at the Lawrenceville site using a Met One SASS $PM_{2.5}$ speciation sampler (x1) and a $PM_{2.5}$ Federal Reference Method sampler (x2) between 6/30/01 and 3/29/05.

average concentration. These biases are correctable using the calibration lines shown in **Figures 13 (b) and 14 (b)**. Constant common imprecision estimates for the comparisons between the FRM sampler and the SFS and SASS samplers were 1.1 µg/m^3 (9.0%) and 1.4 µg/m^3 (8.7%), respectively, just slightly greater than the 0.5-1.0 µg/m^3 range referenced above for collocated FRM samplers. These results indicate that, after calibration to account for relative bias, data from the FRM and speciation samplers could be used interchangeably to represent ambient $PM_{2.5}$ mass concentrations with little impact on overall data quality.

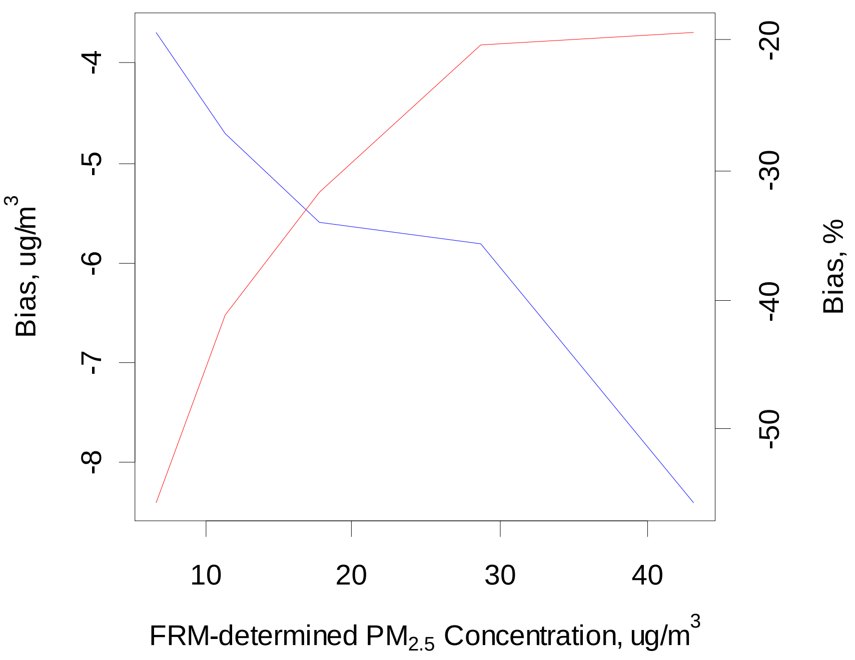


Figure 15: Bias of the TEOM monitor relative to the FRM monitor at the Franciscan University of Steubenville site, as a function of the FRM-determined $PM_{2.5}$ concentration. The blue line represents bias in µg/m^3; the red line represents bias in %.

Eight of the sites that monitored for $PM_{2.5}$ speciation in the Pittsburgh region included TEOM monitors to continuously measure ambient $PM_{2.5}$ mass concentrations. In the TEOM, $PM_{2.5}$ is collected on a Teflon-coated glass fiber filter that sits on the end of a tapered, oscillating glass tube. As $PM_{2.5}$ accumulates on the filter over time, its mass is determined by measuring the change in the oscillation frequency of the tube. Until recently, it was common practice to operate TEOM monitors at 50°C to remove moisture from the sampled air and prevent condensation on the filter. However, operation at this temperature may also cause the loss of some semi-volatile particulate matter (e.g., ammonium nitrate, semi-volatile organic compounds, and particle-bound water), resulting in an underestimation of total $PM_{2.5}$ mass relative to the

Federal Reference Method (Allen et al., 1997). The magnitude of this artifact is dependent upon the composition of the sampled $PM_{2.5}$, and therefore varies with location and time.

As shown in **Table 11** six of the eight speciation sites with TEOMs operated these TEOMs at 50°C. Connell et al. (2005b) presented a comparison of $PM_{2.5}$ mass concentrations measured using the collocated FRM and 50°C TEOM monitors at the SCAMP Franciscan University of Steubenville site. For purposes of the comparison, hourly TEOM data were averaged over 24-hr periods corresponding to the FRM sampling schedule; 515 pairs of collocated measurements were included in the analysis. Both methods exhibited similar imprecisions, ranging from 0.0 to 4.2µg/m^3 for the FRM sampler and from 1.7 to 4.0 µg/m^3 for the TEOM monitor for FRM-determined ambient $PM_{2.5}$ concentrations of 6.6 to 43.2 µg/m^3. However, as shown in **Figure 15**, $PM_{2.5}$ mass concentrations determined by the TEOM were substantially biased in the negative direction relative to those determined by the FRM, likely because of losses of semi-volatile material at the 50°C operating temperature.

To reduce losses of semi-volatile material, TEOM monitors can be equipped with a Sample Equilibration System (SES) or Filter Dynamics Measurement System (FDMS) and operated at 30°C; the TEOM monitor at Schenley Park (and for a time, the TEOM monitor at Bruceton) was equipped with an SES and operated

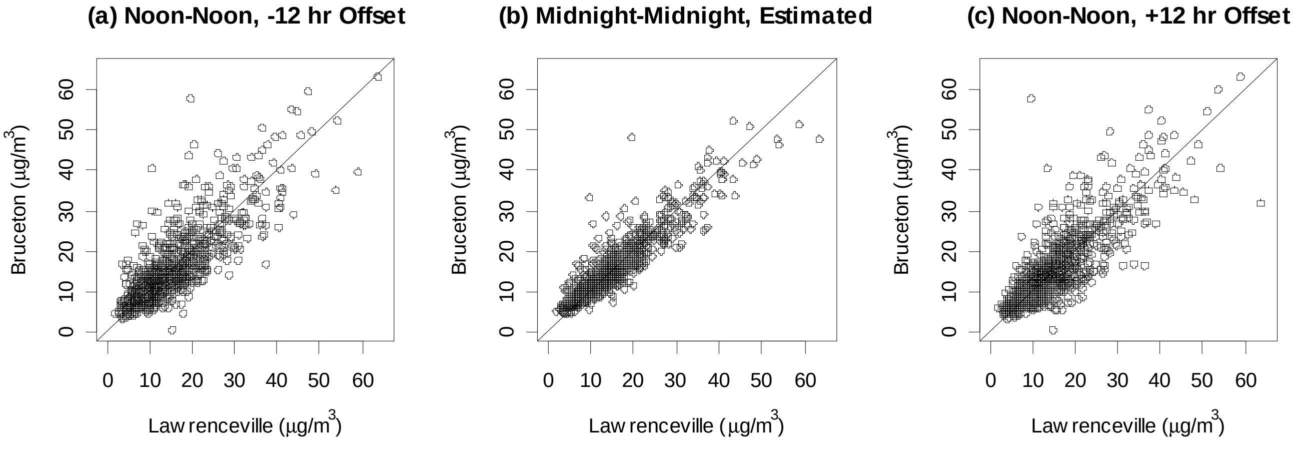


Figure 16: Scatterplots comparing 24-hour FRM $PM_{2.5}$ concentrations measured from midnight-to-midnight at the Lawrenceville site with 24-hour FRM $PM_{2.5}$ concentrations (a) measured from noon-to-noon at the Bruceton site with a -12-hour offset relative to the Lawrenceville measurements, (b) estimated from midnight-to-midnight at the Bruceton site on the basis of noon-to-noon measurements made there, and (c) measured from noon-to-noon at the Bruceton site with a +12-hour offset relative to the Lawrenceville measurements.

at 30^{o}C, and the TEOM monitor at the EPA's Steubenville site was equipped with an FDMS. Nevertheless, $PM_{2.5}$ concentrations determined using these TEOM monitors may still exhibit small biases relative to those determined using FRM monitors.

Hence, before using TEOM data to represent $PM_{2.5}$ concentrations for purposes of a retrospective epidemiology study, any biases in these data must be removed (e.g., via LVM). All of the $PM_{2.5}$ speciation monitoring sites in the Pittsburgh region that included TEOM monitors (and, with the exception of AQS site No. 420050001, all of the $PM_{2.5}$ mass monitoring sites that included TEOM monitors) also included FRM monitors; hence, calibrations can be performed easily using pairwise data from these collocated monitors.

As discussed above, 24-hr integrated $PM_{2.5}$ samples from the NETL/OST Bruceton monitoring site were generally collected from 12:00 p.m. to 12:00 p.m., and 24-hr integrated samples from the five SCAMP monitoring sites were generally collected from 9:00 a.m. to 9:00 a.m. These abnormal sampling schedules affect the comparability of daily $PM_{2.5}$ data from these sites with daily $PM_{2.5}$ data from other monitoring sites in the Pittsburgh region and with daily health outcomes data, which are collected from 12:00 a.m. to 12:00 a.m. The disparity is of particular concern for the NETL/OST Bruceton site, which is one of the largest sources of existing $PM_{2.5}$ data and archived $PM_{2.5}$ samples from the Pittsburgh region during the time period of interest, but which operated on a sampling schedule that caused most of these data and samples to be offset by a half-day from other available exposure and health data.

Table 12: Summary of paired differences between $PM_{2.5}$ concentrations measured using an FRM monitor at the Bruceton site and $PM_{2.5}$ concentrations measured using an FRM monitor at the Lawrenceville site between 1/01 and 12/03. Results include a comparison of paired differences computed using $PM_{2.5}$ concentrations measured from noon-to-noon at the Bruceton site with paired differences computed using midnight-to-midnight concentrations estimated by averaging these noon-to-noon measurements. (All data from the Lawrenceville site were measured from midnight-to-midnight).

	Time Period Represented by Bruceton Data		
	Noon-to-Noon -12-hr offset (Case 1)	Noon-to-Noon +12-hr offset (Case 2)	Midnight-to-Midnight Estimate (Case 3)
Number of Paired Differences, BRU – LAW	965	965	943
Mean Paired Difference, BRU – LAW	-0.5	-0.4	-0.5
Standard Deviation of Paired Differences, BRU – LAW	5.6	5.5	3.6
Variance of Paired Difference, BRU – LAW	31.6	30.5	13.0
F-test Results (relative to Case 3)			
F	2.44	2.35	NA
p-value	<0.0001	<0.0001	NA

Daily $PM_{2.5}$ concentrations between 12:00 a.m. and 12:00 a.m. at the NETL/OST and SCAMP monitoring sites could be estimated by computing time-weighted averages from the two 24-hr concentrations that represent a portion of each day. For the Bruceton site, the time-weighted average concentration on a given day would be the simple arithmetic mean of the 24-hr concentration determined from 12:00 p.m. on the previous day to 12:00 p.m. on the day of interest and the 24-hour concentration determined from 12:00 p.m. on the day of interest to 12:00 p.m. on the following day. Midnight-to-midnight average $PM_{2.5}$ concentrations estimated in this way using FRM data from the Bruceton site were compared with midnight-to-midnight average $PM_{2.5}$ concentrations measured using an FRM sampler at the Lawrenceville site. To assess the effect of averaging the Bruceton data, the noon-to-noon $PM_{2.5}$ concentrations measured at the Bruceton site were also compared with the midnight-to-midnight concentrations from the Lawrenceville site. Comparisons were performed both using the Bruceton data that were lagged by 12 hours in the negative direction relative to the Lawrenceville data and using the Bruceton data that were lagged by 12 hours in the positive direction relative to the Lawrenceville data. **Figure 16** presents scatterplots showing these three comparisons. It is visually evident that use of the estimated midnight-to-midnight data from the Bruceton site improved the comparability (i.e., decreased the scatter) between 24-hour $PM_{2.5}$ concentrations from this site and 24-hour $PM_{2.5}$ concentrations from the Lawrenceville site. To confim this observation, paired differences between $PM_{2.5}$ concentrations at Bruceton and $PM_{2.5}$ concentrations at Lawrenceville were computed for each of the three cases shown in **Figure 16**, and F-tests were applied to compare the variability in these paired differences among cases. As shown in **Table 12**, the paired differences between $PM_{2.5}$ concentrations from the Bruceton and Lawrenceville sites were significantly less variable when estimated midnight-to-midnight data from the Bruceton site were used than when measured noon-to-noon data were used. (The absolute values of the paired differences were also significantly less when estimated midnight-to-midnight data from the Bruceton site were used). Although not shown, similar results were obtained when $PM_{2.5}$ data from the Bruceton site were compared with those from the Schenley Park site. These results suggest that 24-hour integrated data (including both $PM_{2.5}$ total mass data and compositional data derived from integrated $PM_{2.5}$ samples) that were not collected from midnight-to-midnight should be combined to derive midnight-to-midnight averages (e.g., using time-weighted means) prior to use in the epidemiology study. It may be possible to further improve these midnight-to-midnight estimates using TEOM data where available; this possibility should be investigated when constructing the exposure database for use in the study.

2.4.1.2 $PM_{2.5}$ Ions

Table 11 identifies eleven different methods that were employed by the 19 $PM_{2.5}$ speciation monitoring sites in the greater Pittsburgh region to determine ambient concentrations of inorganic ionic components of $PM_{2.5}$ (e.g., SO_4^{2-}, NO_3^-, NH_4^+, etc.) between 1999 and 2005. Ion concentrations were measured using both integrated and semi-continuous methods. At all of the AQS speciation sites, integrated $PM_{2.5}$ samples for ionic analysis were collected from 12:00 a.m. to 12:00 a.m. on nylon filters using Met One SASS speciation samplers that were equipped with MgO denuders. Concentrations of SO_4^{2-}, NO_3^-, NH_4^+, K^+, and Na^+ were determined from these samples by ion chromatography (IC). Additional ion concentration data at the Lawrenceville site are available from the UORVP and IMPROVE sampling networks. UORVP collected 6- or 24-hour integrated $PM_{2.5}$ samples for ionic analysis on quartz filters using Desert Research Institute SFS samplers (or $PM_{2.5}$ FRM samplers) that operated from midnight to midnight. These samples were then analyzed by IC to determine concentrations of SO_4^{2-}, NO_3^-, and Cl^-, by automated colorimetry to determine concentrations of NH_4^+, and by atomic absorption to determine concentrations of Na^+ and K^+. IMPROVE collected integrated $PM_{2.5}$ samples for ionic analysis from 12:00 a.m. to 12:00 a.m. on nylon filters using an IMPROVE sampler that was equipped with a denuder, and determined concentrations of SO_4^{2-}, NO_3^-, and Cl^- from these samples by IC.

Apart from the AQS sites, ionic data from the PAQS Schenley Park site and the NETL/OST Bruceton site are of particular interest for the proposed epidemiology study, because these sites have an abundance of daily $PM_{2.5}$ speciation data and were located in Allegheny County, where much of the Pittsburgh region's population is located. Twenty-four-hour average concentrations of SO_4^{2-}, NO_3^-, and NH_4^+ at the Schenley Park site were determined from integrated $PM_{2.5}$ samples collected using a CMU $PM_{2.5}$ speciation sampler that operated from approximately 12:00 a.m. to 12:00 a.m. (at times, several separate samples were collected during this period to improve the time resolution of the data). The sampler was equipped with MgO and citric-acid-coated denuders to scrub nitric acid and ammonia gases, respectively, from the sampled air stream. Sulfate was determined by IC from $PM_{2.5}$ samples collected on Teflon filters; nitrate was determined by IC from the Teflon-filter-based samples and from samples collected on nylon backup filters (which were used to account for volatilized nitrate), and ammonium was determined by IC from the Teflon-filter-based samples and from samples collected on citric-acid-impregnated cellulose fiber backup filters (which were used to account for volatilized ammonium). Twenty-four-hour average concentrations of SO_4^{2-}, NO_3^-, NH_4^+, K^+, Na^+, and several other inorganic ions at the Bruceton site were determined by IC from Teflon-filter-based $PM_{2.5}$ samples collected using a $PM_{2.5}$ FRM sampler or an Andersen RAAS2.5-400 $PM_{2.5}$ speciation sampler that generally operated from 12:00 p.m. to 12:00 p.m. (except during several EPA sampling intensives, when it operated from 12:00 a.m. to 12:00 a.m.). Twenty-four-hour SO_4^{2-} and NO_3^- concentrations at the Bruceton site were also determined by IC at times from integrated samples collected on Teflon and quartz filters using a PC-BOSS sampler. This sampler, which is designed to collect particles with diameters ranging from about 0.1 – 2.3 m (Modey and Eatough, 2004), is equipped with a denuder to remove reactive gases (e.g., nitric acid) from the sample steam, and also includes a backup Nylasorb filter for determining the amount of semi-volatile nitrate lost from the particles during sampling. Finally, both the Schenley Park and Bruceton sites included semi-continuous measurements of fine particulate SO_4^{2-} and NO_3^- using Rupprecht & Patashnick 8400S and 8400N monitors. In these instruments, $PM_{2.5}$ is collected on a platinum (8400S) or NiChrome (8400N) flash strip, and flash volatilization is applied every ten minutes to convert the S (assumed to be SO_4^{2-}) or NO_3^- contained in the sample to NO_x (8400N) or SO_2 (8400S), respectively, which are then measured using gas analyzers.

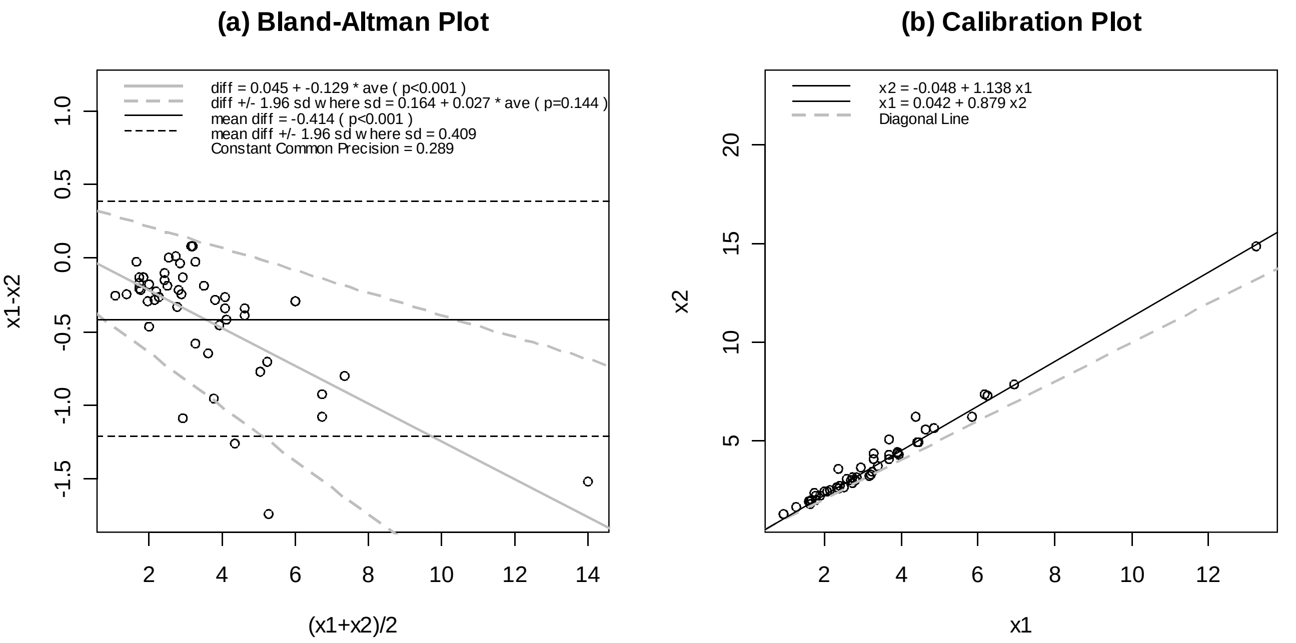


Figure 17: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily SO_4^{2-} concentrations measured at the Lawrenceville site using a Desert Research Institute Sequential Filter Sampler (x1) and a Met One SASS $PM_{2.5}$ speciation sampler (x2) between 10/1/02 and 2/27/03.

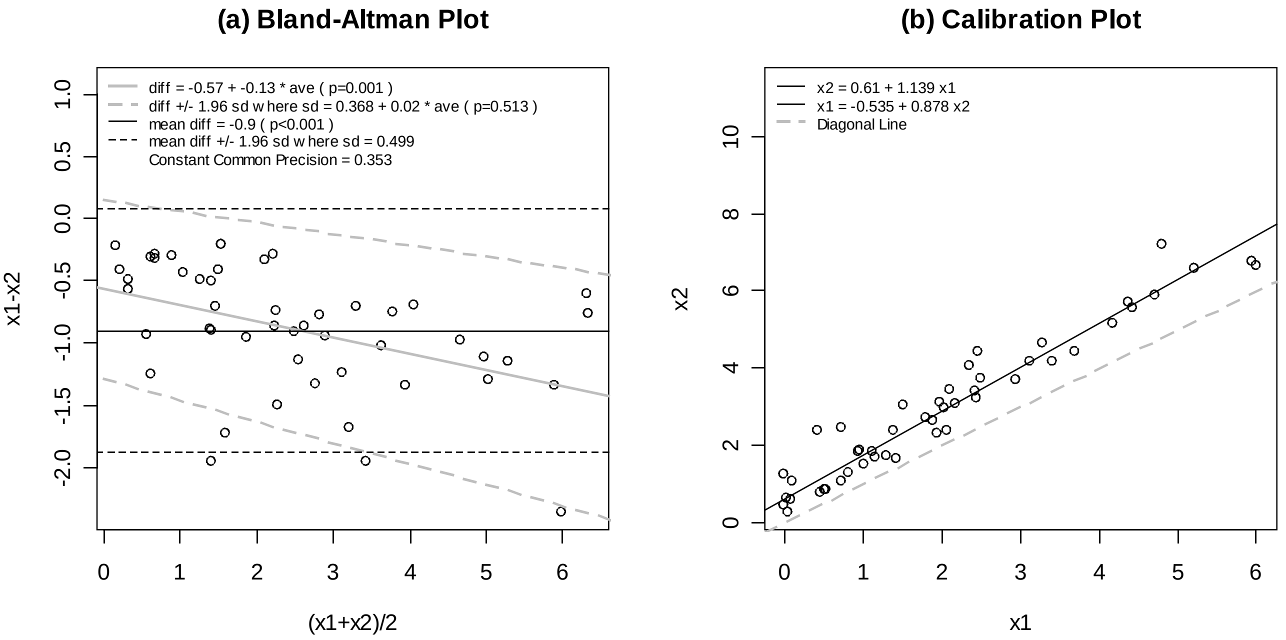


Figure 18: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily NO_3- concentrations measured at the Lawrenceville site using a Desert Research Institute Sequential Filter Sampler (x1) and a Met One SASS $PM_{2.5}$ speciation sampler (x2) between 10/1/02 and 2/27/03.

Figure 17 presents a Bland-Altman plot comparing 24-hour average fine particulate SO_4^{2-} concentrations determined by IC from quartz-filter-based samples collected by the SFS sampler at the Lawrenceville site with 24-hour average SO_4^{2-} concentrations determined by IC from nylon-filter-based samples collected by the SASS sampler at that site. **Figure 18** presents a similar plot for fine particulate NO_3^-. As shown in these figures, the SFS sampler exhibited a statistically significant, nonconstant bias relative to the SASS sampler for both SO_4^{2-} and NO_3^-. On average, SO_4^{2-} concentrations determined by the SFS sampler were 0.4 $\mu g/m^3$ less than those determined by the SASS sampler, and NO_3^- concentrations determined by the SFS sampler were 0.9 $\mu g/m^3$ less than those determined by the SASS sampler. The larger relative bias for NO_3^- as compared to SO_4^{2-} likely results from losses of semivolatile NO_3^- from the quartz filters used by the SFS sampler. (The nylon filters used by the SASS sampler are employed to prevent these losses). Irrespective of their causes, the biases between the two samplers are correctable via the calibration lines presented in **Figures 17 (b) and 18 (b)**. Constant common imprecision estimates for the comparisons between the SFS and SASS samplers were 0.3 $\mu g/m^3$ (8.1%) for SO_4^{2-} and 0.4 $\mu g/m^3$ (13.9%) for NO_3^-. These estimates are similar (on a percentage basis) to those presented above for $PM_{2.5}$ mass measurements. Results indicate that, after calibration to account for relative bias, data from the SFS and SASS samplers could be used interchangeably to represent ambient fine particulate ion concentrations. **Figures 19 and 20** present Bland-Altman plots comparing concentrations of SO_4^{2-} and NO_3^-, respectively, measured at the Bruceton monitoring site using a PC-BOSS sampler with concentrations of these species measured there using a $PM_{2.5}$ FRM or Andersen RAAS2.5-400 sampler. Concentrations of SO_4^{2-} measured using the PC-BOSS sampler on average were about 1.0 $\mu g/m^3$ less than those measured using the FRM or Andersen RAAS2.5-400 sampler, whereas concentrations of NO_3^- measured using the PC-BOSS sampler on average were 0.1 $\mu g/m^3$ greater than those measured using the FRM or Andersen RAAS2.5-400 sampler. These statistically significant relative biases likely reflect a combination of differences between the samplers, including the smaller range of particle sizes collected by the PC-BOSS (i.e., 0.1-2.3 $\mu g/m^3$ vs. <2.5 μ g/m^3), the use of a particle concentrator and multi-channel diffusion denuder in the PC-BOSS to scrub gases including SO_2 and HNO_3 that could otherwise contribute to positive sampling artifacts (the FRM and Andersen RAAS2.5-400 samplers at Bruceton were not equipped with denuders), and the use of a Nylasorb filter in the PC-BOSS to collect volatilized NO_3^-. Again, these biases can be corrected using the calibration lines presented in **Figures 19 (b) and 20 (b)**. Common imprecision estimates of 23% for SO_4^{2-} and 37% for NO_3^- computed from the comparison between the PC-BOSS and FRM or Andersen RAAS2.5-400 samplers at the Bruceton site were greater than the common imprecision estimates reported above for the comparison between the SFS and SASS samplers at the Lawrenceville site.

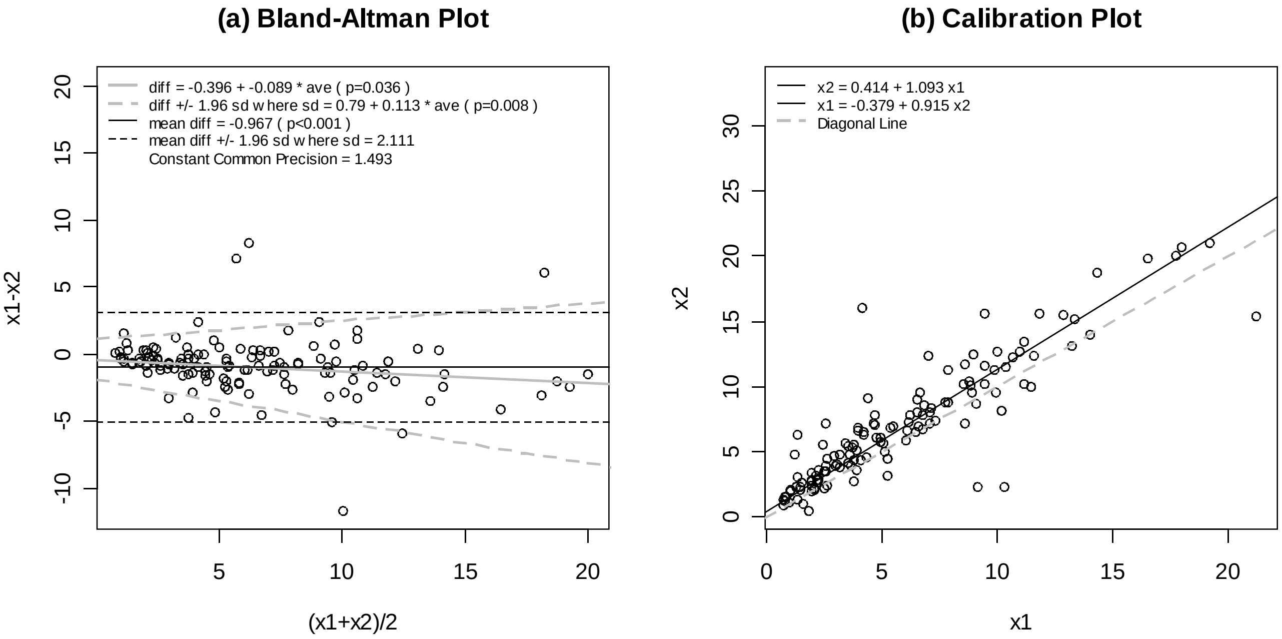


Figure 19: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily SO_4^{2-} concentrations measured at the Bruceton site using a PC-BOSS sampler (x1) and a $PM_{2.5}$ FRM sampler or Andersen RAAS2.5-400 speciation sampler (x2) between 11/2/99 and 11/8/00.

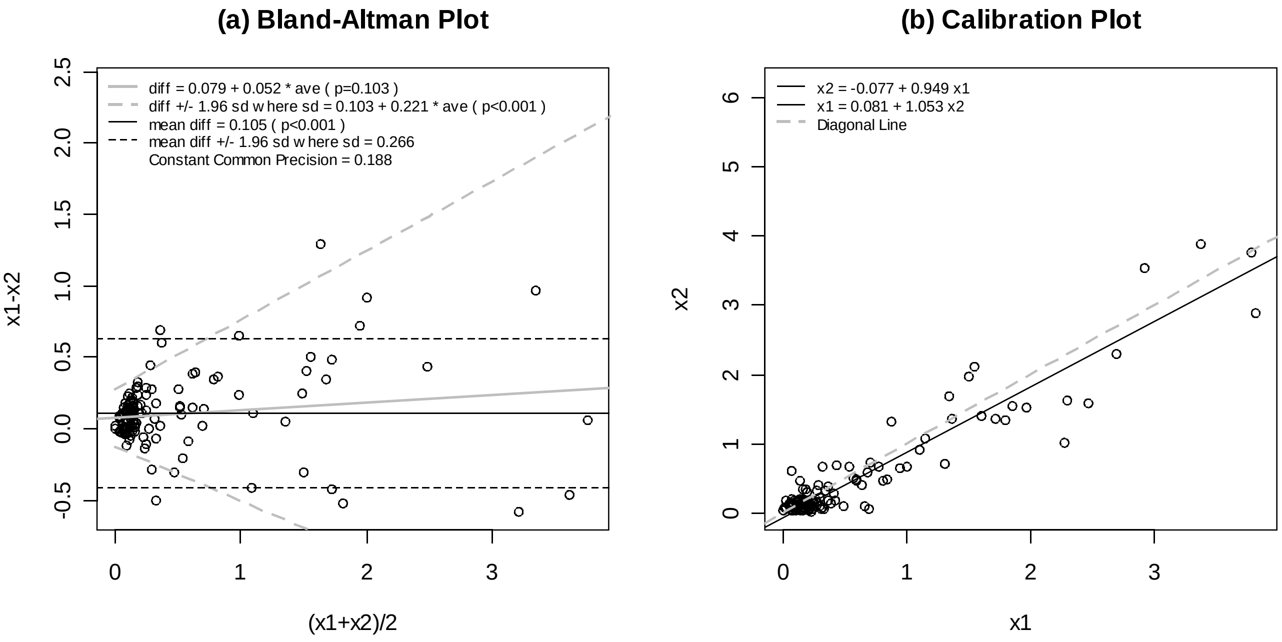


Figure 20: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily NO_3^- concentrations measured at the Bruceton site using a PC-BOSS sampler (x1) and a $PM_{2.5}$ FRM sampler or Andersen RAAS2.5-400 speciation sampler (x2) between 11/2/99 and 11/8/00.

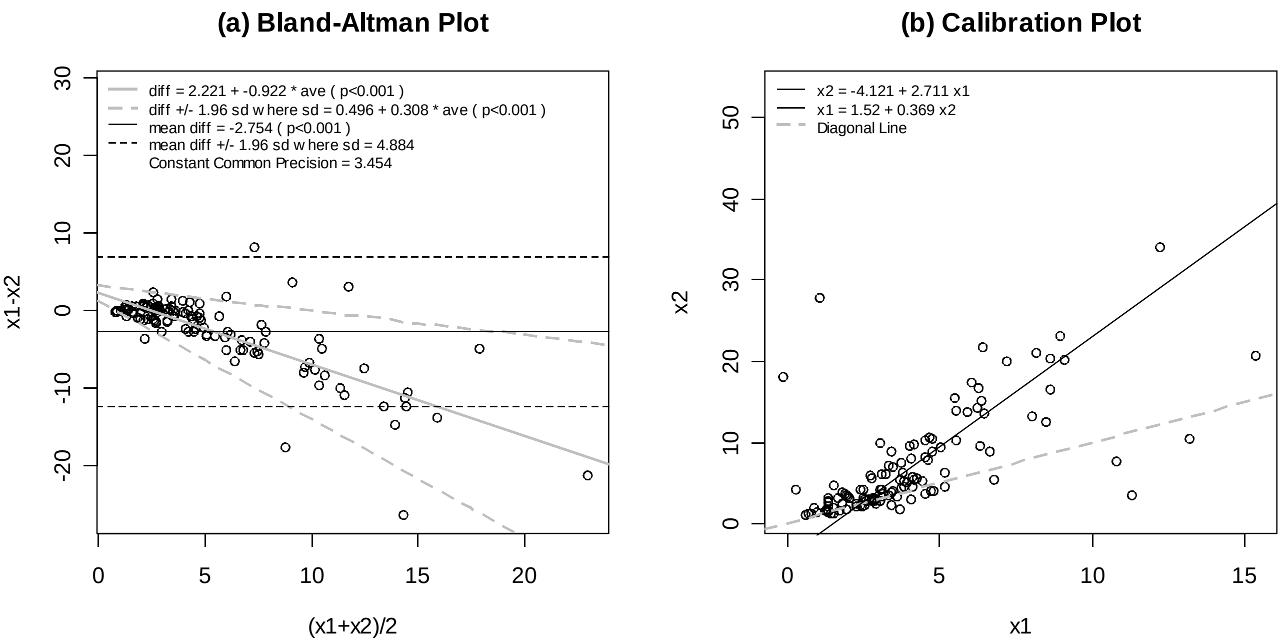


Figure 21: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily SO_4^{2-} concentrations measured at the Bruceton site using an R&P 8400S sampler (x1) and a PM2.5 FRM sampler or Andersen RAAS2.5-400 speciation sampler (x2) between 7/3/01 and 10/10/02.

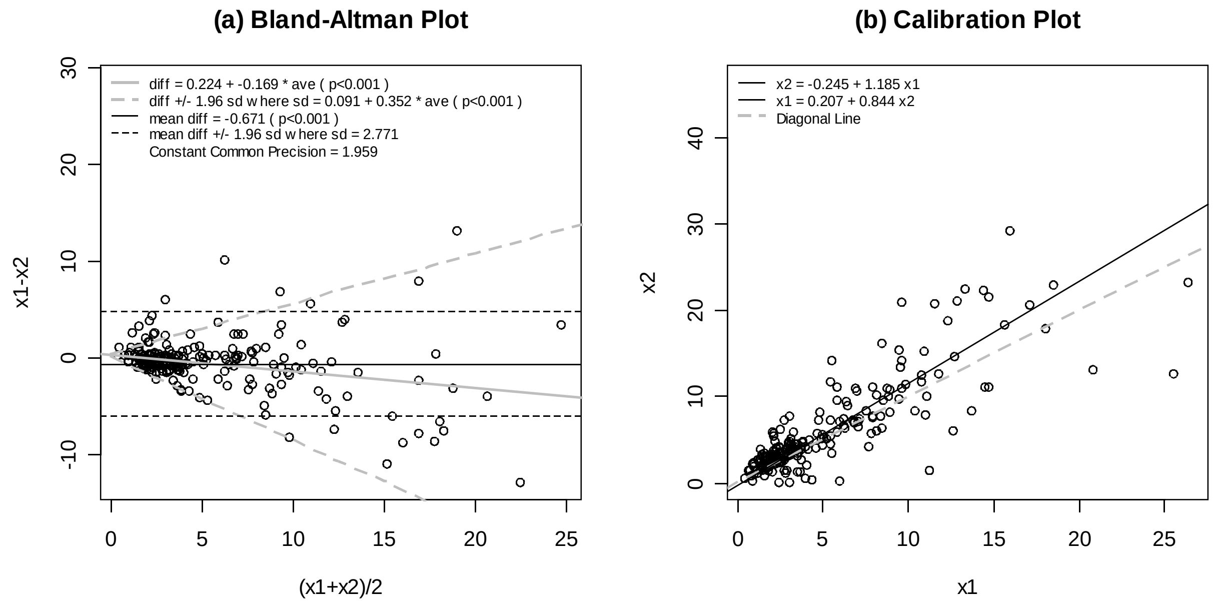


Figure 22: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily SO_4^{2-} concentrations measured at the Schenley Park site using an R&P 8400S sampler (x1) and a CMU $PM_{2.5}$ speciation sampler (x2) between 7/2/01 and 7/21/02.

Finally, **Figures 21 and 22** compare semi-continuous fine particulate SO_4^{2-} measurements made using Rupprecht & Patashnick 8400S sulfate monitors at the Bruceton and Schenley Park sites with integrated SO_4^{2-} measurements made at these sites by IC analysis of Teflon-filter-based $PM_{2.5}$ samples. As shown in the figures, the semi-continuous measurements in both cases exhibited a statistically significant, nonconstant bias relative to the integrated measurements. On average, SO_4^{2-} concentrations measured by the 8400S monitor at the Bruceton site were 2.8 $\mu g/m^3$ less than those determined from $PM_{2.5}$ samples collected using the $PM_{2.5}$ FRM or Andersen RAAS2.5-400 sampler at that site, and SO_4^{2-} concentrations measured by the 8400S monitor at the Schenley Park site were 0.7 $\mu g/m^3$ less than those determined from $PM_{2.5}$ samples collected using the CMU speciation sampler at that site. More importantly, however, the common imprecisions of 64.0% and 37.0% computed from the comparisons of semi-continuous and integrated measurements at the Bruceton and Schenley Park sites, respectively, were substantially greater than the common imprecisions of 8.1% – 23% reported above for pairs of integrated SO_4^{2-} measurement methods. (As discussed later in this report, some data from the Bruceton monitoring site have not undergone extensive QA/QC; hence, invalid outliers may still be present in these data, possibly contributing to the relatively large common imprecision computed for semi-continuous and integrated SO_4^{2-} measurements from this site).

Overall, the results presented in **Figures 17 -22** suggest that fine particulate ion data from various $PM_{2.5}$ FRM and $PM_{2.5}$ speciation samplers can likely be used interchangeably to represent exposures in a $PM_{2.5}$

epidemiology study without substantially increasing the uncertainty in the data, provided that these data are first calibrated to adjust for relative bias. Greater uncertainty is introduced if ion data from the PC-BOSS sampler, which collects particles between 0.1 and 2.3 µm in diameter, are used interchangeably with ion data from the $PM_{2.5}$ FRM and speciation samplers, and a substantial amount of noise is introduced if integrated and semi-continuous ion measurements are used interchangeably. Hence, ion concentrations determined using the semi-continuous and PC-BOSS samplers should be used only when necessary to provide estimates for days that would otherwise have no available data. It is important to recognize that the comparison of ion measurement methods presented here was not exhaustive and is intended to serve only as an example. Prior to specifying ion data for use in the epidemiology study, additional calibrations will need to be performed (using LVM) to allow ion concentrations determined using different methods or at different geographical locations to be corrected to a common basis.

2.4.1.3 $PM_{2.5}$ Elemental and Organic Carbon

Table 11 identifies nine different methods that were employed by the 19 $PM_{2.5}$ speciation monitoring sites in the greater Pittsburgh region to determine ambient concentrations of fine particulate elemental and organic carbon between 1999 and 2005. As with concentrations of fine particulate ionic species, concentrations of EC and OC were measured using both integrated and semi-continuous methods. At all of the AQS speciation sites, integrated $PM_{2.5}$ samples for carbonaceous analysis were collected from 12:00 a.m. to 12:00 a.m. on quartz filters using Met One SASS speciation samplers. Concentrations of EC and OC were determined from these samples by thermal optical transmittance (TOT). Additional EC and OC concentration data at the Lawrenceville site are available from the UORVP and IMPROVE sampling networks. UORVP collected 6- or 24-hour integrated $PM_{2.5}$ samples for carbonaceous analysis on quartz filters using Desert Research Institute SFS samplers (or $PM_{2.5}$ FRM samplers) that operated on a midnight-to-midnight schedule, and IMPROVE collected 24-hour integrated $PM_{2.5}$ samples for carbonaceous analysis on quartz filters using IMPROVE samplers that similarly operated from midnight to midnight. Unlike the quartz-filter-based $PM_{2.5}$ samples from the AQS sites, which were analyzed for EC and OC by TOT, the quartz-filter-based $PM_{2.5}$ samples collected by UORVP and IMPROVE were analyzed for EC and OC by thermal optical reflectance (TOR).

TOT and TOR each determine OC and EC by measuring the amount of carbon that is evolved (via volatilization or oxidation) from the sample as a function of temperature and the O_2 content of the atmosphere surrounding the sample. The amount of light reflected by (TOR) or transmitted through (TOT) the filter-based sample is also measured. As temperature increases in the absence of oxygen, the sample darkens as some OC is charred, causing reflectance / transmittance to decrease. When oxygen is added and the temperature is further increased, the sample lightens as the charred OC and EC are combusted away, causing the reflectance / transmittance to increase such that it returns to and eventually exceeds its original value. The split between OC and EC is operationally defined; all carbon evolved before the reflectance / transmittance returns to its original value is considered to be OC, and all carbon evolved after that point is considered to be EC. In addition to using different techniques to optically monitor changes in light absorption by the sample, TOT and TOR typically use different thermal profiles characterized by different temperature ramping rates, temperature plateaus, and residence times at each plateau. As a result of these differences, whereas concentrations of total carbon (i.e. EC + OC) determined by different thermal/optical methods are typically comparable, the distinction between how much of this carbon is EC and how much of it is OC varies among the methods (Chow et al., 2004; Schauer et al., 2003; Schmid et al., 2001).

Again, apart from the AQS sites, carbonaceous data from the PAQS Schenley Park site and the NETL/OST Bruceton site are of particular interest for the proposed epidemiology study because of these sites' location in Allegheny county and their abundance of daily $PM_{2.5}$ speciation data. Twenty-four-hour average concentrations of EC and OC at the Schenley Park site were determined by TOT from integrated $PM_{2.5}$ samples collected on quartz filters using a CMU TQQQ sampler that operated from approximately 12:00 a.m. to 12:00 a.m. (at times, several separate samples were collected during this period to improve the time resolution of the data). In addition to the primary quartz filter, the TQQQ sampler included a quartz backup filter as well as a separate channel containing a Teflon filter with a quartz backup filter in order to permit an assessment of sampling artifacts. Only data from the primary quartz filter are examined in the analyses presented later in this section, as the speciation samplers (e.g., the SASS and IMPROVE samplers) used by a majority of the monitoring sites in the Pittsburgh region to collect $PM_{2.5}$ samples for carbonaceous analysis were not equipped with backup filters. Most filter-based measurements of EC and OC at the Bruceton site were obtained from samples collected using a PC-BOSS sampler. This sampler was equipped with a diffusion denuder to remove volatile organic compounds (VOCs) from the sample stream and included both a primary quartz filter to collect carbonaceous particles and a charcoal impregnated glass fiber backup filter to collect semivolatile organic material. EC and OC were determined from the filter-based samples by temperature-programmed volatilization (Eatough et al., 1993).

The Schenley Park and Bruceton monitoring sites each included semi-continuous measurements of fine particulate EC and OC as well. EC and OC concentrations were monitored at the Schenley Park site using a Sunset In-Situ Thermal/Optical Carbon Analyzer that was equipped with a multi-channel parallel plate diffusion denuder to scrub gas-phase VOCs from the sample stream. The instrument collected $PM_{2.5}$ samples on quartz fiber filters, and determined EC and OC concentrations from these samples in the field every 1-4 hours using a built-in thermal optical transmittance analyzer. EC and OC concentrations were monitored at the Bruceton site using a Rupprecht & Patashnick Series 5400 Ambient Carbon Particulate Monitor. As discussed in **Section 2.2.2**, these carbon measurements are critical to the feasibility of the retrospective epidemiology study, because they provide the largest source of otherwise scarce daily EC and OC data from the Pittsburgh region during the time period of interest for the study. The Series 5400 carbon monitor consists of two separate sampling trains that are used in alternating fashion to collect 3-hour $PM_{2.5}$ samples on parallel plate impactors for analysis. At the end of each 3-hour sampling period, the chamber containing the $PM_{2.5}$ sample is heated successively to temperature plateaus of 340°C and 750°C in an atmosphere of ambient air. A non-dispersive infrared sensor is used to measure the concentration of CO_2 produced by combustion of carbon contained in the sample. All OC in the sample is assumed to combust during the 340°C temperature plateau, and all EC is assumed to combust during the 750°C temperature plateau. Thus, as with the TOT and TOR techniques, the Series 5400 monitor uses an operational definition to estimate the split between elemental and organic forms of carbon. Unlike TOT and TOR, the Series 5400 monitor does not include any correction for pyrolysis. Rather, any pyrolyzed OC that is formed is measured as part of the EC fraction.

Although not indicated in **Table 11** the monitoring programs that operated in the Pittsburgh region between 1999 and 2005 did not all follow consistent procedures for blank-correcting data prior to reporting these data. For example, all $PM_{2.5}$ speciation data reported by the PAQS program (i.e., for the Schenley Park monitoring site) were corrected to account for possible sample contamination during collection, handling, storage, and analysis, whereas none of the $PM_{2.5}$ speciation data reported by the AQS sites were blank corrected (although blank concentrations were determined and are available for future blank correction, if desired). Particularly high blank concentrations are often observed for carbon measurements; hence, discrepancies in blank correction procedures may affect the relative bias (if blanks tend to be constant) or imprecision (if blanks tend to be variable) among these measurements.

Because of the limited availability of carbon data from the Pittsburgh region during the time period of interest for the proposed epidemiology study, which results largely from the fact that these data generally cannot be obtained retrospectively via analysis of archived $PM_{2.5}$ samples, and because of the methodological discrepancies discussed above regarding the determination of EC and OC, the current feasibility assessment included a relatively thorough evaluation of the comparability of carbon data that might be used in the epidemiology study. **Figure 23** presents a Bland-Altman plot comparing 24-hour average fine particulate OC concentrations determined by TOR from quartz-filter-based samples collected by the SFS sampler at the Lawrenceville site with 24-hour average OC concentrations determined by TOT from quartz-filter-based samples collected by the SASS sampler at that site. **Figure 24** presents a similar plot for fine particulate EC. Concentrations of both OC and EC determined by TOR from the SFS samples exhibited statistically significant, nonconstant biases relative to those determined by TOT from the SASS samples. These biases can be corrected via the calibration lines provided in **Figures 23 (b) and 24 (b).** It is more noteworthy, however, that the constant common imprecision estimates for the comparison of the SFS and SASS OC measurements and for the comparison of the SFS and SASS EC measurements, which were each 28% (1.1 $\mu g/m^3$ for OC and 0.26 $\mu g/m^3$ for EC), were substantially greater than the imprecision estimates of 8.1% and 14% computed for comparisons of SO_4^{2-} and NO_3^- concentrations, respectively, determined using these samplers. The greater imprecision for the carbon measurements as compared to the ion measurements likely arises primarily because of inconsistencies between the TOT and TOR analytical techniques used to analyze samples collected by the SASS and SFS samplers, respectively.

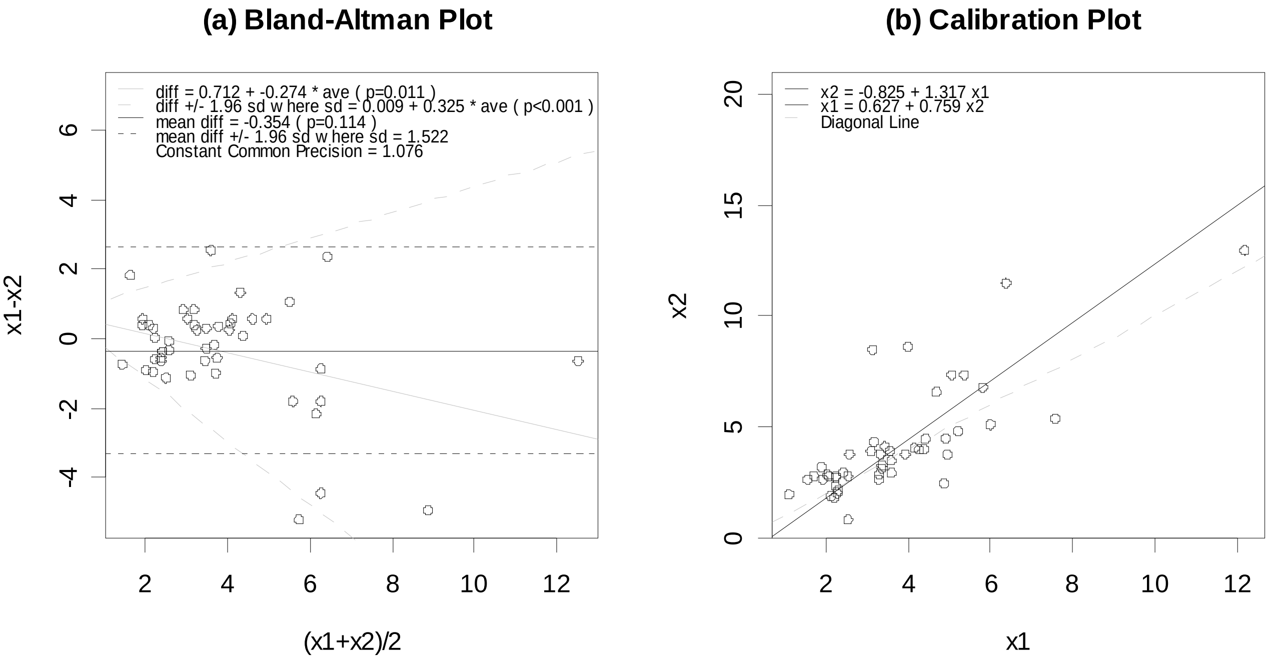


Figure 23: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily OC concentrations measured at the Lawrenceville site using a Desert Research Institute Sequential Filter Sampler with TOR analysis (x1) and a Met One SASS $PM_{2.5}$ speciation sampler with TOT analysis (x2) between 10/1/02 and 2/27/03.

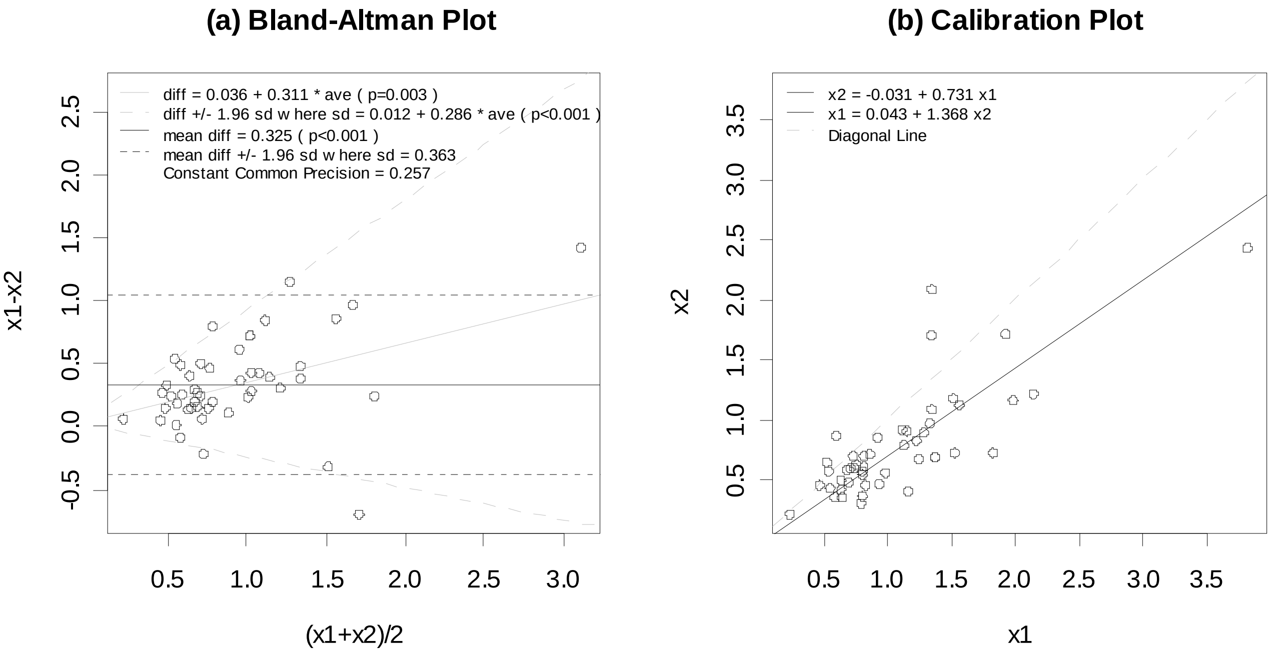


Figure 24: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily EC concentrations measured at the Lawrenceville site using a Desert Research Institute Sequential Filter Sampler with TOR analysis (x1) and a Met One SASS $PM_{2.5}$ speciation sampler with TOT analysis (x2) between 10/1/02 and 2/27/03.

Figures 25 and 26 present the results of Bland-Altman analyses comparing concentrations of OC and EC, respectively, determined by TOT from quartz-filter-based $PM_{2.5}$ samples collected using the CMU TQQQ sampler at the Schenley Park site with concentrations of these species determined by TOT from quartz-filter-based $PM_{2.5}$ samples collected using the SASS sampler at the Lawrenceville site. Concentrations of OC and EC at the Schenley Park site exhibited a statistically significant, negative bias relative to concentrations of these species at the Lawrenceville site. On average, OC concentrations at Schenley Park were about 1.5 $\mu g/m^3$ less than OC concentrations at Lawrenceville, and EC concentrations at Schenley Park were about 0.2 $\mu g/m^3$ less than EC concentrations at Lawrenceville. These biases likely result in part from the fact that data from the Schenley Park site were blank-corrected, whereas data from the Lawrenceville site were not. They may also reflect true differences in ambient concentrations between the two sites, which were located about 3 km from each other. Regardless of the cause of the relative bias, the results presented in **Figures 25 and 26** suggest that, after calibration, TOT data from the Lawrenceville and Schenley Park sites could be used interchangeably to represent ambient OC for use in the proposed epidemiology study. The constant common imprecision estimated for the OC concentrations determined by TOT at these two sites was 19% (0.8 $\mu g/m^3$), which is less than the constant common imprecision of 28% (1.1 $\mu g/m^3$) reported above for collocated determinations of OC by TOT and TOR at the Lawrenceville site. The constant common imprecision estimated for the EC concentrations determined by TOT at the Lawrenceville and Schenley Park sites was 30% (0.21 $\mu g/m^3$). This greater percent imprecision for EC than for OC is expected, because analytical uncertainty is typically greater for EC than for OC (Pun et al., 2004), and because concentrations of EC, which has only primary (local) emission sources, are expected to be more spatially variable than concentrations of OC, which has both primary (local) and secondary (regional) sources.

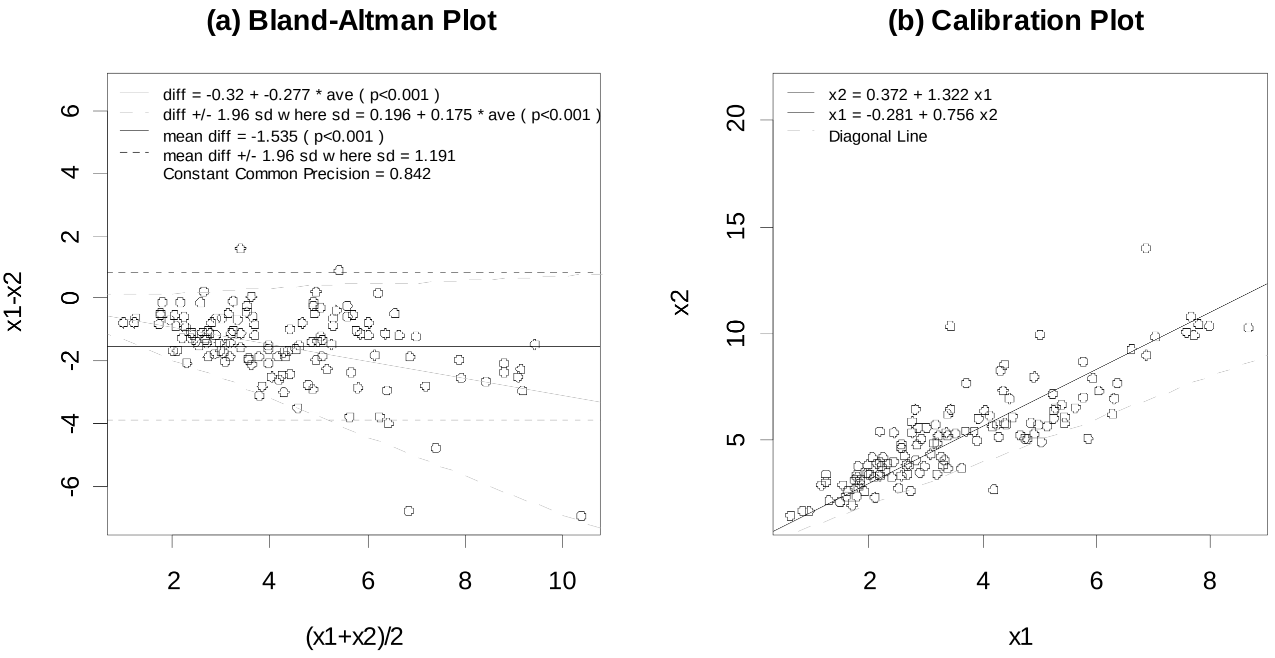


Figure 25: Bland-Altman plot (a) and corresponding calibration plot (b) for pairwise daily OC concentrations measured at the Schenley Park site using a CMU TQQQ sampler with TOT analysis (x1) and at the Lawrenceville site using a Met One SASS $PM_{2.5}$ speciation sampler with TOT analysis (x2) between 6/30/01 and 7/31/02.

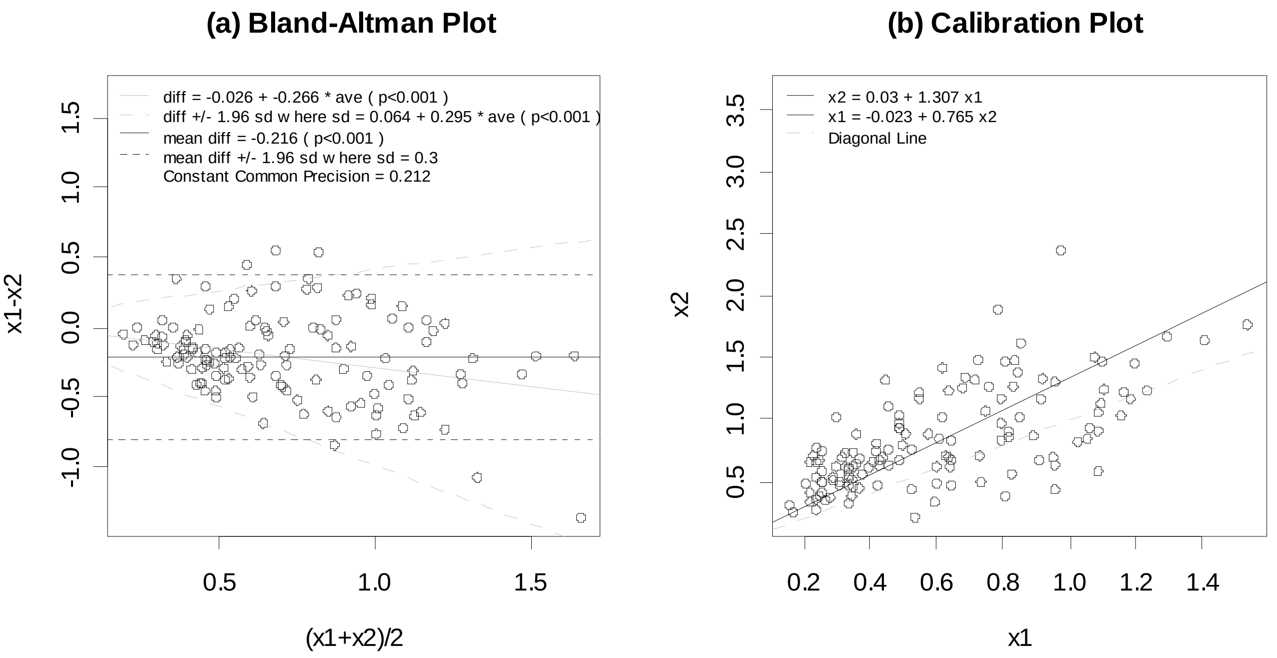


Figure 26: Bland-Altman plot (a) and corresponding calibration plot (b) for pairwise daily EC concentrations measured at the Schenley Park site using a CMU TQQQ sampler with TOT analysis (x1) and at the Lawrenceville site using a Met One SASS $PM_{2.5}$ speciation sampler with TOT analysis (x2) between 6/30/01 and 7/31/02.

Unlike semi-continuous fine particulate sulfate measurements, which as discussed in **Section 2.4.1.2** were characterized by substantial noise relative to integrated filter-based measurements, semi-continuous measurements of fine particulate OC and EC made at the Bruceton and Schenley Park sites were generally comparable (after correction for relative bias) to integrated filter-based measurements of these carbonaceous species. **Figures 27 and 28** present the results of Bland-Altman analyses comparing concentrations of OC and EC, respectively, determined by the R&P Series 5400 ambient carbon particulate monitor at the Bruceton site with corresponding concentrations of these species determined by TOT from quartz-filter-based $PM_{2.5}$ samples collected using the SASS sampler at the Lawrenceville site. **Figures 29 and 30** present similar comparisons between semi-continuous OC and EC measurements from the Bruceton site and corresponding integrated OC and EC measurements at the Schenley Park site. (Semi-continuous data were aggregated to compute daily averages for comparison with the integrated measurements). The semi-continuous OC and EC measurements at the Bruceton site exhibited sizable, statistically significant, nonconstant biases relative to the integrated measurements at the Lawrenceville and Schneley Park sites, likely because of differences both in measurement methods and in site locations. On average, OC concentrations measured by the R&P Series 5400 monitor at the Bruceton site were 1.8 $\mu g/m^3$ less than those measured by the TQQQ sampler at the Schenley Park site and 3.6 $\mu g/m^3$ less than those measured by the SASS sampler at the Lawrenceville site. EC concentrations measured by the R&P Series 5400 monitor at the Bruceton site were on average 0.27 $\mu g/m^3$ less than those measured by the TQQQ sampler at Schenley Park and 0.54 $\mu g/m^3$ less than those measured by the SASS sampler at Lawrenceville. These relative biases, although large, are correctable using the calibration curves shown in **Figures 27(b), 28(b), 29(b) and 30(b)**. Common imprecision estimates for comparisons between the semi-continuous carbon measurements made at the Bruceton site and the integrated carbon measurements made at the Lawrenceville and Schenley Park sites were similar to those reported above for comparisons between integrated measurements at the latter two sites. Imprecision estimates for the comparisons of semi-continuous OC from the Bruceton site with integrated OC from the Lawrenceville and Schenley Park sites were 1.2 $\mu g/m^3$ and 0.9 $\mu g/m^3$, respectively, consistent with the estimates of 1.1 $\mu g/m^3$ and 0.8 $\mu g/m^3$ shown in **Figures 23 and 25** for comparisons among integrated OC measurements. Likewise, imprecision estimates for the comparisons of semi-continuous EC from the Bruceton site with integrated EC from the Lawrenceville and Schenley Park sites were 0.23 $\mu g/m^3$ and 0.18 $\mu g/m^3$, respectively, consistent with the estimates of 0.26 $\mu g/m^3$ and 0.21 $\mu g/m^3$ shown in **Figures 24 and 26** for comparisons among integrated EC measurements. Hence, the comparisons presented in **Figures 27-30** suggest that semi-continuous measurements of EC and OC made at the Bruceton monitoring site, once corrected for relative bias, are generally commensurate with integrated measurements of these species made at other monitoring sites in the Pittsburgh region. Therefore, these semi-continuous measurements, which are the largest source of ambient carbon data available from the Pittsburgh region during the time period of interest for the proposed epidemiology study, are an appropriate source of exposure data for the study.

Semi-continuous OC and EC measurements from the Schenley Park monitoring site are also suitable for use in the epidemiology study. **Figues 31 and 32** present the results of Bland-Altman analyses comparing concentrations of OC and EC, respectively, determined at the Schenley Park site using the Sunset In-Situ Thermal/Optical Carbon Analyzer with concentrations determined there using the collocated CMU TQQQ sampler with TOT analysis. For both OC and EC, there was a statistically significant but correctable relative bias between the two methods. Otherwise, however, the methods agreed as well or better than the methods compared above. Agreement between the semi-continuous and integrated measurements at the Schenley site was particularly strong for OC, for which the estimated common imprecision was only 0.4 µg/m^3 (about 14%). The common imprecision estimate for EC measurements was 0.23 µg/m^3 (about 34%), consistent with the imprecision estimates reported above for EC.

Thus, the results presented here collectively suggest that OC measurements that were collected at different monitoring sites or using different measurement techniques generally can be used interchangeably to represent exposures in Allegheny County after correction for relative biases. EC concentrations measured at different sites or using different techniques are characterized by larger amounts of random error than are OC concentrations, owing to the greater spatial variability and larger analytical uncertainty associated with EC. Hence, the possibility of exposure misclassification for $PM_{2.5}$ components such as EC must be considered when designing the proposed epidemiology study.

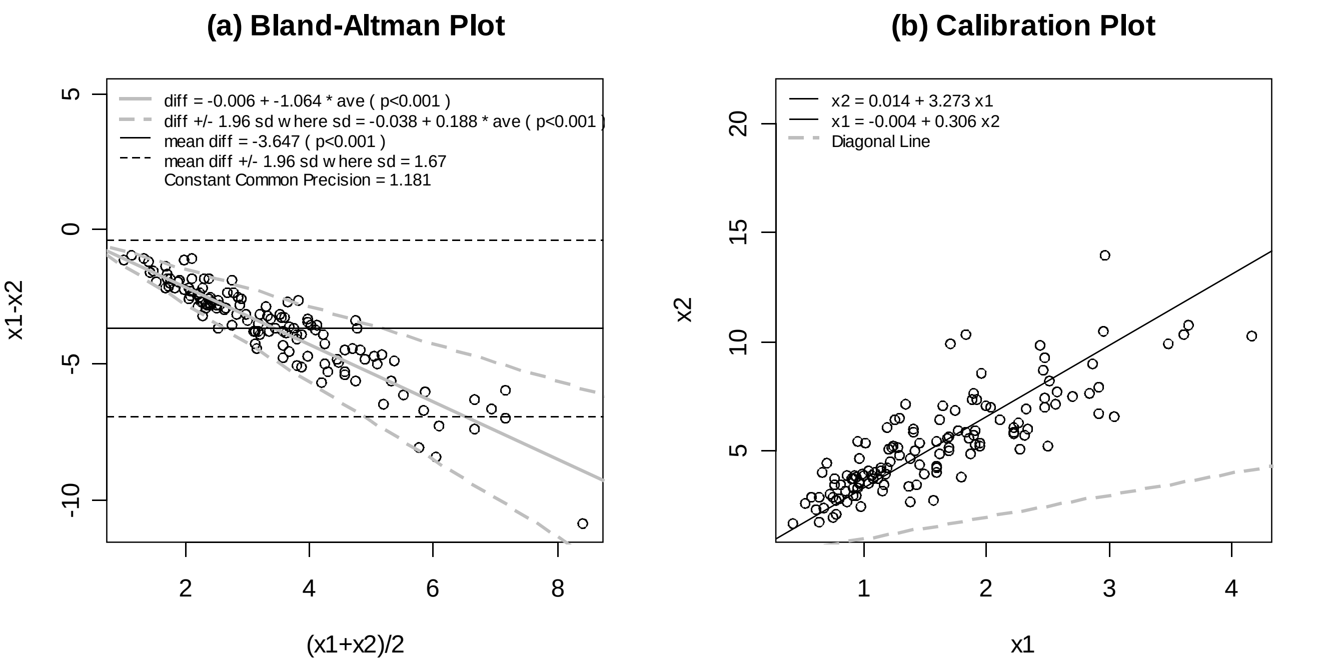


Figure 27: Bland-Altman plot (a) and corresponding calibration plot (b) for pairwise daily OC concentrations measured at the Bruceton site using an R&P Series 5400 Ambient Carbon Particulate Monitor (x1) and at the Lawrenceville site using a Met One SASS $PM_{2.5}$ speciation sampler with TOT analysis (x2) between 7/1/01 and 10/26/02.

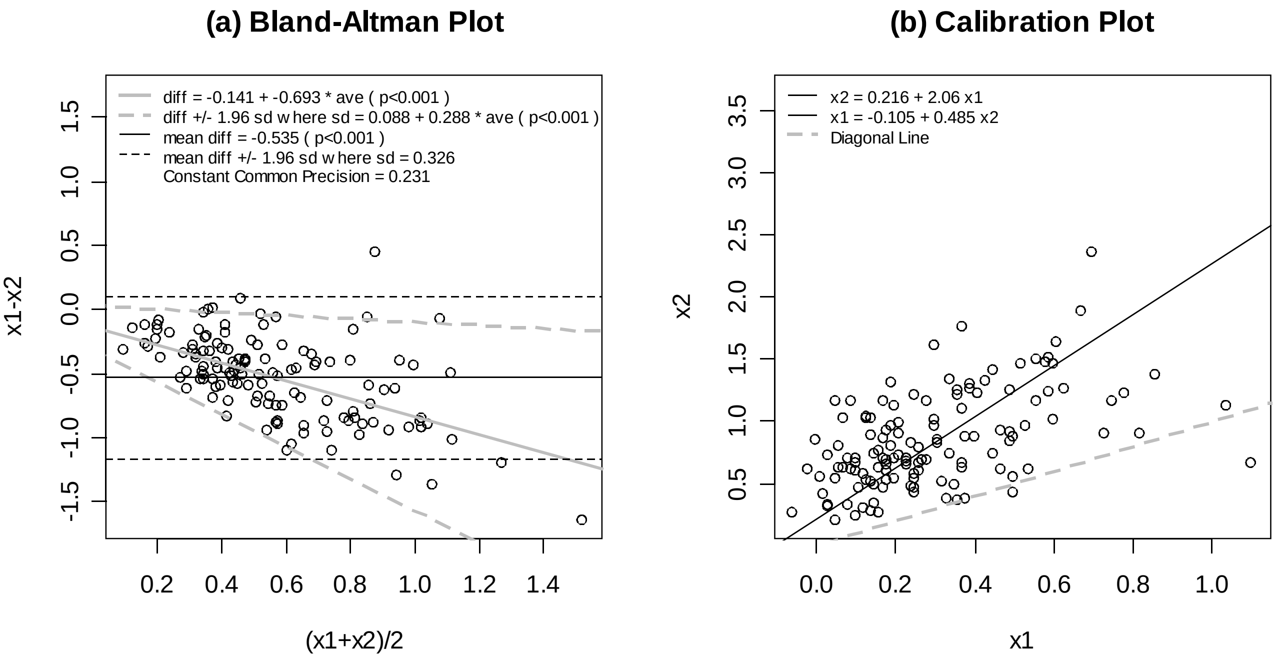


Figure 28: Bland-Altman plot (a) and corresponding calibration plot (b) for pairwise daily EC concentrations measured at the Bruceton site using an R&P Series 5400 Ambient Carbon Particulate Monitor (x1) and at the Lawrenceville site using a Met One SASS $PM_{2.5}$ speciation sampler with TOT analysis (x2) between 7/1/01 and 10/26/02.

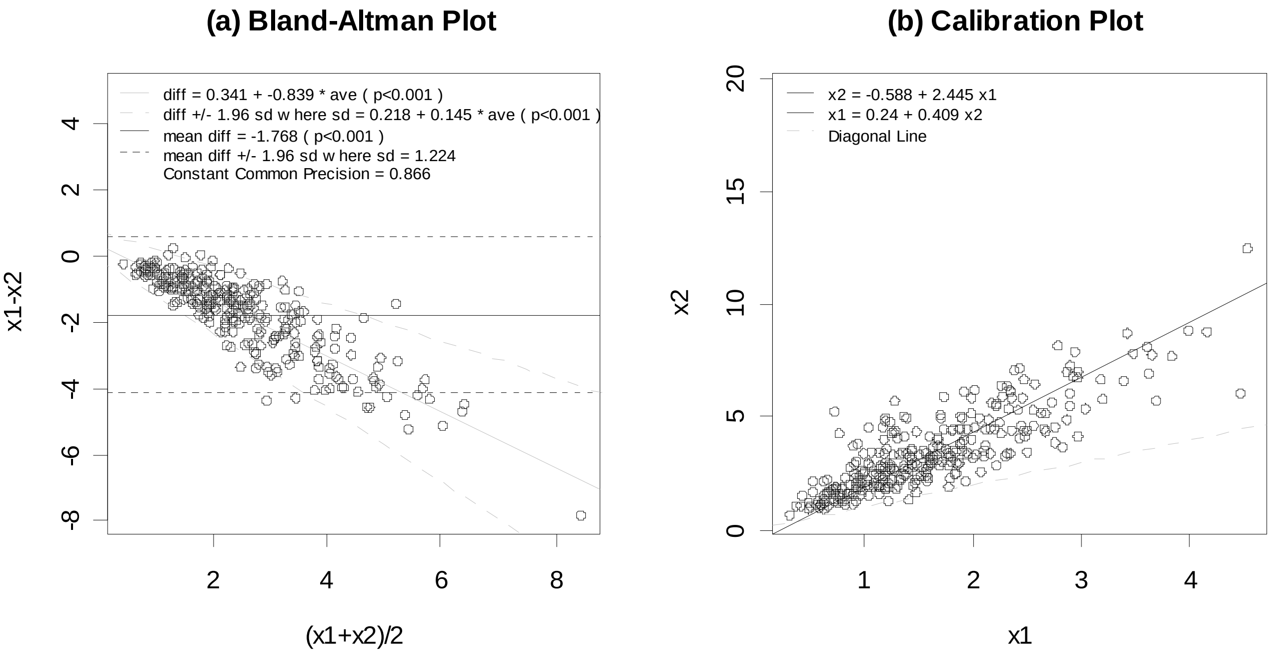


Figure 29: Bland-Altman plot (a) and corresponding calibration plot (b) for pairwise daily OC concentrations measured at the Bruceton site using an R&P Series 5400 Ambient Carbon Particulate Monitor (x1) and at the Schenley Park site using a CMU TQQQ sampler with TOT analysis (x2) between 7/1/01 and 7/31/02.

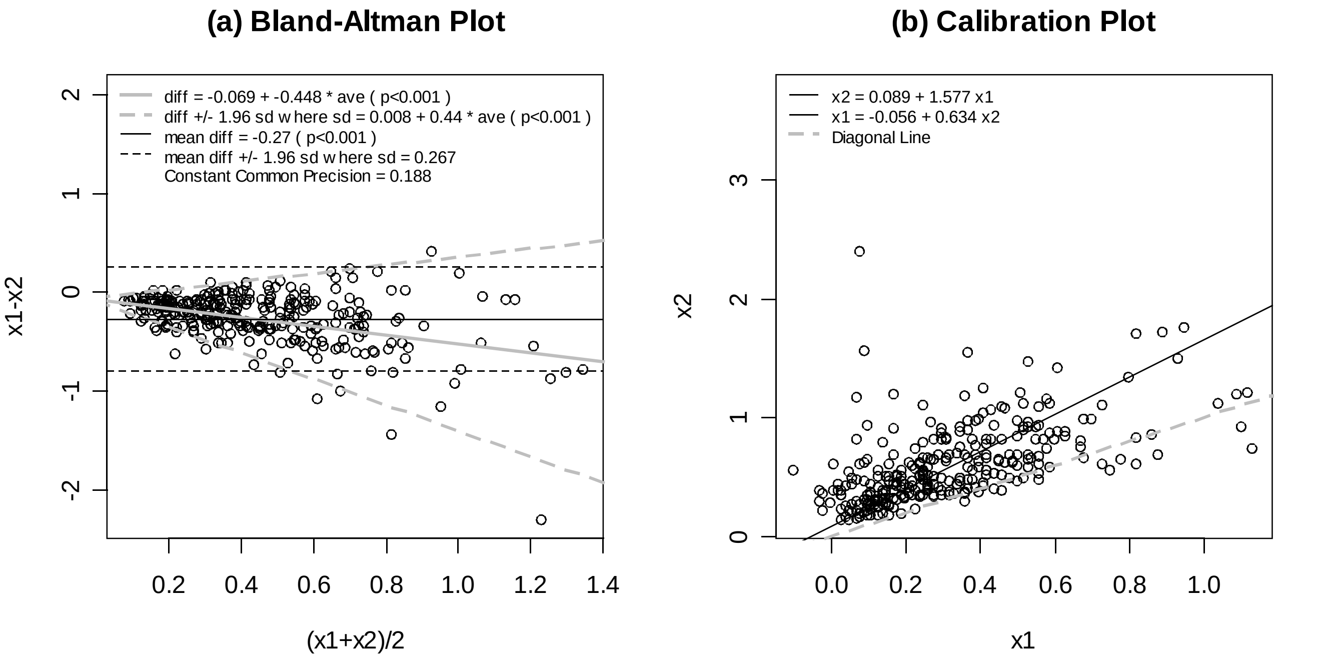


Figure 30: Bland-Altman plot (a) and corresponding calibration plot (b) for pairwise daily EC concentrations measured at the Bruceton site using an R&P Series 5400 Ambient Carbon Particulate Monitor (x1) and at the Schenley Park site using a CMU TQQQ sampler with TOT analysis (x2) between 7/1/01 and 7/31/02.

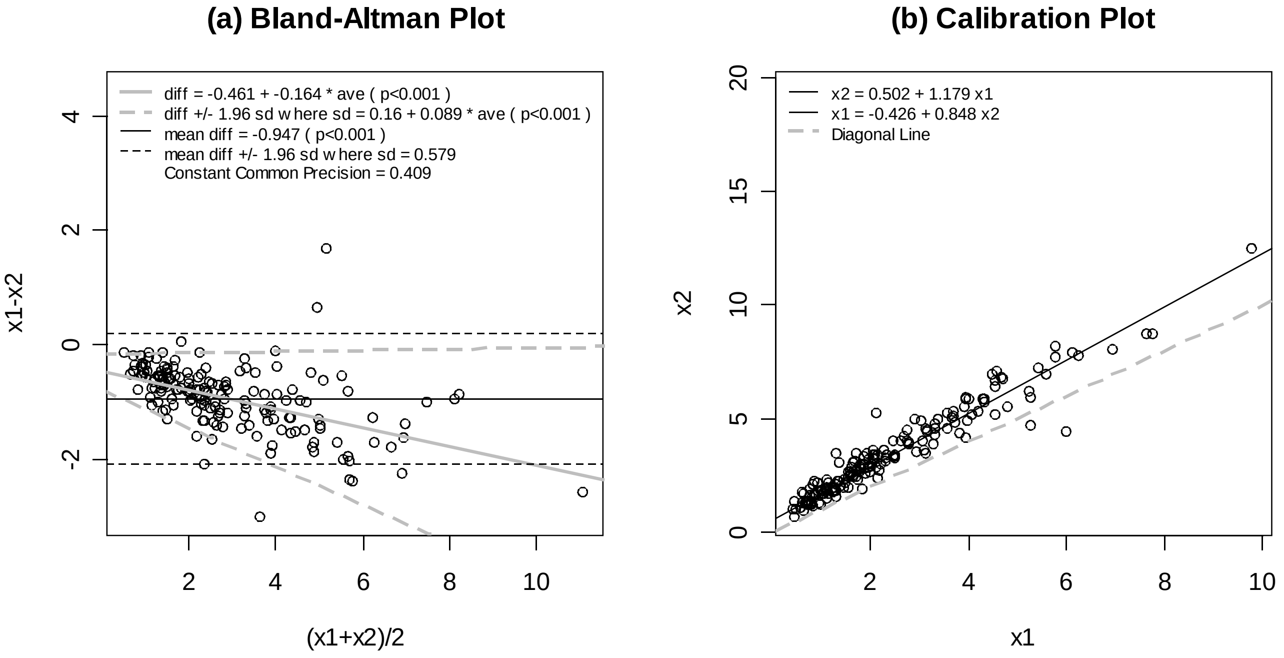


Figure 31: Bland-Altman plot (a) and corresponding calibration plot (b) for pairwise daily OC concentrations measured at the Schenley Park site using a Sunset In-Situ Thermal/Optical Carbon Analyzer (x1) and a CMU TQQQ sampler with TOT analysis (x2) between 7/2/01 and 7/31/02.

2.4.1.4 $PM_{2.5}$ Trace and Crustal Elements

Table 11 identifies seven different methods employed by monitoring sites in the 35-county greater Pittsburgh region to determine ambient concentrations of fine particulate trace and crustal elements between 1999 and 2005. All of the trace and crustal element measurements were made by laboratory analysis of integrated, filter-based $PM_{2.5}$ samples. With the exception of the PAQS Schenley Park site, all of the monitoring sites collected $PM_{2.5}$ samples for elemental analysis on Teflon filters using a $PM_{2.5}$ FRM or speciation sampler with no denuder. At Schenley Park, samples for elemental analysis were collected on cellulose filters using a high-volume (Hi-Vol) $PM_{2.5}$ sampler. As with $PM_{2.5}$ total mass, $PM_{2.5}$ ions, and $PM_{2.5}$ EC and OC, 24-hour integrated $PM_{2.5}$ samples for elemental analysis were generally collected from 12:00 p.m. to 12:00 p.m. at the Bruceton site, from 9:00 a.m. to 9:00 a.m. at the Franciscan University of Steubenville site, and from 12:00 a.m. to 12:00 a.m. at all other sites. Hence, per the discussion in **Section 2.4.1.1**, for the Bruceton and Franciscan University of Steubenville sites, midnight-to-midnight average concentrations would need to be estimated from the available elemental data prior to using these data in the epidemiology study.

There are several limitations associated with characterizing exposures to trace and crustal elemental components of $PM_{2.5}$ for purposes of an epidemiology study. First, these elements, like elemental carbon, are primary pollutants that are expected to have spatially variable concentrations resulting from the influence of localized emission sources (or lack thereof). Hence, the behavior of ambient concentrations measured at a given monitoring site may not reflect the behavior of ambient concentrations measured in other parts of the region. Moreover, unlike the ionic and carbonaceous components of $PM_{2.5}$ discussed above, which are generally present in the ambient air in $\mu g/m^3$ quantities, many elemental components of $PM_{2.5}$ are present only in ng/m^3 quantities. As a result, integrated $PM_{2.5}$ samples contain only very small amounts of these elements (especially if they were collected at low sampling flow rates or over short sampling periods), and the ability to detect the elements depends strongly on the sensitivity of the analytical method being used.

Several different analytical methods were used to determine elements from filter-based $PM_{2.5}$ samples collected in the Pittsburgh region between 1999 and 2005. All elemental determinations at the AQS monitoring sites, CASTNet monitoring sites, and UORVP monitoring sites were performed using X-ray fluorescence spectroscopy (XRF). Elemental determinations at the NETL/OST Bruceton monitoring site were performed using proton induced X-ray emission spectroscopy (PIXE), and elemental determinations at the IMPROVE monitoring sites were performed using both XRF and PIXE, depending on the element being determined. The PAQS and SCAMP programs each used inductively coupled plasma - mass spectrometry (ICP-MS) to determine elements in $PM_{2.5}$ samples. PAQS used conventional low-resolution ICP-MS, whereas SCAMP used dynamic reaction cell (DRC) ICP-MS. Because of differences in the capabilities of these analytical methods and in the objectives of the groups conducting the measurements, the suite of elements that was routinely determined when analyzing $PM_{2.5}$ samples varied from group to

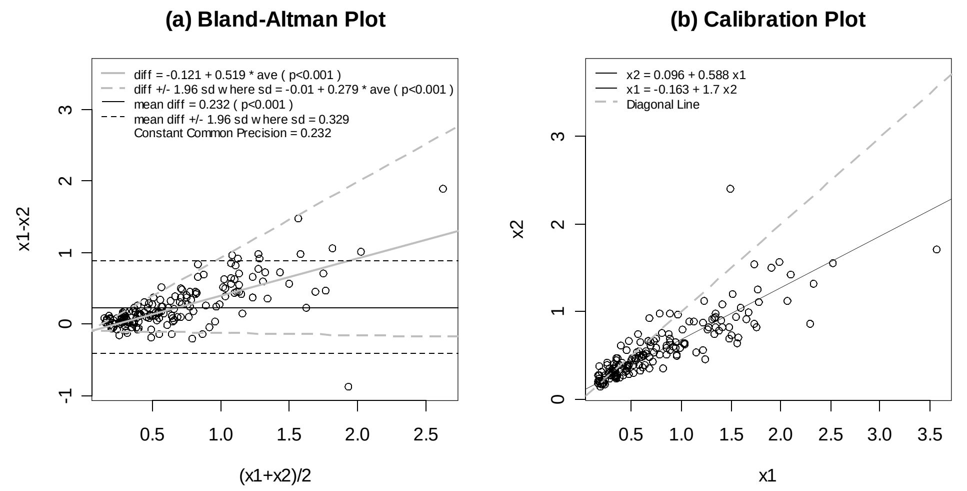


Figure 32: Bland-Altman plot (a) and corresponding calibration plot (b) for pairwise daily EC concentrations measured at the Schenley Park site using a Sunset In-Situ Thermal/Optical Carbon Analyzer (x1) and a CMU TQQQ sampler with TOT analysis (x2) between 7/2/01 and 7/31/02.

group. As part of the inventory of $PM_{2.5}$ speciation data described in Section 2.1 above, we developed a list of 40 elemental components of $PM_{2.5}$ that would be of interest for the proposed retrospective epidemiology study because of their possible implications for human health and/or source apportionment analysis. **Table 13** shows which of these elements were routinely determined by each of the seven monitoring campaigns that performed $PM_{2.5}$ speciation sampling in the Pittsburgh region between 1999 and 2005.

Table 13: Summary of elements that were routinely determined by the various campaigns that monitored $PM_{2.5}$ speciation in the 35-county greater Pittsburgh region between 1999 and 2005.

Monitoring Campaign	AQS	CASTNet	IMPROVE	NETL/OST	PAQS	SCAMP	UORVP
Analytical Method	XRF	XRF	XRF, PIXE	PIXE	ICP-MS	DRC ICP-MS	XRF
Ag	X	X			X		X
Al	X	X	X	X		X	X
As	X	X	X	X	X	X	X
Au	X						X
Ba	X	X		X	X	X	X
Be					X		
Br	X	X	X	X			X
Ca	X	X	X	X	X	X	X

Monitoring Campaign	AQS	CASTNet	IMPROVE	NETL/OST	PAQS	SCAMP	UORVP
Analytical Method	XRF	XRF	XRF, PIXE	PIXE	ICP-MS	DRC ICP-MS	XRF
Cd	X	X			X	X	X
Ce	X				X		
Cl	X	X	X	X			X
Co	X	X		X	X	X	X
Cr	X	X	X	X	X		X
Cs	X				X		
Cu	X	X	X	X	X	X	X
Fe	X	X	X	X	X	X	X
Ga	X	X		X	X		X
Hg	X	X					X
K	X	X	X	X	X	X	X
La	X	X					X
Li					X		
Mg	X	X	X		X	X	X
Mn	X	X	X	X	X	X	X
Mo	X	X			X		X
Na	X	X	X				X
Ni	X	X	X	X	X	X	X
P	X	X	X	X			X
Pb	X	X	X	X	X	X	X
Rb	X	X	X	X	X		X
S	X	X	X	X			X
Sb	X	X			X		X
Se	X	X	X	X	X	X	X
Si	X	X	X	X			X
Sn	X	X				X	X
Sr	X	X	X	X	X		X
Ti	X	X	X	X	X	X	X
V	X	X	X	X	X	X	X
W	X						
Zn	X	X	X	X	X	X	X
Zr	X	X	X	X			X

As shown in the table, 12 of the elements (As, Ca, Cu, Fe, K, Mn, Ni, Pb, Se, Ti, V, and Zn) were routinely determined by all seven campaigns; seven of the elements (Al, Ba, Co, Cr, Mg, Rb, and Sr) were routinely determined by six of the seven campaigns, and seven of the elements (Br, Cd, Cl, Ga, P, S, and Si) were routinely determined by five of the seven campaigns.

Most of the PM$_{2.5}$ elemental data from the Pittsburgh region between 1999 and 2005 were determined by XRF analysis of Teflon-filter-based samples. In XRF, the filter-based PM$_{2.5}$ sample is irradiated by high-energy X-rays, which cause inner orbital electrons to be ejected from the atoms contained in the sample.

Electrons from higher energy, outer orbitals then move to fill the vacancies left by these ejected electrons, resulting in the release of X-ray photons with energies equal to the differences in energy between the outer and inner orbitals. The energies of the emitted photons are unique to each element, and the number of photons emitted corresponds to the amount of the element present in the sample; hence, by counting the number of photons emitted by the sample as a function of energy, concentrations of elements in the sample can be quantified. PIXE is similar to XRF, except that it uses protons rather than X-rays to irradiate the sample. XRF and PIXE are a nondestructive techniques; hence the $PM_{2.5}$ sample is preserved and can be analyzed by other methods once elemental analysis is completed. Moreover, as discussed in Section 2.2, these methods are capable of determining a large suite of elements, including those with atomic numbers between 11 (sodium) and 92 (uranium). However, XRF techniques that have conventionally been used to analyze $PM_{2.5}$ samples in many cases do not have sufficient sensitivity to determine certain elements at the low concentrations in which they are found in these samples.

As an example of this, **Table 14** compares mean ambient air concentrations of trace and crustal elements determined by XRF from 24-hour integrated $PM_{2.5}$ samples collected at the AQS Lawrenceville site with mean method detection limits (MDLs) reported for the determination of these elements by XRF. The percentage of daily observations for which the ambient air concentration was less than the MDL and the median signal-to-noise ratio (i.e., daily ambient air concentration divided by the corresponding measurement uncertainty) are also indicated for each element. AQS began reporting MDLs and measurement uncertainties in July 2003; hence, the comparison is based on data collected between July 14, 2003, and December 27, 2005. For 25 of the 38 elements, ambient air concentrations were less than the MDL for a majority (i.e., >50%) of observations. This suggests that XRF did not have sufficient sensitivity to determine these elements; hence, the reported concentrations likely contain substantial noise. The signal-to-noise ratios presented in **Table 14** confirm this statement; all but two of these 25 elements had median signal-to-noise ratios of ≤ 1.0. Of the thirteen remaining elements, five (Ca, Fe, K, S, Zn) were characterized by less than 10% of observations below the MDL; two (Br, Si) were characterized by 10-20% of observations below the MDL, and six (As, Cr, Cu, Mn, Pb, Se) were characterized by 20-50% of observations below the MDL. These 13 elements had median signal-to-noise ratios ranging from 2.3 (As) to 19.4 (S). (For comparison, only 2.8% of the EC concentrations and none of the $PM_{2.5}$, SO_4^{2-}, NO_3^-, or OC concentrations measured at the Lawrenceville site between July 14, 2003, and December 27, 2005, were less than the corresponding MDL. Median signal-to-noise ratios for these species ranged from 2.6 for EC to 12.3 for S). Hence, in spite of the large suite of elements routinely determined by XRF, only the 13 elements identified above (and possibly Ni and Ti, which had median signal-to-noise ratios of 2.2-2.5 and just over 50% of their observations below the MDL) were measured with sufficient sensitivity to warrant their inclusion in an epidemiology study.

Table 14: MDLs and signal-to-noise ratios for $PM_{2.5}$ elements determined by XRF at the Lawrenceville site between 7/14/03 and 12/27/05.

Element	**Average Ambient Concentration (µg/m^3)**	**Average MDL (µg/m^3)**	**Average Ambient Concentration ÷ Average MDL**	**Percent of Observations <MDL**	**Median Signal-to-Noise Ratio[a]**
Ag	0.0026	0.0103	0.3	92.3%	0.0
Al	0.0156	0.0196	0.8	66.8%	0.0
As	0.0029	0.0023	1.2	47.8%	2.3
Au	0.0013	0.0063	0.2	93.9%	0.0
Ba	0.0103	0.0350	0.3	79.8%	0.0
Br	0.0045	0.0019	2.4	17.4%	5.1
Ca	0.0432	0.0064	6.8	1.2%	7.9
Cd	0.0027	0.0115	0.2	93.1%	0.0
Ce	0.0083	0.0524	0.2	86.2%	0.0
Cl	0.0167	0.0095	1.8	59.9%	0.0
Co	0.0002	0.0016	0.1	97.6%	0.0
Cr	0.0047	0.0021	2.3	41.7%	3.4
Cs	0.0027	0.0256	0.1	96.4%	0.0
Cu	0.0052	0.0022	2.3	25.5%	5.0
Fe	0.1400	0.0023	60.8	0.4%	18.1
Ga	0.0005	0.0040	0.1	96.0%	0.0
Hg	0.0015	0.0047	0.3	85.8%	0.0
K	0.0792	0.0087	9.2	0.0%	8.7
La	0.0060	0.0406	0.1	87.9%	0.0
Mg	0.0052	0.0248	0.2	92.3%	0.0
Mn	0.0075	0.0022	3.4	23.5%	5.4
Mo	0.0012	0.0078	0.2	97.2%	0.0
Na	0.0388	0.0794	0.5	76.1%	0.0
Ni	0.0019	0.0017	1.1	55.5%	2.2
P	0.0022	0.0098	0.2	93.1%	0.0
Pb	0.0111	0.0050	2.2	25.1%	3.8
Rb	0.0003	0.0022	0.1	97.2%	0.0
S	1.7894	0.0117	152.6	0.0%	19.4
Sb	0.0059	0.0251	0.2	93.9%	0.0
Se	0.0056	0.0028	2.0	34.8%	3.6
Si	0.0714	0.0143	5.0	17.8%	6.7
Sn	0.0059	0.0198	0.3	90.7%	0.1
Sr	0.0011	0.0026	0.4	81.8%	0.7
Ti	0.0051	0.0044	1.2	53.8%	2.5
V	0.0015	0.0028	0.5	83.0%	1.0
W	0.0020	0.0116	0.2	95.1%	0.0
Zn	0.0303	0.0023	13.4	0.4%	13.8
Zr	0.0008	0.0039	0.2	91.1%	0.0

[a] *Signal-to-noise ratio was computed as the ratio of the 24-hr ambient air concentration observed at the Lawrenceville site to the corresponding measurement uncertainty reported for that observation in AQS.*

Table 15: Comparison of average MDLs reported by PAQS for the determination of elements in $PM_{2.5}$ samples by ICP-MS at the Schenley Park site with average MDLs reported by AQS for the determination of elements in $PM_{2.5}$ samples by XRF at the Lawrenceville site.

Parameter	Average MDL reported by PAQS for ICP-MS analyses at Schenley Park, 7/01 - 8/02	Average MDL reported by AQS for XRF analyses at Lawrenceville, 7/03 - 12/05	Average ICP-MS MDL ÷ Average XRF MDL
Ag	0.00005	0.01028	0.005
As	0.00009	0.00234	0.036
Ba	0.00078	0.03500	0.022
Ca	0.08031	0.00639	12.572
Cd	0.00008	0.01151	0.007
Ce	0.00034	0.05236	0.007
Co	0.00010	0.00160	0.062
Cr	0.00052	0.00210	0.247
Cs	0.00013	0.02563	0.005
Cu	0.00227	0.00223	1.017
Fe	0.06160	0.00230	26.771
Ga	0.00016	0.00397	0.041
K	0.01973	0.00866	2.279
Mg	0.00710	0.02483	0.286
Mn	0.00035	0.00218	0.159
Mo	0.00010	0.00783	0.012
Ni	0.00030	0.00167	0.182
Pb	0.00028	0.00505	0.056
Rb	0.00009	0.00216	0.041
Sb	0.00006	0.02508	0.003
Se	0.00012	0.00278	0.044
Sr	0.00021	0.00255	0.080
Ti	0.00100	0.00436	0.229
V	0.00012	0.00282	0.042
Zn	0.00577	0.00227	2.547

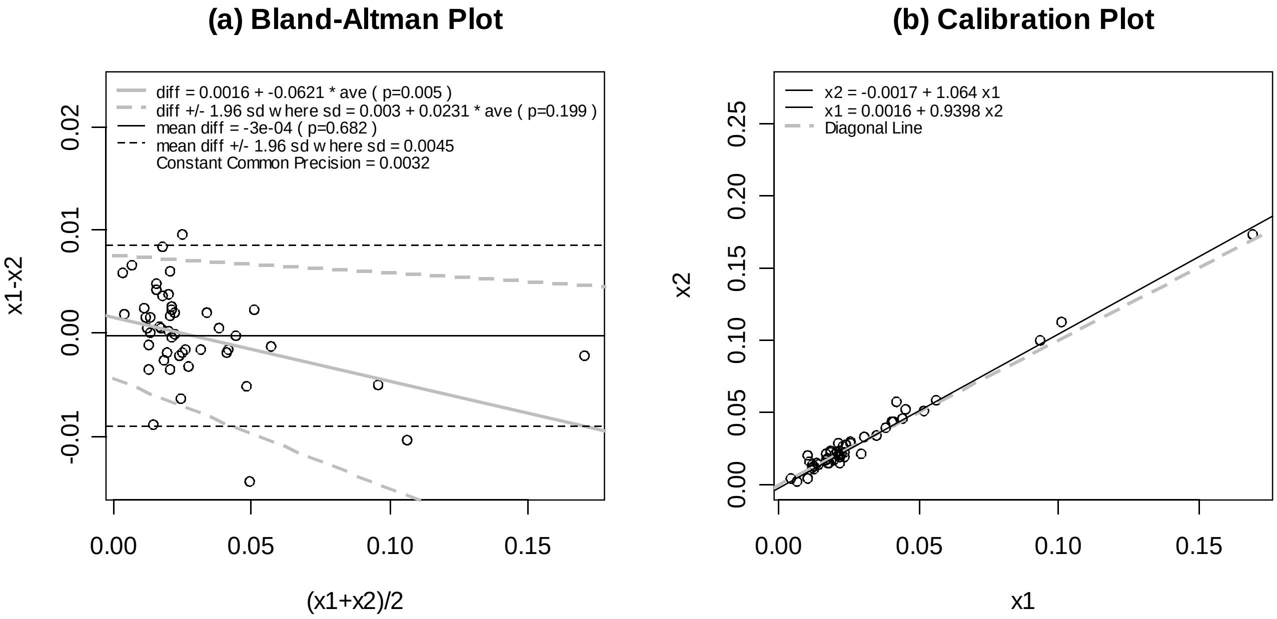


Figure 33: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily Zn concentrations measured at the Lawrenceville site using a Desert Research Institute Sequential Filter Sampler with XRF analysis (x1) and a Met One SASS $PM_{2.5}$ speciation sampler with XRF analysis (x2) between 10/1/02 and 2/27/03.

Figures 33, 34. 35. and 36 present the results of Bland-Altman analyses comparing collocated measurements of Zn, Se, Ni, and V, respectively, determined via XRF from Teflon-filter-based $PM_{2.5}$ samples collected by the UORVP and ACHD monitoring programs at the Lawrenceville site. As evidenced in the figures, the agreement between collocated measurements was best for Zn, which as shown in **Table 14** typically had concentrations at the Lawrenceville site that were about 10 times greater than the XRF MDL, and poorest for V, which typically had concentrations that were only about half as great as the XRF MDL. Constant common imprecision estimates were 11% for Zn, 23% for Se, 51% for Ni, and 120% for V. These results are consistent with the conclusions drawn from **Table 14** regarding the insufficient sensitivity of XRF for determining certain elements, and they indicate that whereas concentrations of some elements (e.g., Zn) determined by XRF from $PM_{2.5}$ samples are likely appropriate for representing exposures in a retrospective epidemiology study, concentrations of many other elements (e.g., V) represent little more than random noise. Hence, even if a relationship truly existed between ambient concentrations of V and adverse health effects, this relationship likely would be masked by measurement error in an epidemiology study, preventing its detection. It is also important to recognize that the examples used here to illustrate the limited sensitivity of XRF for certain elements were based on data from the Lawrenceville site, which was located in an urban area where ambient concentrations tend to be high. Sensitivity is an even greater problem for $PM_{2.5}$ elemental data from suburban or rural areas that are characterized by lower ambient concentrations.

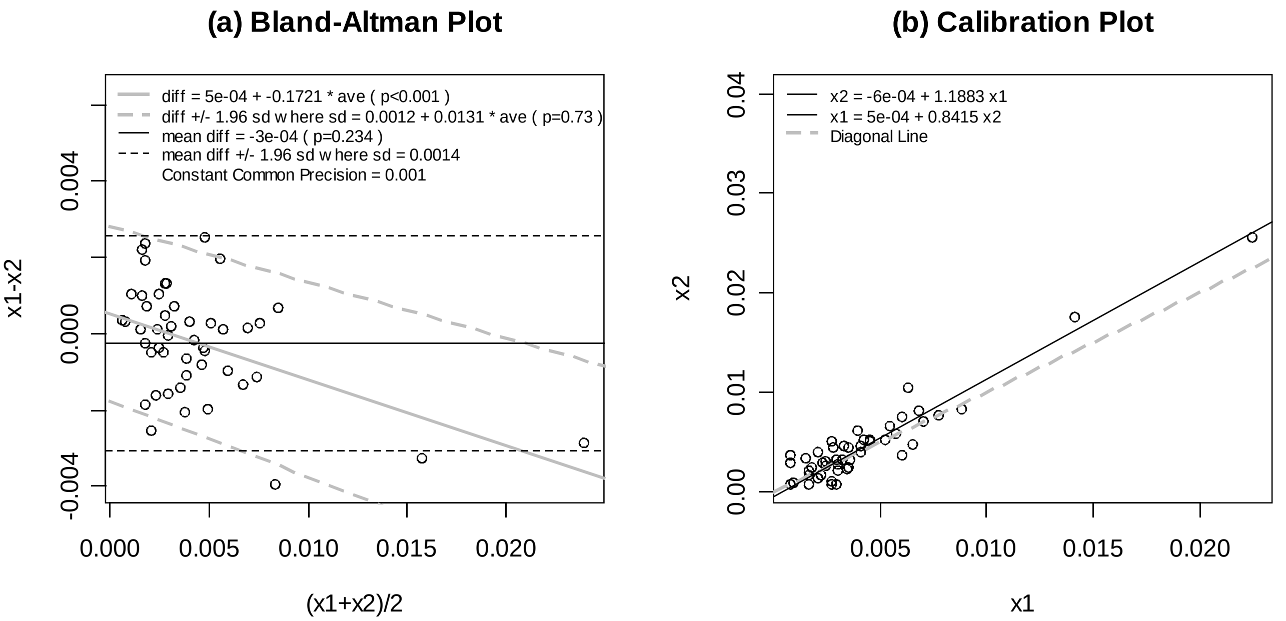


Figure 34: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily Se concentrations measured at the Lawrenceville site using a Desert Research Institute Sequential Filter Sampler with XRF analysis (x1) and a Met One SASS $PM_{2.5}$ speciation sampler with XRF analysis (x2) between 10/1/02 and 2/27/03.

In order to attain better sensitivity than that afforded by XRF, the PAQS and SCAMP programs used ICP-MS for determining the elemental composition of $PM_{2.5}$ samples. $PM_{2.5}$ samples are digested in a strong acid solution to prepare them for analysis by ICP-MS. The digestate, which contains the elemental components of the $PM_{2.5}$ sample, is then nebulized and passed through an argon plasma, which converts the analyte atoms into primarily singly charged ions. These ions are focused by a lens system and fed to a quadrupole mass spectrometer, which filters the ions according to their mass-to-charge ratios (m/z), allowing only those with a specific m/z (corresponding to a particular element) to pass through. A discrete dynode detector is used to count the ions that pass through the quadrupole, allowing the concentration of the element to be determined. **Table 15** compares the average MDLs reported by PAQS for the determination of elements in $PM_{2.5}$ samples by ICP-MS with the average MDLs reported in **Table 14** for the determination of these elements in $PM_{2.5}$ samples by XRF. For many elements, the average MDLs for ICP-MS are one to three orders of magnitude less than the average MDLs for XRF, indicating that ICP-MS has substantially better sensitivity than XRF. ICP-MS is subject to limitations, however. For example, isobaric and polyatomic interferences limit the ability of conventional low-resolution ICP-MS to determine isotopes such as ^{28}Si, ^{39}K, ^{40}Ca, and ^{56}Fe. The effect of these interferences is evident in the relatively large ICP-MS MDLs reported in **Table 15** for Ca, Fe, and K, as well as in the absence of Si from the table. Al and Na also were not able to be reliably determined as part of PAQS (Pekney and Davidson, 2005). The DRC ICP-MS employed by SCAMP is intended to reduce these interferences; use of the DRC produced lower detection limits for Ca, Fe, and K than those reported by PAQS for conventional low-resolution ICP-MS (Connell et al., 2005c). However, the DRC ICP-MS still failed to provide reliable measurements of Si and Na. A second limitation of ICP-MS relative to XRF is that ICP-MS is a destructive method, requiring digestion of the $PM_{2.5}$ sample prior to analysis. Hence, the sample is not preserved for further analysis (e.g., to determine concentrations of inorganic ions by IC, another destructive method) as it is with XRF. This limits the applicability of ICP-MS for analyzing archived $PM_{2.5}$ samples as will be required for the proposed epidemiology study, because it will be necessary to extract as much $PM_{2.5}$ speciation information as possible from each of these samples.

Even if ICP-MS is not employed to analyze archived samples, however, existing elemental data determined by ICP-MS at the Schenley Park site are a source of daily exposure information for use in the study. Collocated daily elemental data are not available from the Schenley Park site; hence, **Figure 37** presents the results of a Bland-Altman analysis comparing Zn concentrations determined by ICP-MS at this site with Zn concentrations determined by XRF at the Lawrenceville site. Whereas collocated Zn concentrations determined at the Lawrenceville site were relatively precise, as shown in **Figure 33**, the comparison of concentrations determined at Schenley with concentrations determined at Lawrenceville was characterized by a large amount of random error. The constant common imprecision estimated for this comparison was 85%. This error likely includes measurement error resulting from differences in the samplers and analytical techniques used to determine concentrations at the two sites, as well as error resulting from the spatial variability of ambient Zn concentrations in the Pittsburgh region.

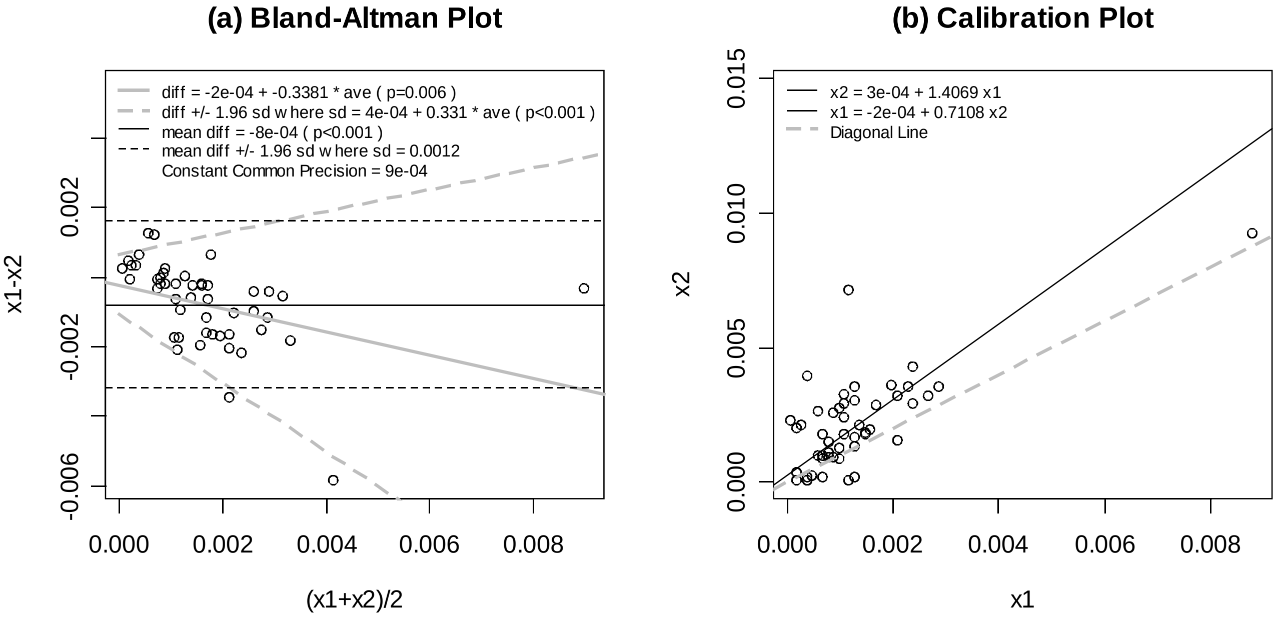


Figure 35: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily Ni concentrations measured at the Lawrenceville site using a Desert Research Institute Sequential Filter Sampler with XRF analysis (x1) and a Met One SASS $PM_{2.5}$ speciation sampler with XRF analysis (x2) between 10/1/02 and 2/27/03.

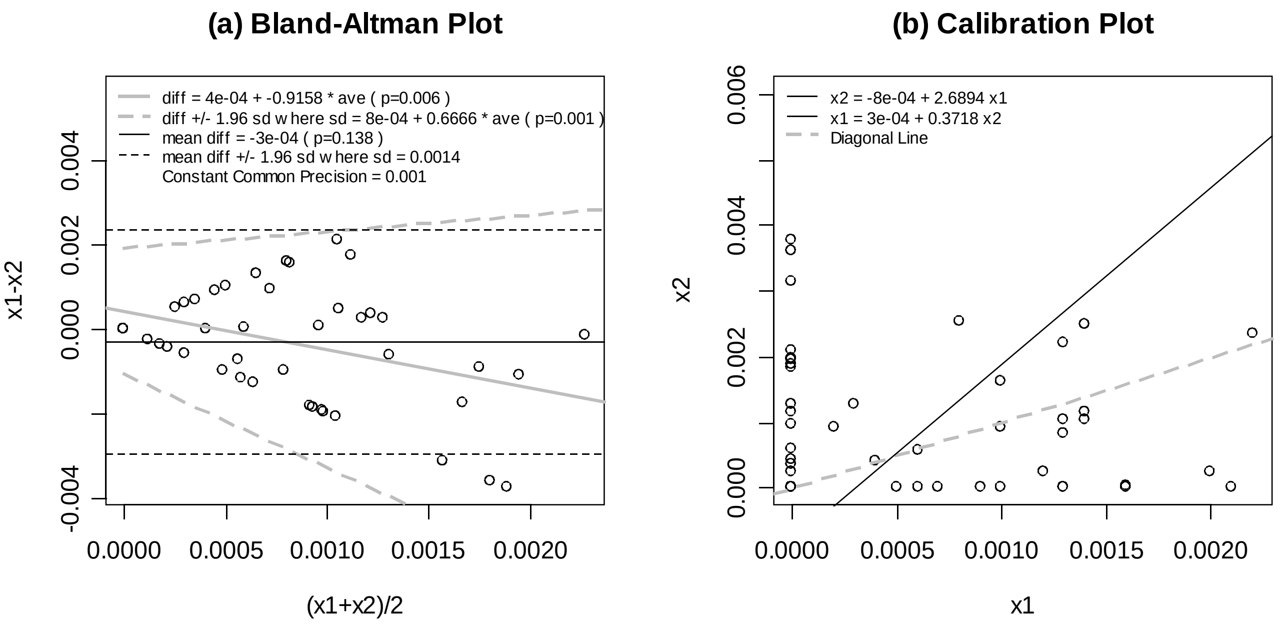


Figure 36: Bland-Altman plot (a) and corresponding calibration plot (b) for collocated daily V concentrations measured at the Lawrenceville site using a Desert Research Institute Sequential Filter Sampler with XRF analysis (x1) and a Met One SASS $PM_{2.5}$ speciation sampler with XRF analysis (x2) between 10/1/02 and 2/27/03.

This spatial variability introduces substantial uncertainty in exposure estimates developed using elemental data from only a single site or a limited number of sites. **Figure 38** presents Bland-Altman results comparing daily Zn concentrations measured at the Lawrenceville site with daily Zn concentrations measured at the Florence site, which was located about 39 km away. At both sites, Zn was determined by XRF from $PM_{2.5}$ samples collected on Teflon filters using Met One SASS speciation samplers. However, as illustrated in **Figure 38**, in spite of the sites' use of identical sampling and analytical methods, pairwise daily Zn concentrations measured at the sites showed poor agreement. The common imprecision estimated from the comparison of these concentrations was 73%. This result, when compared with the common imprecision of only 11% estimated above for collocated determinations of Zn by XRF at the Lawrenceville site, emphasizes the possibility for exposure misclassification resulting from spatially variable concentrations of elements such as Zn in the Pittsburgh region. Such exposure misclassification could mask any associations between these elements and health effects in a time-series epidemiology study.

2.4.2 Effect of Sample Archiving on $PM_{2.5}$ Speciation Measurements

As discussed earlier in **Section 2.3**, the $PM_{2.5}$ speciation data record for the Pittsburgh region during the 1999-2005 time period could be substantially augmented by determining the chemical composition of archived $PM_{2.5}$ samples that were collected in the region during that period. In most cases, only a single Teflon-filter-based sample is available from a given site on a given day; these samples likely would be analyzed first by a nondestructive method such as XRF or PIXE to determine concentrations of trace and crustal elements and then by IC to determine concentrations of inorganic ions. However, in addition to the sampling and analytical errors discussed above in **Section 2.4.1**, concentrations of elements and ions determined from archived $PM_{2.5}$ samples could be affected by errors resulting from contamination or from losses of semi-volatile species during storage. Such errors must be quantified prior to specifying speciation data from archived $PM_{2.5}$ samples for use in the epidemiology study.

The NETL/OST Bruceton monitoring site, which is located in Allegheny County, collected Teflon-filter-based $PM_{2.5}$ samples on an approximately daily basis between July 1999 and June 2004, and has stored these samples under refrigeration since their collection, is an important source of archived $PM_{2.5}$ samples for use in the proposed study. As part of the current feasibility assessment, we conducted a small study to assess whether $PM_{2.5}$ component concentrations determined by analyzing these archived filters, which have been stored for two to seven years, accurately reflect concentrations that would have been determined if the samples had been analyzed shortly after collection. There were several days at the Bruceton site on which collocated Teflon-filter-based $PM_{2.5}$ samples were collected, one of which was analyzed for inorganic ions, and the other of which was archived. With permission from DOE, CONSOL retrieved fifteen of these archived collocated samples, which were collected during 2000 and 2001, and analyzed them for SO_4^{2-}, NO_3^-, and NH_4^+ by ion chromatography. Results were compared with those obtained several years earlier. These comparisons are summarized in the Bland-Altman plots and corresponding calibration plots presented in **Figures 39, 40, and 41** Results for SO_4^{2-} and even for NO_3^-

and NH_4^+, which are more volatile $PM_{2.5}$ components, obtained in 2006 after several years of storage agreed remarkably well with those obtained in 2000 and 2001. The standard deviation of the differences between the CONSOL and NETL measurements was 0.43 μg/m^3 for SO_4^{2-}, 0.18 μg/m^3 for NO_3^-, and 0.34 μg/m^3 for NH_4^+. Assuming equal measurement error for both methods, this translates into measurement imprecisions of about 0.30 μg/m^3 (6.6%) for SO_4^{2-}, 0.13 μg/m^3 (17%) for NO_3^-, and 0.24 μg/m^3 (13%) for NH_4^+. The bias (CONSOL-NETL) for SO_4^{2-} was 0.38 μg/m^3 and for NO_3^- was -0.10 μg/m^3, although the bias for NO_3^- was not statistically significant at a significance level, α, of 0.05. NH_4^+ exhibited an appreciable non-constant bias, such that when NH_4^+ concentrations were about 1 μg/m^3 (average of concentrations determined by CONSOL and NETL), the concentration measured by NETL in 2000 or 2001 was 0.72 μg/m^3 greater than that determined by CONSOL in 2006. For each additional 1 μg/m^3 increase in NH_4^+ concentration, the relative bias decreased by 0.36 μg/m^3, reaching zero at about 3 μg/m^3. Above 3 μg/m^3, NH_4^+ concentrations determined by CONSOL in 2006 were greater than those determined by NETL in 2000 or 2001. Collectively, these results indicate that analysis of refrigerated archived $PM_{2.5}$ samples to determine concentrations of SO_4^{2-}, NO_3^-, and NH_4^+ is feasible, provided that the results are corrected to account for relative biases.

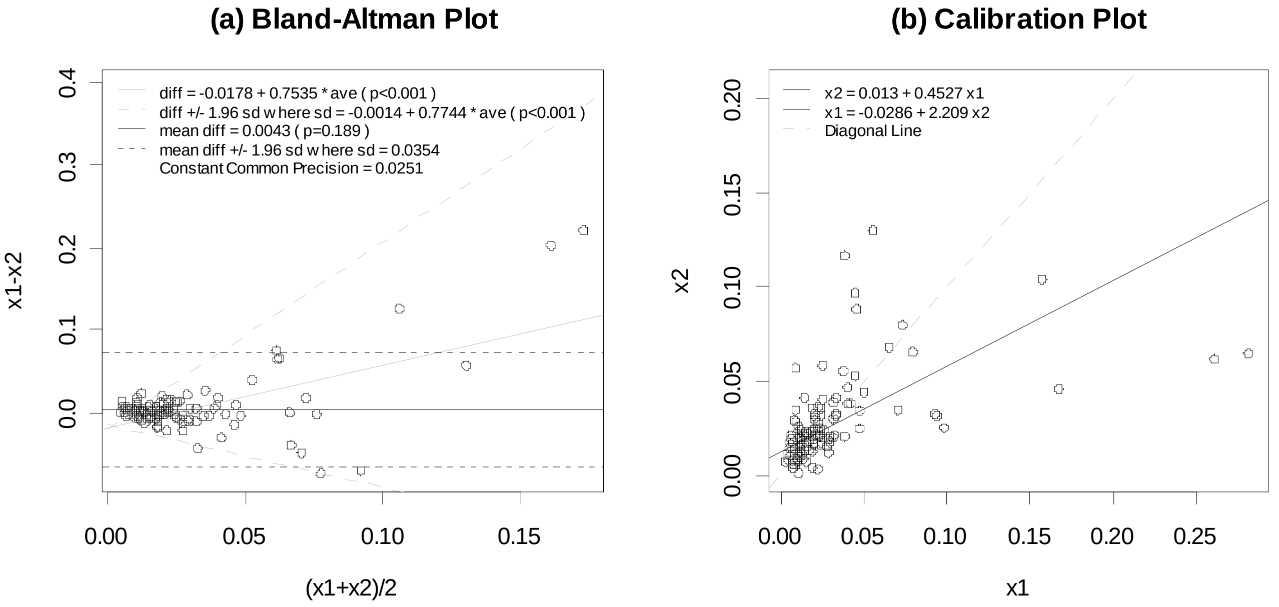


Figure 37: Bland-Altman plot (a) and corresponding calibration plot (b) for pairwise daily Zn concentrations measured at the Schenley Park site using a $PM_{2.5}$ Hi-Vol Sampler with ICP-MS analaysis (x1) and at the Lawrenceville site using a Met One SASS $PM_{2.5}$ speciation sampler with XRF analysis (x2) between 6/30/01 and 7/31/02.

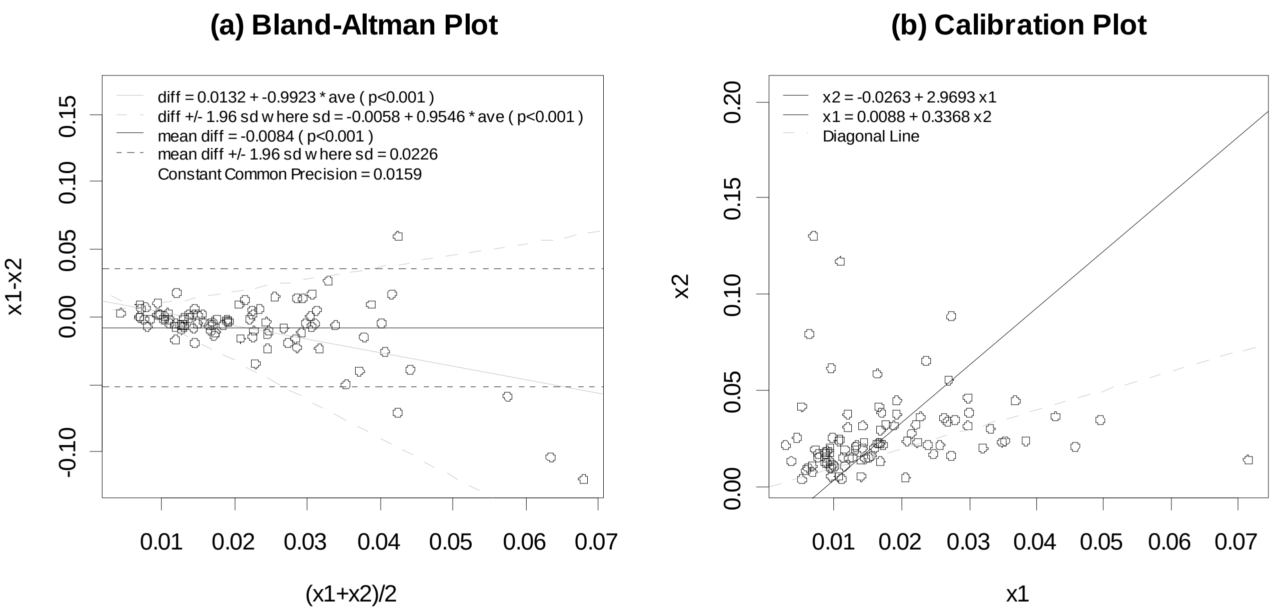


Figure 38: Bland-Altman plot (a) and corresponding calibration plot (b) for pairwise daily Zn concentrations measured at the Florence site (x1) and at the Lawrenceville site (x2) between 6/30/01 and 7/31/02 using Met One SASS $PM_{2.5}$ speciation samplers with XRF analysis.

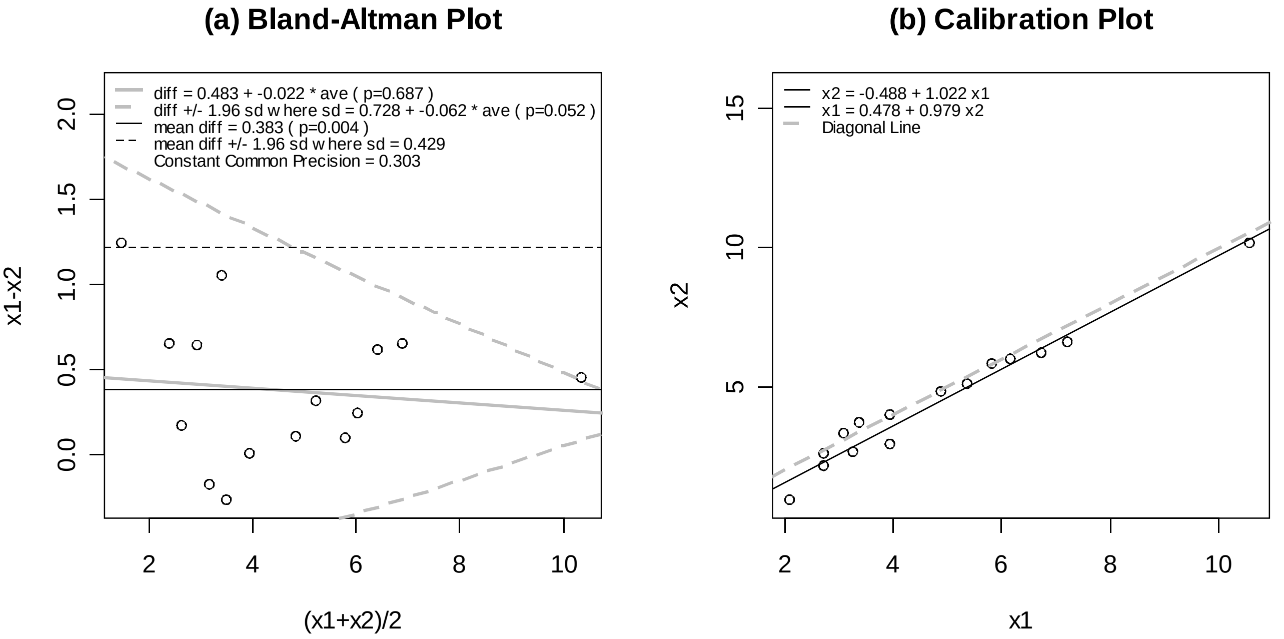


Figure 39: Bland-Altman plot (a) and corresponding calibration plot (b) for daily average sulfate concentrations determined from collocated $PM_{2.5}$ samples collected at the Bruceton site. For each collocated pair, one concentration was determined by NETL in 2000 or 2001 (x2) and the other was determined by CONSOL in 2006 after the sample had been archived under refrigeration for several years (x1).

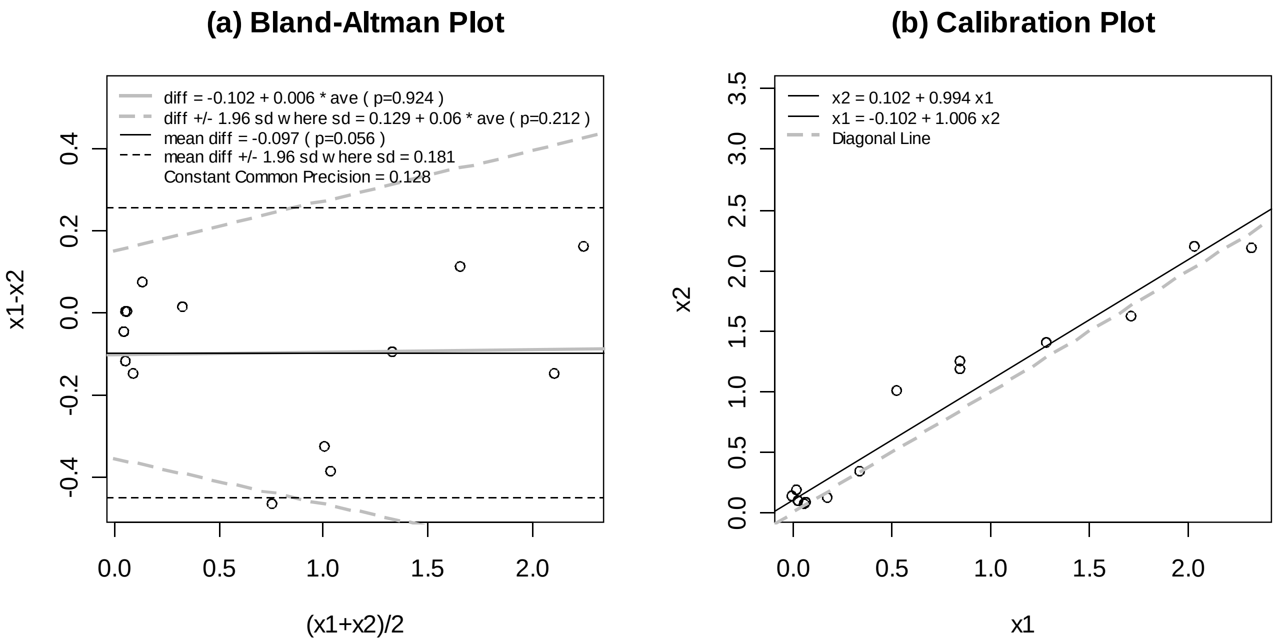


Figure 40: Bland-Altman plot (a) and corresponding calibration plot (b) for daily average nitrate concentrations determined from collocated $PM_{2.5}$ samples collected at the Bruceton site. For each collocated pair, one concentration was determined by NETL in 2000 or 2001 (x2) and the other was determined by CONSOL in 2006 after the sample had been archived under refrigeration for several years (x1).

The results of the analysis of archived $PM_{2.5}$ samples from Bruceton only examined the feasibility of obtaining reliable ion data from samples that have been stored under refrigeration. However, as discussed in **Section 2.3**, many of the archived samples available from the Pittsburgh region during the time period of interest for the proposed epidemiology study have been stored at room temperature, and most of these samples will need to be analyzed for trace and crustal elements as well as ions to provide the speciation data required by the study. Provided that archived $PM_{2.5}$ samples are stored in sealed containers and kept away from possible contamination sources, it is expected that trace and crustal element concentrations determined from these archived samples will accurately reflect ambient concentrations at the time of sample collection, because storage at room temperature is not expected to cause any substantial volatilization of these species. Element concentrations determined as part of the Steubenville Comprehensive Air Monitoring Program are consistent with this expectation. As part of SCAMP, Teflon-filter-based $PM_{2.5}$ samples for elemental analysis were collected in duplicate using separate channels of an Andersen RAAS2.5-400 $PM_{2.5}$ speciation sampler at the Franciscan University of Steubenville site. These samples, which were stored in sealed containers at room temperature prior to analysis, were digested and analyzed in two separate batches, such that all samples from the second speciation sampler channel were analyzed seven to eleven months after the samples from the first speciation sampler channel. In spite of the difference in storage periods (as well as the fact that the results were obtained from two separate samples that were collected on different filters and digested and analyzed independently on different days, all of which are sources of error), results from the second set of filters agreed remarkably well with results from the first set. To exemplify this, **Figures 42 and 43** present Bland-Altman plots for collocated measurements of Fe and collocated measurements of As, respectively, from the Franciscan University site. For both elements, there was little or no bias between the collocated measurements. The mean paired difference between the two sets of results for Fe was not significantly different from zero, and the mean paired difference between the two sets of results for As was just barely significant ($p = 0.049$) at $\alpha = 0.05$. Moreover, the results were characterized by relatively little random error. Constant common imprecision estimates were 19% (0.053 $\mu g/m^3$) for Fe and 21% (0.0003 $\mu g/m^3$) for As, and a substantial portion of this imprecision likely resulted from variability in filter background concentrations, digestion efficiency, instrument calibration, etc. These results suggest that differences in sample storage time likely had little, if any, effect on the elemental results obtained as part of SCAMP.

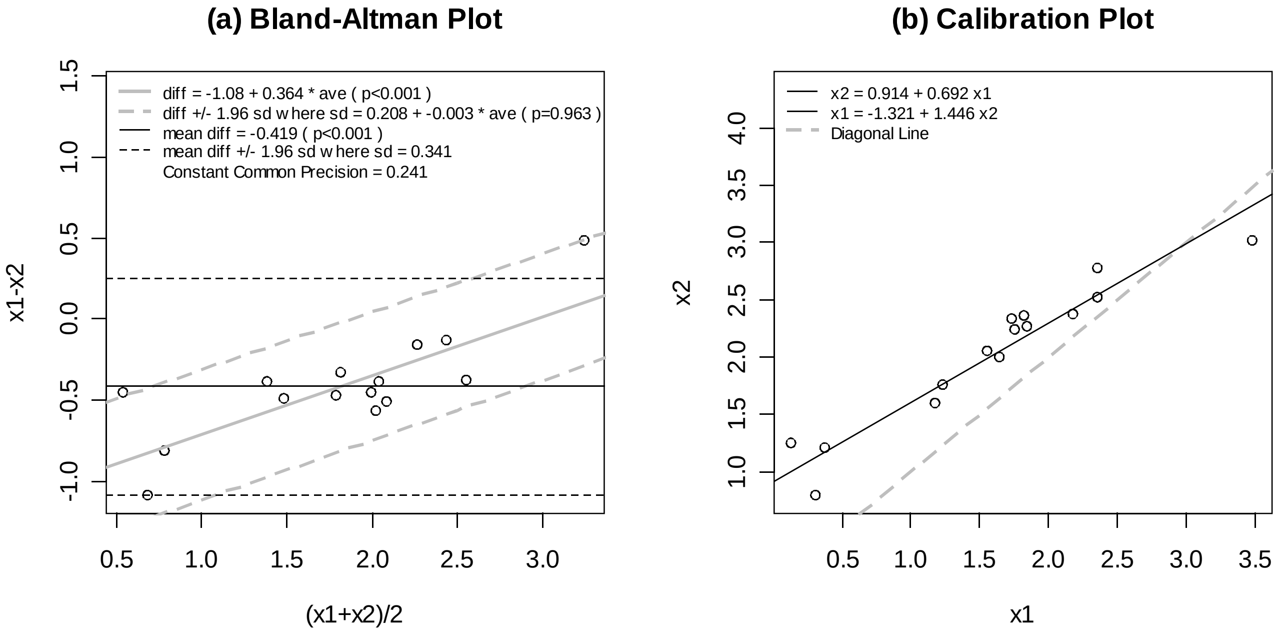


Figure 41: Bland-Altman plot (a) and corresponding calibration plot (b) for daily average ammonium concentrations determined from collocated $PM_{2.5}$ samples collected at the Bruceton site. For each collocated pair, one concentration was determined by NETL in 2000 or 2001 (x2) and the other was determined by CONSOL in 2006 after the sample had been archived under refrigeration for several years (x1).

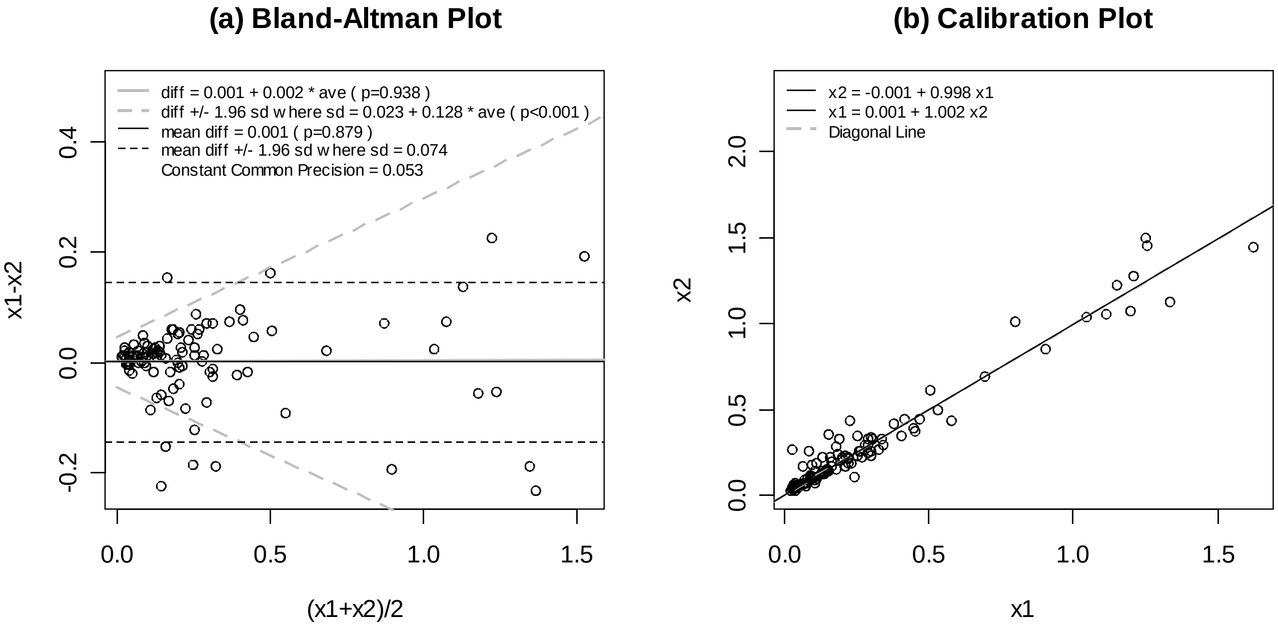


Figure 42: Bland-Altman plot (a) and corresponding calibration plot (b) for daily average Fe concentrations determined by ICP-MS from collocated $PM_{2.5}$ samples collected using separate channels of a speciation sampler at the Franciscan University of Steubenville site. All samples were stored at room temperature prior to analysis; however samples from the second speciation sampler channel (x1) were stored for seven to eleven months longer than samples from the first speciation sampler channel (x2).

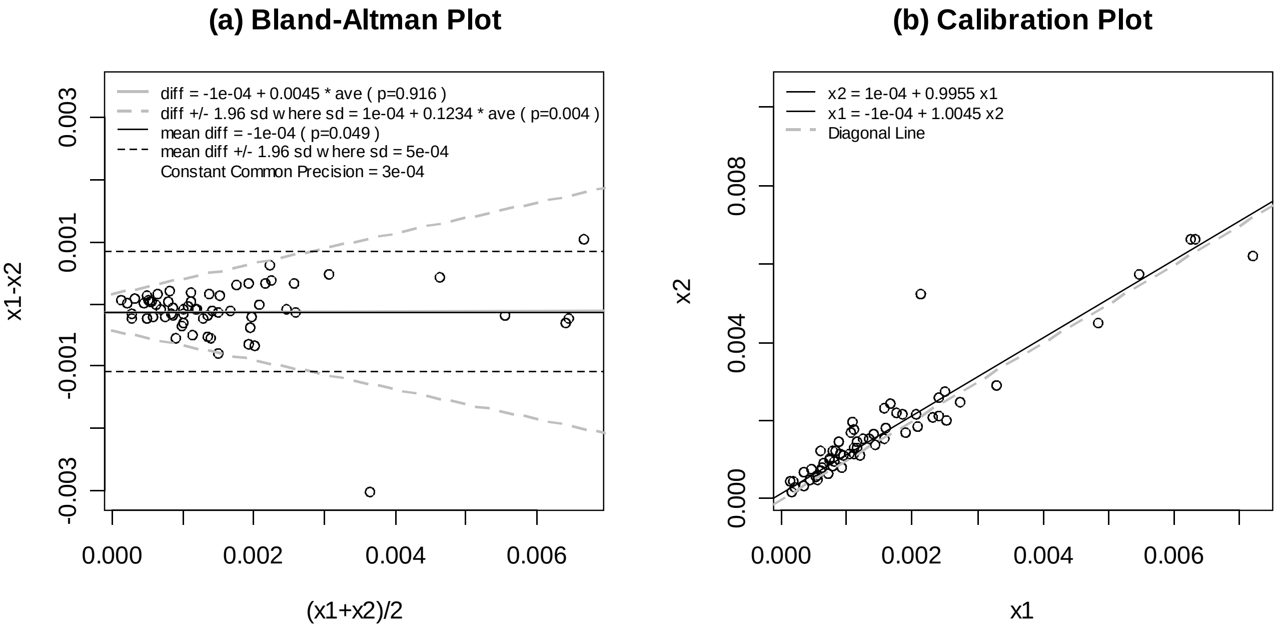


Figure 43: Bland-Altman plot (a) and corresponding calibration plot (b) for daily average As concentrations determined by ICP-MS from collocated $PM_{2.5}$ samples collected using separate channels of a speciation sampler at the Franciscan University of Steubenville site. All samples were stored at room temperature prior to analysis; however samples from the second speciation sampler channel (x1) were stored for seven to eleven months longer than samples from the first speciation sampler channel (x2).

Most of the archived $PM_{2.5}$ samples that would be used to provide elemental data for an epidemiology study have been stored for a longer period of time than the SCAMP samples were. Furthermore, as discussed earlier in **Section 2.3**, these samples are being kept in a variety of locations that are likely characterized by a variety of conditions (e.g., temperature, humidity, possibility for contamination, etc.). Thus, prior to beginning large-scale analysis of archived $PM_{2.5}$ samples to provide chemical speciation data for the proposed study, comparisons such as those presented above for inorganic ions at the Bruceton site should be performed, to as great an extent as possible, for each species at each monitoring site from which archived samples will be analyzed. These comparisons are necessary to establish the validity of the archived sample data and to allow any artifacts resulting from the use of these samples to be corrected.

It is also important to reiterate that, in spite of its superior sensitivity for the determination of many trace elements, ICP-MS likely will not be used to analyze archived $PM_{2.5}$ samples because it requires complete digestion of these samples, rendering them unavailable for subsequent analysis by IC. Rather, XRF is a probable choice for elemental analysis, because it is nondestructive. While XRF generally has poorer sensitivity than ICP-MS, the sensitivity of XRF varies considerably as a function of the details of the technique being used. Kim et al. (2005) compared MDLs for five XRF spectrometers operated by three different laboratories to perform $PM_{2.5}$ elemental analyses; while the averages of these MDLs were generally comparable to the average MDLs reported in **Table 14** for existing data from the Lawrenceville site, the lowest MDLs in many cases were almost an order of magnitude less than the MDLs reported in **Table 14**. For example, whereas average MDLs for determinations of As and Ba at the Lawrenceville site were 0.0023 $\mu g/m^3$ and 0.0350 $\mu g/m^3$, respectively, the lowest MDLs reported by Kim et al. (2005) for XRF determination of these elements were about 0.00052 $\mu g/m^3$ and 0.0023 $\mu g/m^3$, respectively (assuming a sampling flow rate of 6.7 L/min, as used by the Met One SASS sampler at the Lawrenceville site). Hence, use of a more sensitive XRF technique, if affordable, could reduce the error associated with elemental determinations and increase the number of quantifiable elements available for inclusion in the epidemiological models.

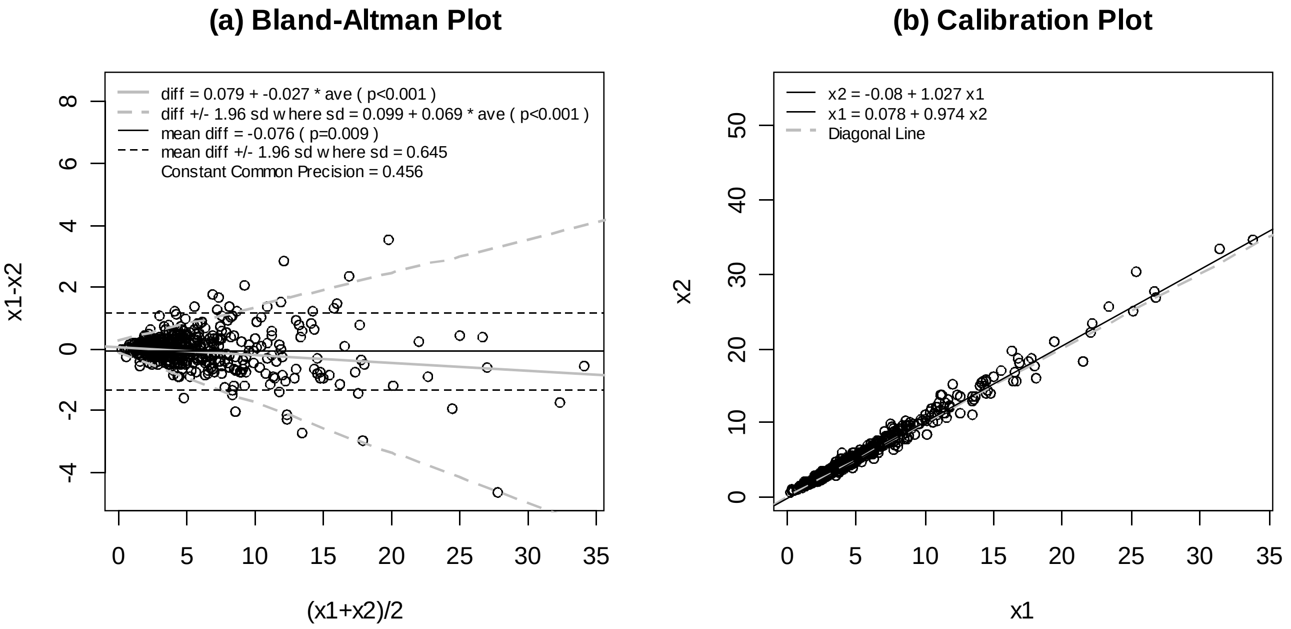


Figure 44: Bland-Altman plot (a) and corresponding calibration plot (b) comparing 24-hr average concentrations of SO_4^{2-} estimated from XRF sulfur measurements (x1) with 24-hr average concentrations of SO_4^{2-} determined by IC at the AQS Lawrenceville site between 6/30/01 and 12/27/05.

In addition to being nondestructive, a strength of XRF is its ability to accurately and precisely determine sulfur, which can be used to estimate concentrations of SO_4^{2-} (under the assumption that all fine particulate S is present as SO_4^{2-}). Hence, for archived $PM_{2.5}$ samples that have not been refrigerated, if preliminary comparisons of archived sample data with preexisting data indicate that determination of NH_4^+ and NO_3^- from these samples is not feasible, then IC analysis may not be required. **Figure 44** presents a Bland-Altman plot comparing SO_4^{2-} concentrations determined by IC with SO_4^{2-} concentrations estimated from XRF sulfur measurements at the Lawrenceville monitoring site. Although the bias between these collocated measurements was statistically significant, it was very small (1.3%). On average, SO_4^{2-} concentrations estimated from the XRF data were about 0.1 μg/m^3 less than SO_4^{2-} concentrations measured by IC. The two methods were also reasonably precise, having a constant common precision of about 0.5 μ g/m^3 (7.8%), which is similar to that reported in **Section 2.4.1.2** for collocated measurements of SO_4^{2-} by IC at the Lawrenceville site. Thus, these results indicate that SO_4^{2-} concentrations estimated from XRF sulfur measurements can be used interchangeably with SO_4^{2-} concentrations determined by IC for purposes of the proposed epidemiology study. Therefore, if SO_4^{2-} is the only inorganic ionic species that needs to be determined from an archived $PM_{2.5}$ sample, its concentration can be estimated from XRF data, saving the cost of IC analysis. In this scenario, following XRF analysis, the sample could be digested for analysis by ICP-MS to determine concentrations of elements that are of interest for inclusion in the epidemiology study but are not readily quantified by XRF (e.g., Al, Ba, Cd, Mg, V, etc.). This would add substantially to the cost of analysis, however.

2.4.3 Comparison of QA/QC Procedures

A review of the quality assurance and quality control procedures followed by the monitoring campaigns that operated in the Pittsburgh region between 1999 and 2005 indicates that, for the most part, these procedures were sufficient to ensure that the data available from these campaigns are of reasonably high quality for use in a retrospective epidemiology study. A detailed description and comparison of these various QA/QC protocols is beyond the scope of this report, as some of the individual protocols are themselves greater than 100 pages in length. However, in general, sampling and analytical activities performed by monitoring sites in the Pittsburgh region were conducted according to standard operating procedures and QA/QC protocols that ensured consistent practices in the field and in the laboratory, and included routine use of logbooks (to document observations, exceptions, maintenance activities, QA/QC items, etc.), sampler audits (to check/calibrate flow, temperature, pressure, etc.), field and trip blanks (to assess sample contamination during handling, transport, and storage), standard reference materials (to confirm the accuracy of analytical techniques), collocated sampling and/or replicate analysis (to assess precision), etc. Moreover, with the exception of data from the NETL/OST Bruceton site, data reported by the various monitoring campaigns have been reduced, screened, and qualified to indicate measurements that are suspect or invalid because of instrument malfunction, abnormal sampling conditions, or noncompliance with QA/QC criteria. Hence, these data require little additional vetting prior to use in the study.

A noteworthy inconsistency among the QA/QC procedures followed by the monitoring campaigns that operated in the Pittsburgh region between 1999 and 2005 is in the flagging procedures used to indicate the validity of reported data. This inconsistency can easily be eliminated, however, by converting all reported flags to the set of standard flags developed by the North American Research Strategy for Tropospheric Ozone (NARSTO). These NARSTO standard flags, which are shown in **Table 16** are general enough to be widely applicable to datasets that were qualified according to more complex flagging procedures, yet provide sufficient detail to indicate whether data should be included or excluded from the epidemiology study. Data labeled as V0, V1, V2, V3, V4, V5, V6, or V7 would be included in the study, and data labeled as M1, M2, or H1 would be excluded from the study (although efforts would be made to assess and validate any data labeled as H1 so that a "V" or "M" flag could be properly assigned to these data). Data estimated via the calibration and geostatistical techniques described elsewhere in this report would be assigned V2 or V3 flags as appropriate. Because of problems related to the use of censored data in statistical models, data labeled as V7 (i.e., censored data) would be evaluated to determine its likely impact on the epidemiological models and to determine whether the uncensored values for observations below the detection limit could be obtained. Finally, statistically influential points flagged as V4, V5, or V6 would be investigated (e.g., by comparison with collocated measurements, review of logbooks and comments provided by data originator, etc.) prior to inclusion in the epidemiological models in order to confirm or refute their validity.

Table 16: NARSTO standard data qualification flags.

Flag	Description
V0	Valid value
V1	Valid value but comprised wholly or partially of below detection limit data
V2	Valid estimated value
V3	Valid interpolated value
V4	Valid value despite failing to meet some QC or statistical criteria
V5	Valid value but qualified because of possible contamination (e.g., pollution source, laboratory contamination source)
V6	Valid value but qualified due to non-standard sampling conditions (e.g., instrument malfunction, sample handling)
V7	Valid value but set equal to the detection limit (DL) because the measured value was below the DL
M1	Missing value because no value is available
M2	Missing value because invalidated by data originator
H1	Historical data that have not been assessed or validated

Data reported by the Pittsburgh Air Quality Study have already been qualified using NARSTO standard flags. Data reported by CASTNet have been qualified using a modified version of these flags. The CASTNet flags include the 11 NARSTO standard flags plus the following five flags:

- I0 – Invalid value, unknown reason

- I1 – Invalid value, known reason
- I2 – Invalid value, -999
- NA – Not available from source data
- M3 – Missing value due to clogged filter

The CASTNet flags could easily be made consistent with the NARSTO standard flags by converting the I0, I1, and M3 flags to M2 flags and by converting the I2 and NA flags to M1 flags. Data collected by the IMPROVE network are qualified using a unique set of IMPROVE validation flags that differ from the NARSTO and CASTNet flags; however, these flags are translated to the same set of flags used by the CASTNet program before the data are reported in the Visibility Information Exchange Web System (VIEWS) database, as shown in **Table 17**. Hence, these flags also could be made consistent with the NARSTO standard flags simply by converting any I1 and M3 flags to M2 flags.

Data from AQS and from the UORVP program are qualified using flagging systems that are appreciably more detailed than the NARSTO flagging system. **Table 18** presents the data validation flags used by AQS, and **Table 19** summarizes the major data validation flags used by UORVP. (In addition to the major flags shown in **Table 19**, the UORVP system included sub-flags to provide additional details, such that data could be qualified according to any combination of more than 120 different sampling and analytical flags). Although it will not be as straightforward as for the CASTNet and IMPROVE flags, the AQS and UORVP flags can be converted to NARSTO standard flags rather easily. In general, all AQS data flagged with a null data qualifier would be assigned an M1 or M2 flag, depending on the specific qualifier code, and all other data would be assigned a V0, V1, V4, V5, or V6 flag, depending on whether these data were above or below the MDL and on the specific qualifiers (if any) used to describe the data. The UORVP program flags any suspect data with an "S" or "s" flag and any invalid data with a "V" or "v" flag. (Upper-case flags are used to qualify field data, and lower-case flags are used to qualify laboratory data). Hence, in general, any UORVP data marked with a "V" or "v" flag would be assigned an M1 or M2 flag; any data marked with an "S" or "s" flag would be qualified with a "V4", "V5", or "V6" flag, and all other data would be qualified with one of the seven NARSTO "V" flags, again depending on whether these data were above or below the MDL and on the specific qualifiers (if any) used to describe the data.

Table 17: IMPROVE native data qualification flags and corresponding VIEWS flags.[a]

IMPROVE Native Flag	Description	VIEWS Flag
AA	Organic artifact corrected.	V5
AP	Possible organic artifact.	V5
BI	Incorrect installation of sample cartridge during weekly change.	M2

IMPROVE Native Flag	Description	VIEWS Flag
CG	Clogging filter - flow rate less than 18 L/min for more than 1 hour. This affects the cut point of the particle but the concentrations are correct.	V2
CL	Clogged filter - flow rate less than 15 L/min for more than 1 hour.	M3
DE	Derived or calculated value.	V0
EP	Equipment problem.	M1
LF	Low/high flow rate. The average flow rate results in a cyclone cut point outside of the 2-3 micro-m range. This corresponds to flow rates < 21.3 L/min or > 24.3 L/min.	V5
MV	Missing module level value.	M1
NA	Not applicable. This is used for missing modules with non-protocol samplers with less than four modules.	M1
NM	Normal.	V0
NR	Not reprocessed. Carbon data between 2000 – 2004 which were not reprocessed to account for negative OP that had originally been reported as zero.	V0
NS	Operator did not install the samples or installed them too late to acquire a valid time. All filters involved.	M1
OL	Offline. In some cases, this is used when the sampler is inoperable due to hurricane or fire. For year 2000, this is used for the period after the Version 1 sampler is removed and before the Version 2 samples begins operation.	M1
PO	Power outage. All filters involved.	M1
QA	QA problem suspected, value held back for further investigation.	I1
QD	Questionable data.	V4
RF	High flow rate. The flow rate is greater than 27 L/min for more than 1 hour. This affects the cut point of the particle but the concentrations are correct.	V5
SA	Used to indicate some sort of 'Sampling Anomaly', though the exact definition has yet to be satisfactorily defined.	V6
SP	An artifact filter was swapped with a sample filter.	V0
SW	Suspected filter swap.	V0
UN	The concentrations failed the data validation for unknown reasons.	M2
XX	The filter is damaged.	M2

[a] *Source: http://vista.cira.colostate.edu/views/Web/Documents/Dataflags.aspx*

Table 18: AQS data qualification flags.[a]

Qualifier Code	Qualifier Description
EX - Exceptional Event Qualifier	
D	SANDBLASTING
F	STRUCTURAL FIRE
H	CHEMICAL SPILLS & INDUST. ACCIDENTS
I	UNUSUAL TRAFFIC CONGESTION
J	CONSTRUCTION/DEMOLITION

Qualifier Code	Qualifier Description
K	AGRICULTURAL TILLING
L	HIGHWAY CONSTRUCTION
M	REROUTING OF TRAFFIC
N	SANDING/SALTING OF STREETS
O	INFREQUENT LARGE GATHERINGS
P	ROOFING OPERATIONS
Q	PRESCRIBED BURNING
R	CLEAN UP AFTER A MAJOR DISASTER
NAT - Natural Event Qualifier	
A	HIGH WINDS
B	STRATOSPHERIC OZONE INTRUSION
C	VOLCANIC ERUPTIONS
E	FOREST FIRE
G	HIGH POLLEN COUNT
S	SEISMIC ACTIVITY
U	SAHARA DUST
NULL - Null Data Qualifier	
AA	SAMPLE PRESSURE OUT OF LIMITS
AB	TECHNICIAN UNAVAILABLE
AC	CONSTRUCTION/REPAIRS IN AREA
AD	SHELTER STORM DAMAGE
AE	SHELTER TEMPERATURE OUTSIDE LIMITS
AF	SCHEDULED BUT NOT COLLECTED
AG	SAMPLE TIME OUT OF LIMITS
AH	SAMPLE FLOW RATE OUT OF LIMITS
AI	INSUFFICIENT DATA (CANNOT CALCULATE)
AJ	FILTER DAMAGE
AK	FILTER LEAK
AL	VOIDED BY OPERATOR
AM	MISCELLANEOUS VOID
AN	MACHINE MALFUNCTION
AO	BAD WEATHER
AP	VANDALISM
AQ	COLLECTION ERROR
AR	LAB ERROR
AS	POOR QUALITY ASSURANCE RESULTS
AT	CALIBRATION
AU	MONITORING WAIVED
AV	POWER FAILURE (POWR)
AW	WILDLIFE DAMAGE

Qualifier Code	Qualifier Description
AX	PRECISION CHECK (PREC)
AY	Q C CONTROL POINTS (ZERO/SPAN)
AZ	Q C AUDIT (AUDT)
BA	MAINTENANCE/ROUTINE REPAIRS
BB	UNABLE TO REACH SITE
BC	MULTI-POINT CALIBRATION
BD	AUTO CALIBRATION
BE	BUILDING/SITE REPAIR
BF	PRECISION/ZERO/SPAN
BG	MISSING OZONE DATA NOT LIKELY TO EXCEED LEVEL OF STANDARD
BH	INTERFERENCE/CO-ELUTION
BI	LOST OR DAMAGED IN TRANSIT
BJ	OPERATOR ERROR
BK	SITE COMPUTER/DATA LOGGER DOWN
QA - Quality Assurance Qualifier	
1	DEVIATION FROM A CFR/CRITICAL CRITERIA REQUIREMENT
2	OPERATIONAL DEVIATION
3	FIELD ISSUE
4	LAB ISSUE
5	OUTLIER
6	QAPP ISSUE
7	BELOW LOWEST CALIBRATION LEVEL
9	NEGATIVE VALUE DETECTED – ZERO REPORTED
V	VALIDATED VALUE
W	FLOW RATE AVERAGE OUT OF SPEC.
X	FILTER TEMPERATURE DIFFERENCE OUT OF SPEC.
Y	ELAPSED SAMPLE TIME OUT OF SPEC.

[a]Source: http://www.epa.gov/ttn/airs/airsaqs/manuals/qualifiers.htm

Data reported by the Steubenville Comprehensive Air Monitoring Program are qualified according to a relatively simple data flagging procedure that classifies all data as valid (“V”), flagged (“F”), or invalid (“I”). For flagged and invalid data, comments are also provided to explain why the “F” or “I” qualifier was applied. Hence, NARSTO standard flags could be applied to the SCAMP data based on the SCAMP data validation flags and associated comments. All data labeled “V” would be assigned NARSTO flags of V0, V1, or V7, depending on whether the data were above or below the MDL and on whether data that were below the MDL were censored. (Ionic and carbonaceous data from SCAMP were censored;

elemental data were not). Data labeled "F" would be assigned NARSTO flags of V4, V5, or V6, and data labeled "I" would be assigned NARSTO flags of M1 or M2, based on the comments accompanying the flag.

Table 19: UORVP data qualification flags.

Qualifier Code	Qualifier Description
Ambient and Source Field Sampling Data Validation Flags	
A	Sampler adjustment or maintenance.
B	Field blank.
D	Sample dropped.
F	Filter damaged or ripped.
G	Filter deposit damaged.
H	Filter holder assembly problem.
I	Inhomogeneous sample deposit.
L	Sample loading error.
M	Sampler malfunction.
N	Foreign substance on sample.
O	Sampler operation error.
P	Power failure during sampling.
Q	Flow rate error.
R	Replacement filter used.
S	Sample validity is suspect.
T	Sampling time error.
U	Unusual local particulate sources during sample period.
V	Invalid sample.
W	Wet sample.
X	No sample was taken this period, sample run was skipped.
Chemical Analysis Data Validation Flags	
b	Blank.
c	Analysis result reprocessed or recalculated.
d	Sample dropped.
f	Filter damaged or ripped.
g	Filter deposit damaged.
h	Filter holder assembly problem.
i	Inhomogeneous sample deposit.
m	Analysis results affected by matrix effect.
n	Foreign substance on sample.
q	Standard.
r	Replicate analysis.
s	Suspect analysis result.
v	Invalid analysis result.
w	Wet sample.

Only some of the data collected by NETL/OST at the Bruceton monitoring site have been reduced and screened for quality. Certain data from the Bruceton site are available only in the raw format in which they were recorded by the data logger in the field or by the analytical laboratory, and others have been reduced to ambient air concentration units and undergone preliminary screening for quality, but have not been qualified using a standard set of data validation flags. Hence, prior to using data from the Bruceton monitoring site in an epidemiology study, these data must be reduced, vetted, and qualified according to a consistent procedure. Unlike data from the other monitoring sites in the Pittsburgh region, in which NARSTO standard flags can be assigned relatively simply on the basis of preexisting data validation flags, these NARSTO flags will need to be assigned to the Bruceton data manually based upon a review of sampling logbooks; sampling temperature, pressure, and flow rate data; instrument error codes; laboratory records, etc. As part of the current feasibility assessment, records from the Bruceton site were gathered and reviewed, and it was determined that all of the information needed to validate these data is available. While this process will be time-consuming, it is necessary to ensure that the quality of the dataset constructed for the Bruceton site is consistent with the quality of the other data being used in the epidemiology study.

2.5 Plan for the Construction of an Air Monitoring Database for the Retrospective Epidemiology Study

Having provided an inventory of existing air monitoring data and archived $PM_{2.5}$ samples that were collected in the greater Pittsburgh region between 1999 and 2005 and would be available for use in a retrospective epidemiology study of $PM_{2.5}$, as well as an evaluation of the quality and comparability of these data and samples, we now present a plan for utilizing the existing data and samples to assemble a database of exposure information suitable for the proposed retrospective study. Development of this plan essentially amounted to solving a constrained optimization problem; the plan had to minimize cost while achieving a quantity and quality of air monitoring data sufficient to enable the performance of a retrospective epidemiology study of $PM_{2.5}$ and its components in the Pittsburgh region.

Based upon the air monitoring data inventory results presented in **Section 2.2** and the health data inventory results presented later in this report, $PM_{2.5}$ speciation data impose the largest constraint on the design of the proposed retrospective epidemiology study. Because of their limited availability, these data largely dictate the time period and geographic region that will form the basis for the study. **Table 20** summarizes the availability of existing $PM_{2.5}$ speciation data by calendar year for each of three geographic regions: the 35-county greater Pittsburgh region that was defined earlier in **Table 1**, the seven-county Pittsburgh Metropolitan Statistical Area (MSA) (which includes Allegheny, Armstrong, Beaver, Butler, Fayette, Washington, and Westmoreland Counties in southwestern Pennsylvania), and the 28 counties from the 35-county region that are not part of the Pittsburgh MSA. As shown in **Table 20**, the record of $PM_{2.5}$ speciation data available from the 28 counties outside of the Pittsburgh MSA is sparse relative to the record available from the seven counties in the Pittsburgh MSA. There are 1372, 1392, and 1537 days between 1999 and 2005 on which sulfate, nitrate, and EC/OC data, respectively, are available from at least one monitoring site in the Pittsburgh MSA, compared with only 766, 766, and 743 days on which these data are available from at least one site outside of the Pittsburgh MSA. Whereas data are available from at least one site in the Pittsburgh MSA on greater than 80% of the days in 2000 and 2001 for EC/OC, on greater than 80% of the days in 2002 for all $PM_{2.5}$ species, and on greater than 80% of the days in 2003 for sulfate and nitrate, in no case are data for any $PM_{2.5}$ species available from the sites located outside of the Pittsburgh MSA on more than 50% of the days in a given calendar year. Moreover, as shown earlier in **Table 10** with the exception of the Holbrook site in Greene County, all of the archived $PM_{2.5}$ samples available from western Pennsylvania were collected at monitoring sites located in the Pittsburgh MSA. Hence, the available $PM_{2.5}$ speciation data suggest that a retrospective epidemiology study focusing on the chemical components of $PM_{2.5}$ should be limited to the Pittsburgh MSA or a smaller region, as there are not sufficient data to allow the development of reliable daily exposure estimates for the rest of the 35-county region. (Selection of a smaller region with a greater density of geographically diverse monitoring sites is expected to reduce the possibility for effect attenuation resulting from exposure misclassification in the time series epidemiology study).

Therefore, the analyses presented in the remainder of this section focus only on data that were collected in the Pittsburgh MSA. (It should be noted, however, that any data available from outside of the Pittsburgh MSA would be appropriate for use in geostatistical modeling to inform the computation of spatially averaged exposure estimates, as discussed later in this report). The monitoring sites in the Pittsburgh MSA that collected $PM_{2.5}$ speciation data between 1999 and 2005 were the Bruceton, Hazelwood, Lawrenceville, Liberty Borough, and Schenley Park sites in Allegheny County, the Florence site in Washington County, and the Greensburg and St. Vincent College sites in Westmoreland County. The population of the MSA, based on the 2000 census, was 2,431,087, which represented more than half of the population of the larger 35-county region.

Table 20: Summary of $PM_{2.5}$ speciation data availability between 1999 and 2005 by species, year, and geographic region.[a]

	Number of Days With:				
	Sulfate	**Nitrate**	**EC/OC**	**Elements**	**Complete Speciation**
Any Site in 35-County Region	1459	1478	1579	1160	1128
1999	92	92	164	79	79
2000	199	194	336	153	145
2001	292	290	362	287	280
2002	360	358	361	325	308
2003	300	340	201	162	162
2004	184	172	123	123	123
2005	32	32	32	31	31

	Number of Days With:				
	Sulfate	**Nitrate**	**EC/OC**	**Elements**	**Complete Speciation**
Any Site in Pittsburgh MSA	1372	1392	1537	997	964
1999	50	49	139	34	34
2000	174	169	327	93	87
2001	275	273	361	241	232
2002	360	358	361	323	305
2003	297	339	196	154	154
2004	184	172	121	121	121
2005	32	32	32	31	31
Any Site Outside of Pittsburgh MSA	766	766	743	735	722
1999	69	69	69	69	69
2000	115	115	101	98	95
2001	178	178	171	168	162
2002	144	144	142	141	137
2003	121	121	121	120	120
2004	122	122	122	122	122
2005	17	17	17	17	17

[a] Inventory for 2005 does not include all data collected in that year. At the time of the inventory, data were available for the AQS sites through 4/10/05 and for the IMPROVE sites through 12/29/04.

There are not enough preexisting $PM_{2.5}$ speciation data available from the Pittsburgh MSA between 1999 and 2005 to provide sufficient statistical power for a retrospective epidemiology study of $PM_{2.5}$ chemical components. The power calculations presented later in this report indicate that a minimum of three years (1,095 days) of daily data are needed for a feasible study. More than three years of data are recommended, as the additional data increase the power of the study and hence its ability to detect smaller effects. The data specified for the study should also contain relatively few missing values, as these detract from the power of the study. Hence, our goal when designing an exposure database for the proposed retrospective epidemiology study was to identify a four-year (1,460-day) or longer period during which 24-hour average data for each $PM_{2.5}$ species of interest (i.e., SO_4^{2-}, NO_3^-, EC and OC, and trace and crustal elements) are available for at least 85% of the days (i.e., such that there are less than 15% missing values for each species).

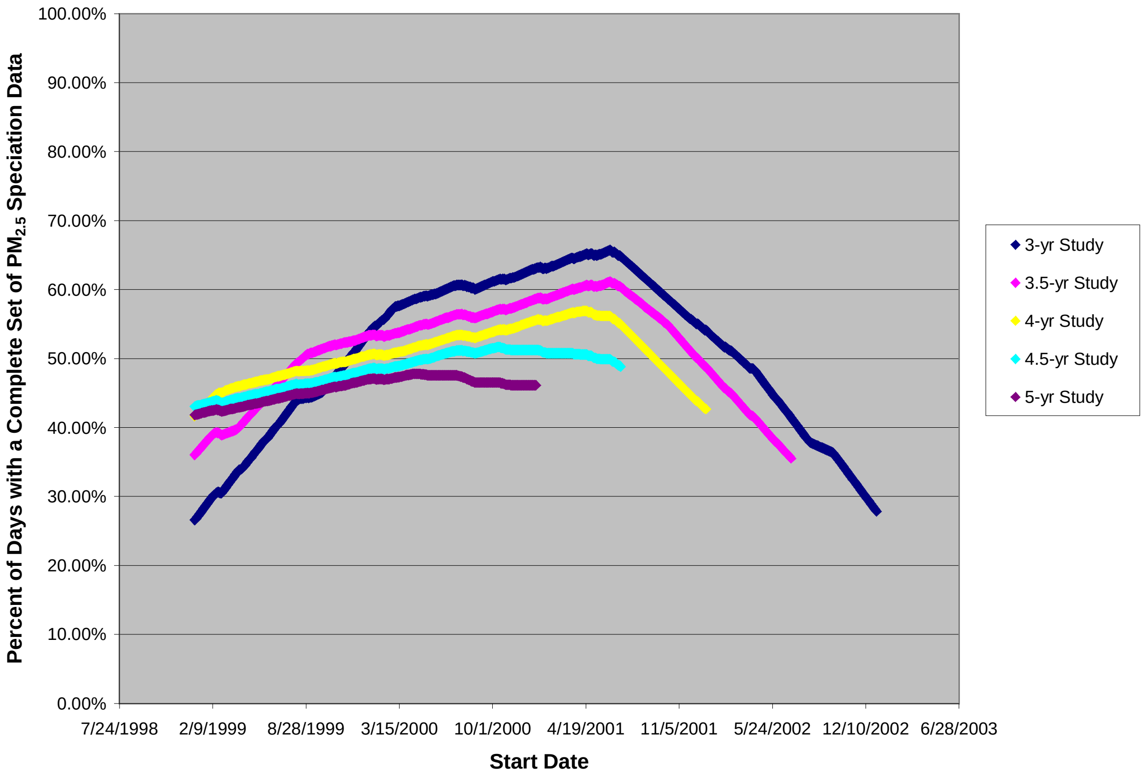


Figure 45: Effect of start date and study length on $PM_{2.5}$ speciation data completeness (i.e., the percentage of days on which sulfate, nitrate, carbon, and elements were each measured at one or more sites in the Pittsburgh MSA, regardless of whether all of the species were measured at the same site) for the case in which only preexisting data, and not data obtained from the analysis of archived $PM_{2.5}$ samples, are used.

It is not possible to satisfy this criterion using only preexisting data from the Pittsburgh MSA. In order to select an optimal study period, we analyzed data completeness (i.e., the percentage of days during the study period for which a complete set of $PM_{2.5}$ speciation data could be obtained by using any combination of existing sulfate, nitrate, carbon, and elemental data from monitoring sites in the Pittsburgh MSA) as a function of study start date and study length. When performing this analysis, we assumed that data from the various monitoring sites (and measurement techniques) in the Pittsburgh MSA could be used interchangeably to represent the exposures of the region's population, such that even if no single monitoring site in the Pittsburgh MSA had a complete set of $PM_{2.5}$ speciation data on a given day, that day would still be considered to have a complete set of speciation data if data from multiple monitoring sites could be combined to provide at least one measurement from somewhere in the region for each $PM_{2.5}$ species of interest. Based on the discussion in **Section 2.4**, the assumption that measurements from different monitoring sites are interchangeable is much more reasonable for $PM_{2.5}$ components such as SO_4^{2-} and OC than it is for trace and crustal element species. Nevertheless, this assumption indicates the maximum data completeness that can possibly be achieved for a given study length. Hence if the criterion for data completeness cannot be satisfied under this assumption, then it will not be satisfied under any other, more restrictive assumption. We also assumed that midnight-to-midnight concentrations would have to be estimated by averaging filter-based data collected at the Bruceton and St. Vincent College sites, as discussed in **Section 2.4**, such that data were only considered to be available from these sites on a given day if data from the prior day were also available.

Results of the analysis are presented in **Figure 45**. These results indicate that, if only preexisting $PM_{2.5}$ speciation data from the Pittsburgh MSA are used, the optimal start date (i.e., the start date that maximizes the number of days on which a complete set of $PM_{2.5}$ speciation data are available) for a 4-year study is April 12, 2001. However, as shown in **Table 21**, none of the $PM_{2.5}$ chemical components satisfy the 85% data completeness criterion for this optimal 4-year study. Rather, data completeness for a 4-year study beginning on April 12, 2001, ranges from 57% for fine particulate trace and crustal elements to 74% for fine particulate sulfate. Even the best possible 3-year exposure database that could be constructed using existing $PM_{2.5}$ speciation data from the Pittsburgh MSA would have only 66% data completeness for trace and crustal elements and 74% data completeness for EC and OC. (The inventory results used to perform this analysis did not include data from the PC-BOSS sampler at the Bruceton monitoring site. These data, if available, could improve data completeness for sulfate, nitrate, and EC/OC, but would have little impact on data completeness for trace and crustal elements, which are the species that most require additional data). Hence, these results indicate that a retrospective epidemiology study of $PM_{2.5}$ chemical components in the Pittsburgh region would only be feasible if archived $PM_{2.5}$ samples could be analyzed to augment the existing $PM_{2.5}$ chemical speciation data record.

Table 21: Data completeness statistics, by $PM_{2.5}$ species, for the optimal study periods determined from ***Figure 45****. Statistics are based on inventories of existing data that are available from sites in the Pittsburgh MSA.*

		Sulfate	**Nitrate**	**EC /OC**	**Trace/Crustal Elements**	**Complete Speciation**
6/10/01-6/8/04 (3 yr)	Days with Data	971	1000	812	727	720
	Study Days	1095	1095	1095	1095	1095
	% Complete	89%	91%	74%	66%	66%
6/10/01-12/8/04 (3.5 yr)	Days with Data	1033	1062	874	789	782
	Study Days	1278	1278	1278	1278	1278
	% Complete	81%	83%	68%	62%	61%
4/12/01-4/10/05 (4 yr)	Days with Data	1083	1112	971	838	831
	Study Days	1460	1460	1460	1460	1460
	% Complete	74%	76%	67%	57%	57%
10/11/00-4/10/05 (4.5 yr)	Days with Data	1106	1131	1147	860	849
	Study Days	1643	1643	1643	1643	1643
	% Complete	67%	69%	70%	52%	52%
4/12/00-4/10/05 (5 yr)	Days with Data	1161	1183	1308	887	872
	Study Days	1825	1825	1825	1825	1825
	% Complete	64%	65%	72%	49%	48%

To determine whether analysis of archived $PM_{2.5}$ samples from the Pittsburgh MSA could provide enough additional chemical speciation data to make the retrospective epidemiology study feasible, we repeated the above analysis, but this time allowed for the inclusion of $PM_{2.5}$ speciation data that could be obtained by analyzing the archived $PM_{2.5}$ samples identified in **Section 2.3**. In addition to the assumptions set forth in the preceding paragraph, the following assumptions were made regarding the analysis of archived $PM_{2.5}$ samples, based on the considerations discussed in previous sections:

- Nitrate data can only be obtained from $PM_{2.5}$ samples that have been stored under refrigeration since collection.

- By the time the retrospective epidemiology study begins, all $PM_{2.5}$ samples being archived by the ACHD and the PA DEP will have been removed from refrigerated storage and transferred to storage at room temperature.

- Inorganic ions can be determined from samples collected on quartz or Teflon filters.

- Carbonaceous species can be determined from samples collected on quartz, but not Teflon, filters.

- Trace and crustal element species can be determined from samples collected on Teflon, but not quartz, filters.

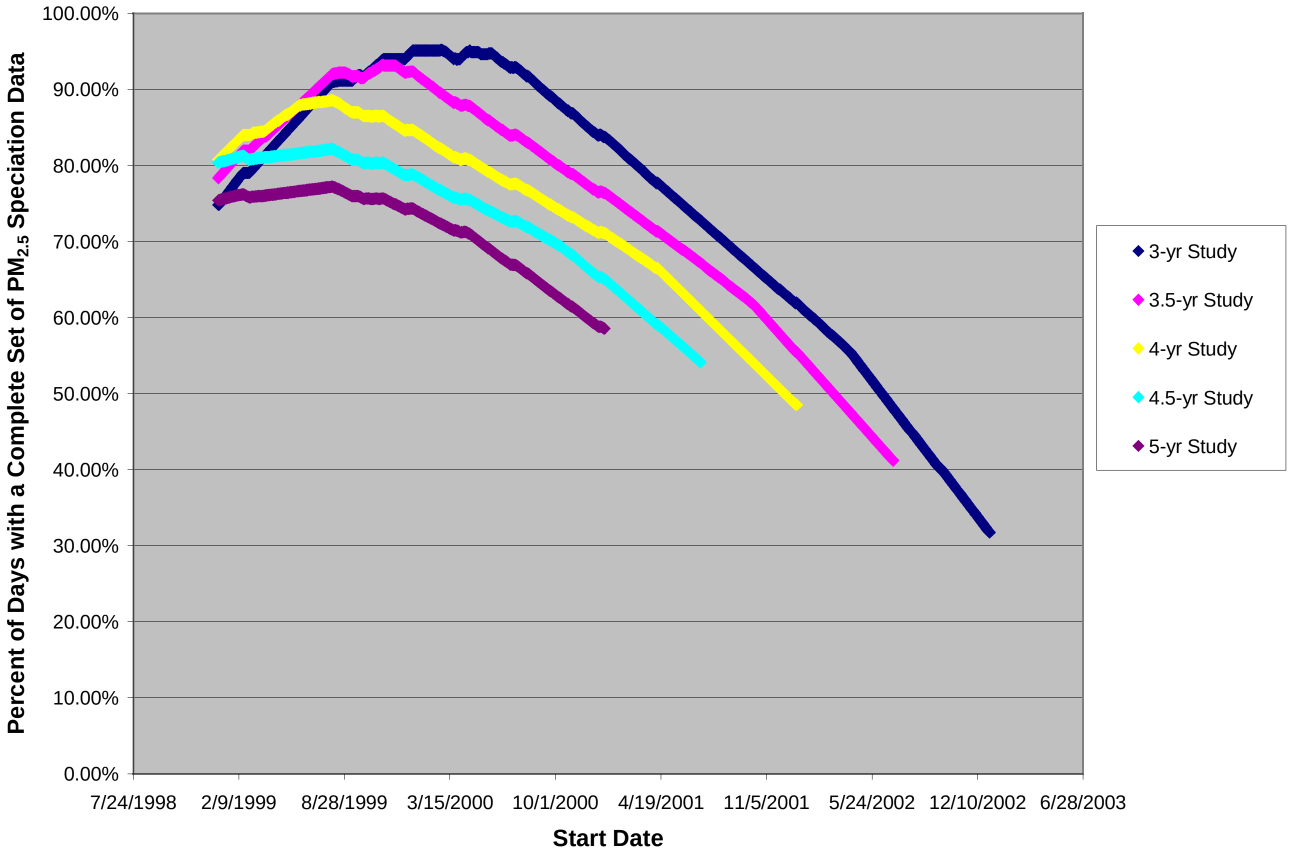


Figure 46: Effect of start date and study length on $PM_{2.5}$ speciation data completeness (i.e., the percentage of days on which sulfate, nitrate, carbon, and elements were each measured at one or more sites in the Pittsburgh MSA, regardless of whether all of the species were measured at the same site) for the case in which both preexisting data and data that can be obtained from the analysis of archived $PM_{2.5}$ samples are used.

Results of the analysis are presented in **Figure 46** and in **Table 22**. These results indicate that, if both preexisting $PM_{2.5}$ speciation data and $PM_{2.5}$ speciation data that can be obtained by analyzing archived $PM_{2.5}$ samples are used, then it is possible to construct a four-year database having at least 85% data completeness for each $PM_{2.5}$ species of interest. Again, this result is based on the assumption that data from the various monitoring sites (and measurement techniques) in the Pittsburgh MSA can be used interchangeably to represent the exposures of the region's population, such that if a given species was measured at any one or more of the Pittsburgh MSA's monitoring sites (or can be determined from an archived sample collected by any of these sites) on a particular day, then that species is considered to have been adequately characterized on that day for purposes of the data completeness calculation. Based on the sensitivity analysis presented in **Figure 46**, the optimal start date for a four-year retrospective study is August 3, 1999. As shown in **Table 22**, if all available archived $PM_{2.5}$ samples that were collected by monitoring sites in the Pittsburgh MSA between August 3, 1999, and August 1, 2003, were analyzed for chemical speciation according to the assumptions outlined above, then each of sulfate, nitrate, EC/OC, and trace and crustal element species would have greater than 90% data completeness during the four-year study period. August 3, 1999, is also the optimal starting date for a 4.5-year retrospective study; such a study would have greater than 85% data completeness for all $PM_{2.5}$ chemical components of interest, per the results presented in **Table 22**. Even when allowing for the use of data obtained from archived $PM_{2.5}$ samples, however, there are insufficient EC and OC data available from the Pittsburgh MSA to enable a five-year retrospective epidemiology study with 85% data completeness for these species. Nevertheless, the results presented in **Figure 46** and in **Table 22** confirm that, by combining existing $PM_{2.5}$ speciation data with data that can obtained by chemically analyzing archived $PM_{2.5}$ samples, it is possible to construct a database of sufficient length for a retrospective epidemiology study of $PM_{2.5}$ chemical components in the Pittsburgh region.

*Table 22: Data completeness statistics, by $PM_{2.5}$ species, for the optimal study periods determined from **Figure 46**. Statistics are based on inventories of existing $PM_{2.5}$ data and archived $PM_{2.5}$ samples that are available from sites in the Pittsburgh MSA.*

		Sulfate	Nitrate	EC /OC	Trace/Crustal Elements	Complete Speciation
2/26/00-2/24/03 (3 yr)	Days with Data	1095	1077	1059	1095	1042
	Study Days	1095	1095	1095	1095	1095
	% Complete	100%	98%	97%	100%	95%
11/7/99-5/7/03 (3.5 yr)	Days with Data	1266	1247	1211	1264	1191
	Study Days	1278	1278	1278	1278	1278
	% Complete	99%	98%	95%	99%	93%
8/3/99-8/1/03 (4 yr)	Days with Data	1430	1410	1344	1413	1293
	Study Days	1460	1460	1460	1460	1460
	% Complete	98%	97%	92%	97%	89%

		Sulfate	**Nitrate**	**EC /OC**	**Trace/Crustal Elements**	**Complete Speciation**
8/3/99-1/31/04 (4.5 yr)	Days with Data	1613	1593	1401	1596	1350
	Study Days	1643	1643	1643	1643	1643
	% Complete	98%	97%	85%	97%	82%
8/4/99-8/1/04 (5 yr)	Days with Data	1796	1733	1460	1779	1409
	Study Days	1825	1825	1825	1825	1825
	% Complete	98%	95%	80%	97%	77%

For economic reasons, it may not be practical to analyze all of the archived $PM_{2.5}$ samples that are available from $PM_{2.5}$ speciation monitoring sites in the Pittsburgh region. The inventory results presented in **Sections 2.2 and 2.3** suggest that, at a minimum, existing data and archived samples from the Bruceton, Lawrenceville, and Schenley Park monitoring sites should be used to develop exposure estimates for the retrospective epidemiology study. All three of these sites are located in Allegheny County, where more than half of the population of the Pittsburgh MSA resides, and all three featured daily collection of $PM_{2.5}$ speciation data or of archived $PM_{2.5}$ samples for at least a one-year period between 1999 and 2005. (The Liberty Borough monitoring site also meets these criteria, but it is sited to monitor the impact of emissions from a coke production facility on the local air quality in a particular portion of the Pittsburgh MSA, and therefore is less useful for estimating the exposures of the larger region's population). As discussed earlier, the Bruceton site is the largest source of fine particulate EC and OC data from the Pittsburgh MSA during the time period of interest for the proposed epidemiology study, and it is also the largest source of archived daily $PM_{2.5}$ samples that are being stored under refrigeration. The Lawrenceville site, which included $PM_{2.5}$ speciation monitoring by three separate groups between 1999 and 2005, features more days with a complete suite of $PM_{2.5}$ chemical component data than do any of the other sites in the Pittsburgh MSA, as well as the largest number of archived daily $PM_{2.5}$ samples of the sites in the Pittsburgh MSA. Finally, the Schenley Park site features the longest contiguous period during which the complete suite of $PM_{2.5}$ chemical component data was measured routinely on a daily basis at any site in the Pittsburgh MSA. To determine whether the Bruceton, Lawrenceville, and Schenley Park sites alone could provide sufficient $PM_{2.5}$ speciation information for the proposed retrospective epidemiology study, the analysis presented above in **Figure 46** and **Table 22** was repeated using only data and archived $PM_{2.5}$ sample inventory results from these three sites. The analysis concluded that the optimal start dates for 3-, 3.5-, 4-, 4.5-, and 5-year studies would be the same as those identified in **Table 22** for studies using data and archived samples from all of the $PM_{2.5}$ speciation monitoring sites in the Pittsburgh MSA. **Table 23** presents data completeness statistics as a function of study length, based on these optimal start dates, for the case in which only data and archived samples from the Bruceton, Lawrenceville, and Schenley Park sites are used to develop the exposure database. Results indicate that even if only these three monitoring sites are used as the basis for developing exposure estimates, it would be possible to construct a 4-year exposure database with greater than 90% data completeness or a 4.5-year exposure database with greater than 85% data completeness for all $PM_{2.5}$ chemical components of interest.

Hence, when evaluating the cost of analyzing archived filter-based $PM_{2.5}$ samples to supplement the existing record of $PM_{2.5}$ speciation data for purposes of a retrospective epidemiology study, we first considered a base case in which only archived samples from the Bruceton and Lawrenceville sites would be analyzed (there are few or no samples from the Schenley Park site requiring analysis), and then examined the incremental cost of analyzing archived samples from each of the other $PM_{2.5}$ speciation monitoring sites in the Pittsburgh MSA. Cost estimates were performed for each of the optimal 3-, 3.5-, 4-, 4.5-, and 5-year study periods identified in **Table 22**. First, to approximate the maximum cost of archived sample analysis for each scenario, we assumed that, for a given study period and group of included monitoring sites, all $PM_{2.5}$ samples available from those sites during that period would be analyzed to determine as much chemical speciation information as possible, regardless of whether such analysis would produce replicate results for a particular site. As such, we assumed that all archived Teflon-filter-based $PM_{2.5}$ samples would be analyzed for trace and crustal elements by XRF and for inorganic ions (i.e., SO_4^{2-}, NO_3^-, NH_4^+) by IC, and that all archived quartz-filter-based $PM_{2.5}$ samples would be analyzed for elemental and organic carbon by TOT and for inorganic ions by IC, irrespective of whether or not these samples were being stored under refrigeration. In addition, we increased the number of samples by 10% prior to costing to account for the cost of analyzing blank samples, which were not included in the data inventory but must be analyzed for QA/QC purposes. Based on the study team's experience with the cost of analyzing filter-based $PM_{2.5}$ samples for chemical speciation and on prices cited by several commercial laboratories (i.e., RTI International, Research Triangle Park, NC and Sunset Laboratories, Tigard, OR) that specialize in $PM_{2.5}$ sample analysis, we assumed prices of $30/sample for IC, $50/sample for TOT, and $70/sample for XRF.

*Table 23: Data completeness statistics, by $PM_{2.5}$ species, as a function of study length for the case in which only data and archived $PM_{2.5}$ samples from the Bruceton, Lawrenceville, and Schenley Park sites are used to develop the exposure database. For each study length, statistics were computed for the optimal period of data availability determined according to the procedure illustrated in **Figure 46**.*

		Sulfate	**Nitrate**	**EC /OC**	**Trace/Crustal Elements**	**Complete Speciation**
2/26/00-2/24/03 (3 yr)	Days with Data	1087	1036	1059	1067	987
	Study Days	1095	1095	1095	1095	1095
	% Complete	99%	95%	97%	97%	90%
11/7/99-5/7/03 (3.5 yr)	Days with Data	1258	1206	1209	1229	1127
	Study Days	1278	1278	1278	1278	1278
	% Complete	98%	94%	95%	96%	88%
8/3/99-8/1/03 (4 yr)	Days with Data	1422	1369	1339	1378	1226
	Study Days	1460	1460	1460	1460	1460
	% Complete	97%	94%	92%	94%	84%
8/3/99-1/31/04 (4.5 yr)	Days with Data	1605	1552	1392	1561	1279
	Study Days	1643	1643	1643	1643	1643
	% Complete	98%	94%	85%	95%	78%

		Sulfate	Nitrate	EC /OC	Trace/Crustal Elements	Complete Speciation
8/4/99-8/1/04 (5 yr)	Days with Data	1787	1692	1449	1743	1336
	Study Days	1825	1825	1825	1825	1825
	% Complete	98%	93%	79%	96%	73%

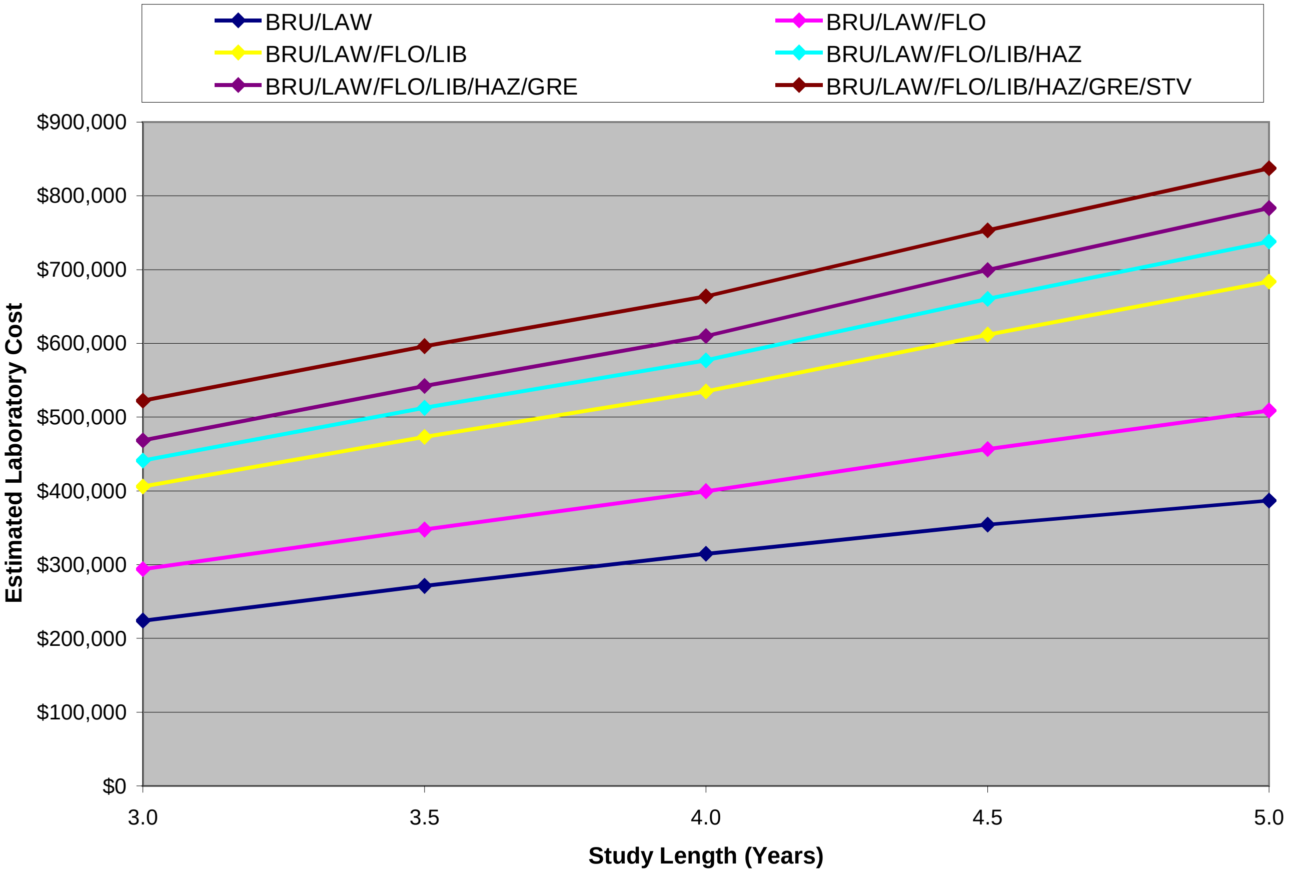


Figure 47: Estimated maximum cost for laboratory analysis of archived $PM_{2.5}$ samples, as a function of the length of the study and the group of monitoring sites from which samples are being analyzed.

Estimated filter analysis costs for this maximum-cost scenario are shown in **Figure 47** as a function of the length of the study and the group of monitoring sites from which archived samples are being analyzed. It is important to recognize that these costs only reflect the cost of laboratory analysis, and do not include the costs associated with collecting and organizing the archived $PM_{2.5}$ samples or the costs associated with reducing the laboratory results and integrating them into the exposure database for use in the retrospective epidemiology study. The incremental cost associated with analyzing samples from any given site can be calculated by differencing the appropriate lines in the plot. As shown in **Figure 47**, the estimated maximum sample analysis cost for a 4-year study under the base case scenario, in which only archived $PM_{2.5}$ samples from the Bruceton and Lawrenceville sites are analyzed to provide supplemental $PM_{2.5}$ speciation data, is approximately $315,000, and the estimated maximum sample analysis cost for a 4.5-year study under this scenario is approximately $355,000. Inclusion of archived $PM_{2.5}$ sample analysis from additional sites substantially increases the cost. The estimated maximum sample analysis cost for a 4.5-year study in which archived $PM_{2.5}$ samples from all of the speciation monitoring sites in the Pittsburgh MSA are analyzed to provide supplemental $PM_{2.5}$ speciation data is about $755,000, or $400,000 greater than the cost of a 4.5-year study under the base case scenario.

Actual costs will likely be less than those shown in **Figure 47**. The estimates presented for the maximum-cost scenario assumed that all archived $PM_{2.5}$ samples will be analyzed for inorganic ions by IC. However, as discussed in **Sections 2.3 and 2.4.2**, reliable NO_3^- and NH_4^+ concentrations may not be able to be obtained from samples that have been stored at room temperature rather than under refrigeration. If it is determined that NO_3^- and NH_4^+ data cannot be obtained from these samples, then concentrations of SO_4^{2-}, the only other major inorganic ion of interest, can be reliably estimated from XRF sulfur determinations, saving the cost of IC analysis. Moreover, the estimates shown in **Figure 47** assumed that all available filters from the sites under consideration would be analyzed for chemical speciation, regardless of whether such analysis produced duplicate results for a given site. While some duplicate results are desired so that existing speciation data can be used to verify the quality of results obtained from archived filter analyses, these results must only be obtained to the extent necessary to produce a statistically valid sample size for the comparison. Hence, for sites from which both existing $PM_{2.5}$ speciation data and an archived $PM_{2.5}$ sample are available on a large number of days, only a subset of the duplicate archived samples must be analyzed. Finally, if it is assumed that for the Bruceton and St. Vincent College sites, which did not sample from midnight to midnight, data from two consecutive days must be averaged to produce a valid 24-hour midnight-to-midnight concentration estimate, then it follows that an archived sample from either of these sites would only need to be analyzed if a sample or valid data point from an adjacent 24-hour period is also available. Based on these considerations, we developed a minimum-cost estimate for the laboratory analysis of archived $PM_{2.5}$ samples using the following assumptions:

1. For each monitoring site, up to 100 archived samples that were collected during the study period on days already having preexisting $PM_{2.5}$ speciation data will be analyzed to produce duplicate results for

use in establishing the validity of the results obtained from the archived samples (by comparison with the existing, collocated speciation data). These duplicate samples will be analyzed to determine as much chemical speciation information as possible, such that all archived Teflon-filter-based $PM_{2.5}$ samples will be analyzed for trace and crustal elements by XRF and for inorganic ions by IC, and all archived quartz-filter-based $PM_{2.5}$ samples will be analyzed for elemental and organic carbon by TOT and for inorganic ions by IC, regardless of whether or not these samples were being stored under refrigeration. (Analysis of non-refrigerated samples for inorganic ions by IC is necessary to determine whether reliable nitrate and ammonium data can be obtained from these samples). Any additional archived samples that would produce duplicate results in excess of 100 will not be analyzed.

2. With the exception of the duplicate samples identified above, archived samples will only be analyzed for ions by IC if they have been stored under refrigeration since collection. This assumption presupposes that the comparison of duplicate results from archived samples and preexisting data will indicate that reliable NO_3^- and NH_4^+ cannot be reliably obtained from non-refrigerated samples, and that SO_4^{2-} concentrations can in all cases be estimated accurately from XRF sulfur determinations.

3. Regarding the Lawrenceville site, for days on which archived Teflon-filter-based $PM_{2.5}$ samples are available from both the ACHD and UORVP monitoring programs (and no preexisting elemental data are available), only the ACHD sample will be analyzed by XRF.

4. Regarding the Bruceton and St. Vincent College Sites, archived samples will only be analyzed if their results can be combined with existing data or archived sample results from an adjacent 24-hour period to produce a valid midnight-to-midnight concentration estimate.

5. As with the maximum-cost estimate presented in **Figure 47**, we increased the number of samples by 10% prior to costing to account for the cost of analyzing blank samples for QA/QC purposes, and assumed prices of $30/sample for IC, $50/sample for TOT, and $70/sample for XRF.

Estimated filter analysis costs for this minimum-cost scenario are shown in **Figure 48** as a function of the length of the study and the group of monitoring sites from which archived samples are being analyzed. Again, these costs only reflect the cost of laboratory analysis, and do not include the costs associated with collecting and organizing the archived $PM_{2.5}$ samples or the costs associated with reducing the laboratory results and integrating them into the exposure database for use in the retrospective epidemiology study. The cost estimates presented in **Figure 48** on the basis of the assumptions outlined above are substantially less than those presented in **Figure 47**. The estimated sample analysis cost for a 4-year study including archived samples from only the Bruceton and Lawrenceville sites is $225,000 under the minimum-cost scenario shown in **Figure 48**, or $90,000 less than the estimated cost of such a study under the maximum-cost scenario shown in **Figure 47**. For a 4-year study in which archived samples from all of the speciation monitoring sites in the Pittsburgh MSA are utilized, the estimated sample analysis cost of about $490,000 under the minimum-cost scenario is about $175,000 less than the estimated maximum cost shown in

Figure 47. The estimates presented in Figures 46 and 47 are intended to represent a realistic range of laboratory costs for various study designs. For a given study length and combination of included monitoring sites, we expect that the actual cost of archived sample analysis will be closer to the minimum presented in **Figure 48** than to the maximum presented in **Figure 47**, although the exact cost will depend especially on whether IC analyses of non-refrigerated samples are required (i.e., whether it is determined that reliable nitrate and ammonium estimates can be obtained from these samples), on the actual number of blank and duplicate samples analyzed, and on the actual per-sample prices for performing the analyses.

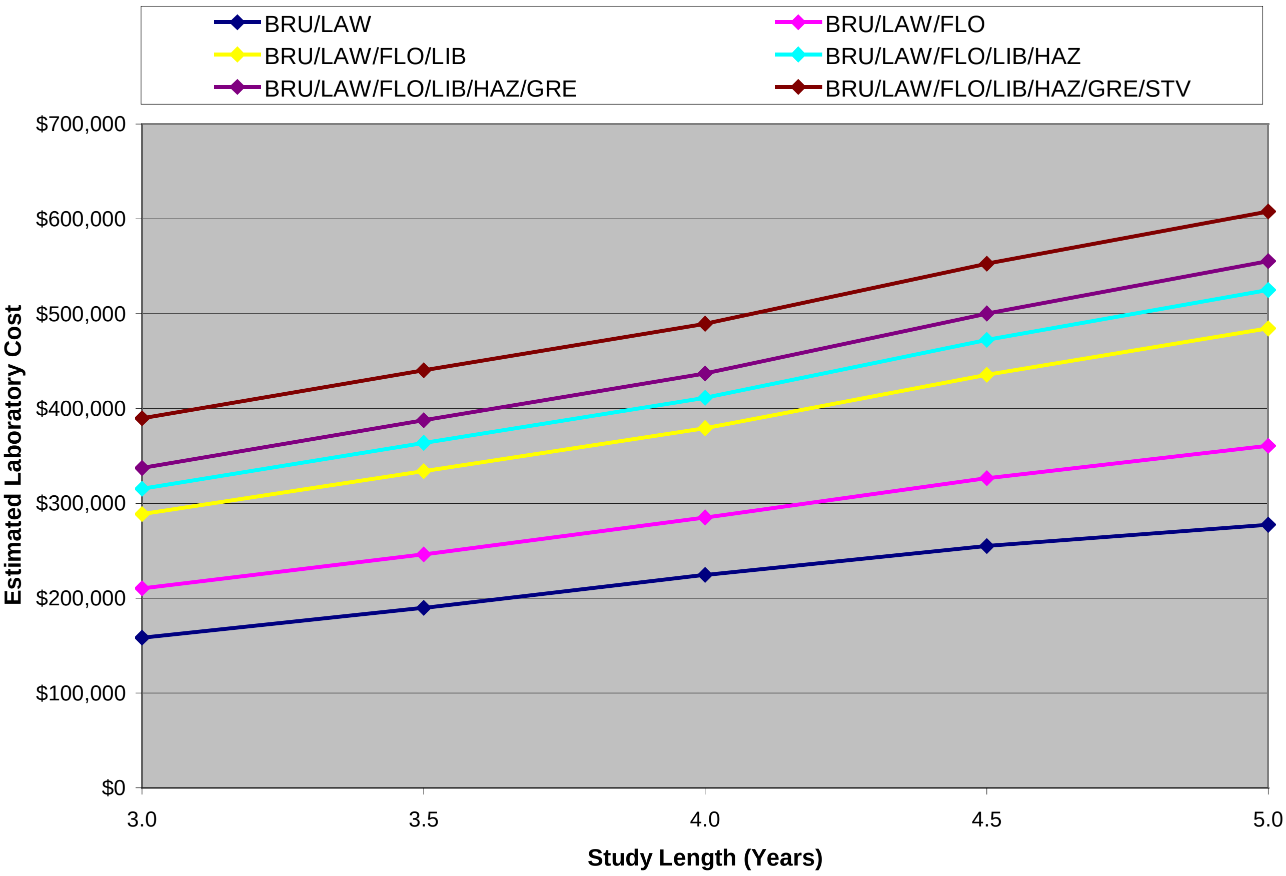


Figure 48: Estimated minimum cost for laboratory analysis of archived $PM_{2.5}$ samples, as a function of the length of the study and the group of monitoring sites from which samples are being analyzed.

As stated earlier in this subsection, the largest limitation regarding the design of the proposed retrospective epidemiology study of $PM_{2.5}$ in the Pittsburgh region is the limited availability of exposure data, particularly for $PM_{2.5}$ chemical components. Selection of a final design for the retrospective epidemiology study requires a careful consideration of the availability, quality, and completeness of these data, the costs associated with obtaining them, and the constraints imposed by statistical power and the spatial distribution of pollutants throughout the study region. Based on the analyses presented throughout this report, we conclude that there are a sufficient quantity and quality of existing data and archived $PM_{2.5}$ samples available to permit a retrospective epidemiology study of $PM_{2.5}$ from coal-fired power plants in the Pittsburgh region, and propose as a reasonable design a four-year (1,460-day) study focusing on the Pittsburgh MSA between August 3, 1999, and August 1, 2003. The study would utilize existing $PM_{2.5}$ speciation data collected by all of the speciation monitoring sites located in the Pittsburgh MSA, as well as additional $PM_{2.5}$ speciation data that would be obtained by analyzing archived $PM_{2.5}$ samples from the Bruceton, Lawrenceville, Liberty Borough, Florence, and St. Vincent College sites. (As discussed earlier, existing $PM_{2.5}$ speciation data collected by monitoring sites outside of the Pittsburgh MSA could also be included in geostatistical models to inform the exposure estimates developed for the MSA). Estimates of $PM_{2.5}$ total mass concentrations, co-pollutant concentrations, and meteorological conditions in the Pittsburgh MSA would be obtained by geostatistically averaging data from the numerous sites identified in **Section 2.2** that collected these data in the greater Pittsburgh region.

Table 24 summarizes the availability of $PM_{2.5}$ speciation data for the proposed 4-year study. As with the statistics presented in **Table 22**, the numbers shown in **Table 24** were tabulated on the premise that nitrate concentrations will only be able to be determined from archived $PM_{2.5}$ samples that have been stored under refrigeration since collection. Under this assumption, the laboratory cost for archived sample analysis (excluding the costs of sample retrieval and data reduction) is estimated to be about $430,000 for our proposed design, per the results presented in **Figure 48**.

Table 24: Summary of data availability, by $PM_{2.5}$ species, for a 4-year retrospective epidemiology study focusing on the Pittsburgh MSA between August 3, 1999, and August 1, 2003, and including chemical speciation analysis of archived $PM_{2.5}$ samples that were collected at the Bruceton, Lawrenceville, Liberty Borough, Florence, and St. Vincent College sites during that period.

	Sulfate	**Nitrate**	**EC/OC**	**Trace/Crustal Elements**
# of Days with Data from 1 Site	112	445	725	135
# of Days with Data from 2 Sites	111	557	416	190
# of Days with Data from 3 Sites	254	222	46	293
# of Days with Data from 4 Sites	412	36	46	412
# of Days with Data from 5 Sites	276	61	57	207
# of Days with Data from 6 Sites	171	53	54	122
# of Days with Data from 7 Sites	48	36	0	40
# of Days with Data from 8 Sites	46	0	0	14
Total # of Days with Data from Any Site(s)	1430	1410	1344	1413
# of Study Days	1460	1460	1460	1460
Data Completeness[a]	97.9%	96.6%	92.1%	96.8%

[a] Data Completeness = Total # of Days with Data from Any Site(s) divided by the # of Study Days

The start date of August 3, 1999, was selected on the basis of the sensitivity analysis presented in **Figure 46**, which identified this as the optimal start date for a 4- to 4.5-year retrospective study of $PM_{2.5}$ chemical components in the Pittsburgh MSA. The 4-year study length is expected to afford sufficient statistical power for the retrospective epidemiology study, based on the analyses presented later in this report. Although a 4.5-year study, if feasible, would provide even greater power, the only fine particulate elemental and organic carbon data that are available from the Pittsburgh MSA between August 1, 2003 (the end date for a 4-year study beginning on August 3, 1999) and January 1, 2004 (the end date for a 4.5-year study beginning on August 3, 1999) were collected every third day, and hence are characterized by numerous, regularly occurring missing values that could be statistically problematic. This, coupled with the added cost of a 4.5-year study, resulted in the selection of a 4-year study.

The proposed study design calls for analysis of archived $PM_{2.5}$ samples from each of the five monitoring sites located in the Pittsburgh MSA that collected $PM_{2.5}$ speciation data or archived $PM_{2.5}$ filters on a daily basis for at least one year during the four-year study period. These sites are well situated to represent the diversity of ambient air pollution in the Pittsburgh MSA, as they include sites located in urban (Lawrenceville), suburban (Bruceton), and industrialized (Liberty Borough) areas of Allegheny County, as well as sites located in more remote areas to the west (Florence) and east (St. Vincent College) of Allegheny County. The data presented in **Table 23** demonstrate that it would be possible to assemble a four-year exposure database for the Pittsburgh MSA with greater than 90% data completeness for all $PM_{2.5}$ species of interest if only archived $PM_{2.5}$ samples from the Bruceton and Lawrenceville monitoring sites were analyzed. However, such a strategy, while less costly than the one being proposed, would result in a large number of days on which data for one or more $PM_{2.5}$ species were only available from a single site in the Pittsburgh MSA. As discussed in **Section 2.4**, for $PM_{2.5}$ trace and crustal elements in particular, concentrations measured at any given site in the Pittsburgh MSA are not necessarily representative of concentrations throughout the rest of the region. Hence, the recommendation to analyze archived $PM_{2.5}$ filters from the Florence, Liberty Borough, and St. Vincent College sites is intended to increase the number of days for which trace and crustal element data are available from multiple sites so that exposures to these species can be more reliably estimated. As shown in **Table 25**, under the proposed design, $PM_{2.5}$ elemental data would be available from at least two sites in the Pittsburgh MSA on about 88% of the study days, and from at least three sites in the Pittsburgh MSA on about 75% of the study days. If only archived samples from the Bruceton and Lawrenceville sites were analyzed, just 67% of the study days would include elemental data from multiple sites, and just 23% would include data from at least three sites. Analysis of archived filters from the Hazelwood and Greensburg sites would slightly increase the number of days with elemental data from multiple monitoring sites; however, this increase is not substantial enough to justify the additional $57,000 or more in laboratory costs associated with performing these analyses.

The analyses presented in **Section 2.4.1.2** indicate that semi-continuous measurements of sulfate and nitrate that were made at the Bruceton and Schenley Park monitoring sites are likely to introduce substantial noise when used interchangeably with integrated measurements of these species, and therefore should only be used when necessary to provide estimates for days that otherwise would have no available data. The statistics presented in **Table 24** include these semi-continuous data; however, there are only two days during the proposed four-year study period on which a semi-continuous measurement is the only source of nitrate data for the Pittsburgh MSA, and there are no days for which a semi-continuous measurement is the only source of sulfate data. Hence, the study could easily be conducted without using any semi-continuous $PM_{2.5}$ ion data. Semi-continuous measurements are the only source of fine particulate EC and OC data on 539 (or about 37%) of the study days; however, as shown in **Section 2.4.1.3**, these semi-continuous carbon measurements were generally comparable (after correction for relative bias) to measurements made by applying thermal optical transmittance to integrated, quartz-filter-based $PM_{2.5}$ samples. Therefore, they are considered appropriate for use in the study.

Table 25: $PM_{2.5}$ trace and crustal element data availability as a function of the sites from which archived $PM_{2.5}$ samples are analyzed for chemical speciation. All percentages are based on a four-year retrospective epidemiology study focusing on the Pittsburgh MSA between August 3, 1999, and August 1, 2003, and assume that any existing $PM_{2.5}$ elemental data from the Pittsburgh MSA are included in the study.

Monitoring Sites Included in Archived Filter Analysis	Percent of Days with Elemental Data from at Least 1 Site in the Pittsburgh MSA	Percent of Days with Elemental Data from at Least 2 Sites in the Pittsburgh MSA	Percent of Days with Elemental Data from at Least 3 Sites in the Pittsburgh MSA
BRU, LAW	94%	67%	23%
BRU, LAW, FLO	95%	73%	52%
BRU, LAW, FLO, LIB	97%	86%	71%
BRU, LAW, FLO, LIB, STV	97%	88%	75%
BRU, LAW, FLO, LIB, STV, GRE, HAZ	97%	88%	76%

Because of remaining uncertainties regarding the quality of elemental results that can be obtained by XRF analysis of archived $PM_{2.5}$ samples and the feasibility of obtaining ammonium and nitrate data from archived samples that have not been kept refrigerated, we recommend that work on assembling the exposure database for the proposed retrospective epidemiology study be carried out in two phases, separated by a decision point. Tasks to be performed under the first phase include:

1. Obtaining and organizing all existing $PM_{2.5}$ speciation data that were collected by monitoring sites in the 35-county greater Pittsburgh region during the study period. This task has largely been completed as part of the current feasibility assessment.

2. Requesting, obtaining, and organizing all archived $PM_{2.5}$ samples (including blanks and duplicates) that were collected at the Bruceton, Lawrenceville, Florence, Liberty Borough, and St. Vincent

College sites between August 3, 1999, and August 1, 2003. Sampler operating data and QA/QC data regarding these samples must also be gathered. For the Florence site in particular, this task must be completed expeditiously, as some samples are scheduled to be discarded in April 2007.

3. For each site, chemically analyzing up to 100 archived $PM_{2.5}$ samples that were collected on days from which collocated $PM_{2.5}$ speciation data are already available. Teflon-filter-based samples will be analyzed for trace and crustal elements by XRF and for inorganic ions by IC, and quartz-filter-based samples will be analyzed for EC and OC by TOT and for inorganic ions by IC.

4. Applying latent variable modeling and Bland-Altman analyses to develop calibrations relating the archived sample results to the existing speciation data.

Based on the quality of these calibrations, a decision will be made regarding plans for analysis of the remaining archived $PM_{2.5}$ samples. This decision point provides a means for avoiding unnecessary project costs by ensuring that only analyses that will contribute valuable data to the study are performed. Tasks to be performed under the second phase would then include:

1. Chemically analyzing the remaining archived $PM_{2.5}$ samples according to the plan decided on above.

2. Reducing and assuring the quality of all data produced by the chemical analysis of archived $PM_{2.5}$ samples.

3. Obtaining and organizing all existing co-pollutant and meteorological data that were collected by monitoring sites in the 35-county Pittsburgh region during the study period.

4. Applying a consistent set of QA/QC standards to the data collected by the various monitoring sites (e.g., per the discussion in **Section 2.4.3**).

5. Mathematically adjusting data (e.g., using the calibrations developed at the end of Phase 1) to account for relative biases resulting from discrepancies in measurement techniques, blank correction practices, archiving procedures, etc.

6. Aggregating data to compute 24-hour, midnight-to-midnight average values for each parameter at each monitoring site (e.g., for data that were measured with a finer-than-daily time resolution or for daily data that were not measured from midnight-to-midnight).

7. Assembling the reduced, validated, daily data from all sites into a final comprehensive database for use in geostatistical and epidemiological modeling.

3 Health Outcomes Data Assessment

3.1 Introduction

A critical component of any retrospective epidemiological study of $PM_{2.5}$ and health outcomes is the identification of readily available and accessible mortality and morbidity databases for the region of interest. Given the trends in improved treatments for disease, mortality alone is unlikely to be a sensitive enough indicator to capture all potential effects of daily changes in air pollution on health. Ideally, in addition to mortality data, daily or even hourly medical information would be available to capture all health-related outcomes in the population potentially related to variations in $PM_{2.5}$ concentrations and/or its components. For the time period of interest, this information would include but not be limited to deaths, hospital admissions, emergency room visits, physicians' office visits, prescriptions, and medication use, symptomatology, and others, preferably all in electronic format. Although the level of detail required for perhaps the "ideal" comprehensive retrospective assessment with multiple, novel outcomes of interest (e.g. limited access to prescriptions, and medication use, symptomatology) might not be available in retrospective datasets, a number of health outcomes, including mortality, hospitalizations and emergency department visits (ED), in the Pittsburgh region are captured by several data collection entities. Some of these entities are unique to this geographic area and will enhance our ability to conduct a retrospective epidemiological assessment of health effects related to $PM_{2.5}$ and its component species.

From previous work and preliminary evaluation, several databases are known to be available from 1999 through 2006 and later that might be used to reconstruct retrospectively the health outcomes profiles of residents in the Pittsburgh region for studies of short term effects of fine particulates and its speciated components. For mortality, potential resources include, but are not limited to the National Center for Health Statistics (NCHS), the Pennsylvania Department of Health Bureau of Health Statistics, Allegheny County Health Department (ACHD), West Virginia Hospital Authority, and the Ohio Department of Health. For morbidity, hospital admissions might be the most well-defined and accessible estimate of the health effects potentially related to $PM_{2.5}$ and its components. These data are systematically collected for the region and are available retrospectively in administrative datasets via the Pennsylvania Health Care Cost Containment Council (PHC4), the West Virginia Hospital Authority and the Ohio Department of Health. Information on daily emergency room visits are available electronically from: the University of Pittsburgh Medical Center (UPMC) Medical Archival System (MARS), a proprietary software system of UPMC; the Real-time Outbreak Disease Surveillance (RODS) Laboratory, a University of Pittsburgh real-time computer-based public health surveillance system; and individual local and regional hospital databases.

To investigate long-term morbidity and mortality effects, the study team investigated the feasibility of entering into an agreement with local Health Plans and Health Maintenance Organizations (HMOs) for restricted access to de-identified and/or identified health care data of subscribers. Also, data from local population-based cohorts assembled in past years for national and regional research studies such as the Cardiovascular Health Study (CHS), Women's Health Initiative (WHI), Study of Women Across the Nation (SWAN), MR. FIT and others were explored as another potential resource for evaluating the health effects of long-term exposure to fine particulates. Medicare billing information, Verispan and IMS Health pharmaceutical use databases, and the UPMC Health Plan Pharmaceutical Database were also assessed as potential resources for a retrospective analysis, particularly in more susceptible age groups (i.e. Medicare-65 years and older; pharmaceutical databases-children).

3.2 Study Area of Interest for Health Outcomes Assessment

The study area of ultimate interest is defined by the population at risk for significant exposure to PM $_{2.5}$ from regional coal fired power plants, potentially in southwestern PA, eastern Ohio and the northern West Virginia panhandle. Allegheny County, Pennsylvania is home to the city of Pittsburgh proper, is the most population dense of all counties in southwestern Pennsylvania and houses the majority of the region's hospitals and industrial entities. Given, however, that the true population at risk for the retrospective assessment was to be determined during this project, we investigated the availability of health outcomes data for an expanded region of interest in this feasibility analysis. For example, the Pittsburgh Metropolitan Statistical Area (MSA) (circa 2000) is also a rather intuitive study area of interest and consists of 7 counties in southwestern Pennsylvania (outlined in in yellow); however, due to the regional nature of $PM_{2.5}$ and the wide distribution of power plants in southwestern Pennsylvania, the "combined core-based statistical area" might also represent the population at risk and includes, in addition to the

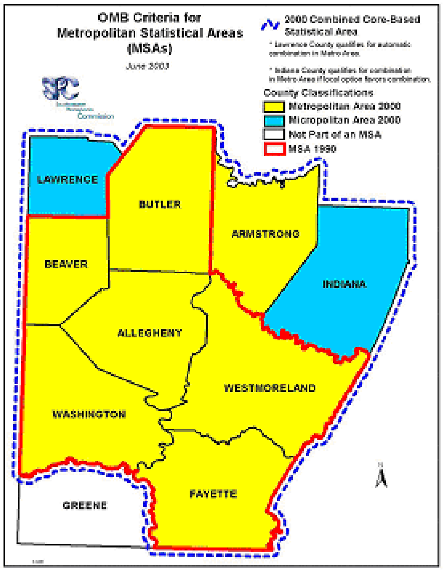

OMB Criteria for
Metropolitan Statistical Areas
(MSAs)
June 2003
2000 Combined Core-Based Statistical Area
* Lawrence County qualifies for automatic combination in Metro Area.
* Indiana County qualifies for combination in Metro Area if local option favors combination.
County Classifications
Metropolitan Area 2000
Micropolitan Area 2000
Not Part of an MSA
MSA 1990
LAWRENCE
BUTLER
ARMSTRONG
BEAVER
INDIANA
ALLEGHENY
WESTMORELAND
WASHINGTON
FAYETTE
GREENE

Figure 49: Pittsburgh metropolitan and combined core base statistical areas (http://www.spcregion.org/about_press_grow.shtm).

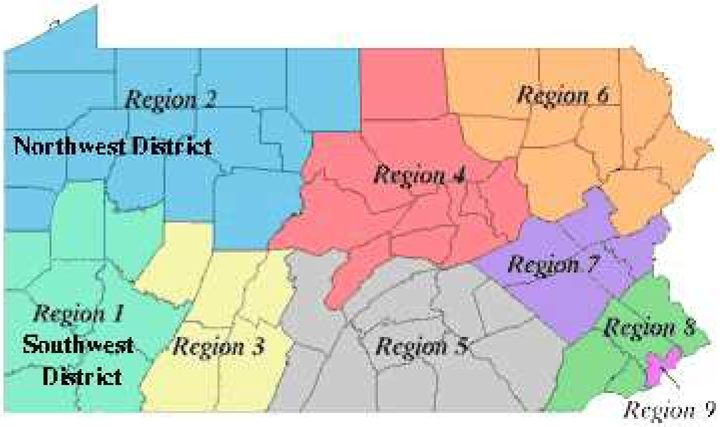


Figure 50: Pennsylvania Health Care Cost Containment Council (PHC4) regions (http://www.phc4.org/dept/dc/state.htm).

Pittsburgh MSA counties, Lawrence and Indiana Counties. Greene County is the southwestern most county in the region and likely to be included within the core statistical area in 2006-2007 (see also **Figure 49**). The various state and local health agencies have individualized definitions of "southwestern Pennsylvania." For example, while the PHC4 recognizes 8 counties in the Southwestern PA region (Region 1) (**Figure 50**), the PADOH considers a total of 11 counties as comprising the Southwest PA district, including Allegheny, Armstrong, Beaver, Butler, Cambria, Fayette, Greene, Indiana, Somerset, Washington and Westmoreland (**Figure 51**). Health data are often compiled by these agencies accordingly. In addition, certain more northern counties in western Pennsylvania, including Lawrence,

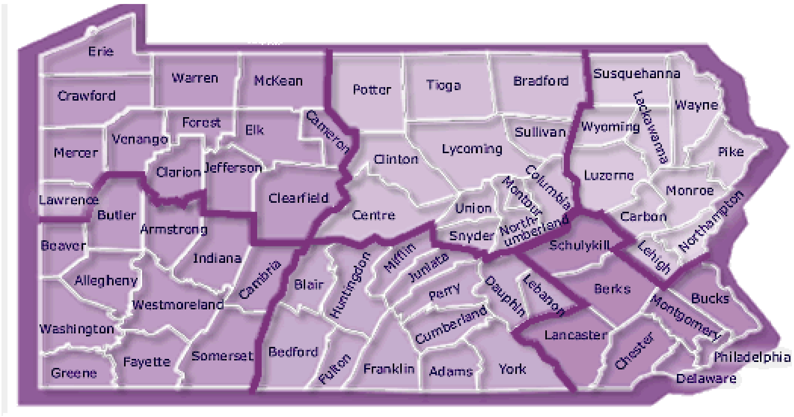


Figure 51: Regional Districts of the Pennsylvania Department of Health

(http://www.dsf.health.state.pa.us/health/cwp/view.asp?a=180&Q=199440).

Mercer, Venango, Clarion and Jefferson (Northwest District; **Figure 50**), might also be significantly affected by regional pollutants from power plants in the Ohio River Valley. Therefore, for the purpose of this feasibility analysis, possible sources of health outcomes data were explored in this expanded "Pittsburgh metropolitan region", including the 10 counties in the Pittsburgh combined core base statistical area, 6 counties in Pennsylvania bordering the core base statistical area (Cambria, Somerset, Jefferson, Clarion, Mercer, Venango) as well as adjacent counties in Ohio and West Virginia.

3.3 Inventory of Existing Health Outcomes Data

A comprehensive inventory and assessment of available mortality and morbidity datasets for the Pittsburgh metropolitan region was completed. As noted in the previous progress reports, a checklist was developed to evaluate variables of interest in existing health outcomes datasets available and accessible for the Pittsburgh region between 1999 and 2006 (**Appendix E**). A summary of the available data sets with identified strengths and weaknesses, particularly for time series analysis with speciated components, is presented below. The evaluation of the various health outcomes in cohort studies in relation to long term effects of air pollution is addressed in a separate section in this final report. A metadata database detailing available mortality and morbidity for the study area of interest was constructed and its data layout is included as an appendix (**Appendix F**).

Given the relative paucity of speciated $PM_{2.5}$ data, specifically from 1999-2001, for areas other than Allegheny and possibly Washington and Westmoreland Counties in Pennsylvania, we suggest that a retrospective study dating back to 1999 would have a more narrow regional focus than a prospective (longitudinal) study that could capitalize on a growing network of speciation monitors. Although we have explored multiple health outcomes data in an expanded region, the existence of both adequately monitored $PM_{2.5}$ and speciation data and the availability and quality of health outcomes data ultimately determined the focus of our proposed retrospective analysis as outlined in the Proposed Study section. The general conclusion is that retrospective mortality and hospitalization data are readily available for the expanded region of interest from 1999 to the present. Emergency department (ED) data in electronic format is likely available from local hospitals for Allegheny County residents dating back to 1999 and for Washington and Westmoreland county residents from 2001 to the present. For the outlying counties, electronic ED data is available only more recently (2004 to present) if at all from smaller community hospitals. The capture of health information from unscheduled physician office visits, pharmaceutical databases, and other less traditional datasets will most likely not be possible for a retrospective assessment but could potentially be compiled for a prospective study

3.4 Inventory and Assessment of Mortality Databases

3.4.1 National Center for Health Statistics (NCHS) Division of Vital Statistics

Mortality data in the United States are relatively well characterized for the 1999-2006 time period of interest for a retrospective study. The United States (US) Vital Statistics System has been operational in some form since the 1950s. The National Center for Health Statistics (NCHS) Division of Vital Statistics (DVS) assumed the responsibility for vital statistics program operations in the 1960s and has continued to serve as the primary compilation and contact agency to the present.

Mortality data are provided through contracts between NCHS and vital registration systems operated in the jurisdictions legally responsible for the registration of vital events. In the US, legal authority resides individually within the 50 states, 2 cities (Washington DC and New York City) and 5 territories (Puerto Rico, the Virgin Islands, Guam, American Samoa, and the Commonwealth of the Northern Mariana Islands.) NCHS compiles national mortality statistics from death certificates provided by these individual registrars. If mortality data from multiple states are required for a retrospective assessment, NCHS might be the repository of choice for all data acquisition.

To achieve the uniformity required for combining data from all states, cities and territories to provide national statistics, certain standards for certificates and reports are recommended by the NCHS as guides for use by individual registration offices. The most current standardized death certificate was revised in 2003. Standardization of mortality data across registrars is critical, particularly if any follow-back epidemiological study of the area of interest might eventually include parts of Ohio and/or West Virginia in addition to Pennsylvania.

Through the National Vital Statistics System, data on vital events are now published in electronic form. Data from public-use versions of these files are provided on CD-ROM. Confidentiality of medical data is a key aspect in the release of health outcomes data files, specifically since the passage of the Health Insurance Portability and Accountability Act of 1996 (HIPAA). In order to prevent disclosure of individuals and institutions, beginning with the 1989 data year for Births and Deaths Public-Use files, NCHS has excluded a) geographic identities of counties, cities, and metropolitan areas with less than 100,000 population and b) exact day of birth and death. Fetal Deaths and Linked Birth/Infant Death Public-Use files exclude identities of counties, cities, and metropolitan areas less than 250,000 population, as well as exact dates. Public-Use files can be requested by using the standard procedures for requesting vital statistics Public-Use files found at NCHS Publications and Information Products. Also NCHS Public-Use Data File Program and NCHS's Data Release Policy web links provide for more information on NCHS policy, including data use restrictions. As noted above, date of death, ZIP code or other potential individual-level identifiers are not included in public use files and require approval for protected access to these additional elements.

Customized (i.e., non-public-use) data files are defined as any files not included in the definition of Public-Use files as stated above. These would include files that identify all counties or smaller cities, or files that provide exact dates of birth or death, or combinations of these. To gain protected access to a customized vital statistics data file, researchers must send a letter, email, or fax with a complete description of the proposed use of the data, including why the data being requested are needed, the exact data items being requested, how the data will be utilized, who will be utilizing the data, and the time frame in which it will be needed, to the Director of Vital Statistics for review. If a request is denied, the requestor will be so notified by official letter.

If a request is approved, it may be determined, for reasons of confidentiality, that it is not appropriate for NCHS to physically release a customized data file directly to the requestor. Such requests will be referred to the NCHS Research Data Center (RDC), which allows for controlled access to the data files without their release. The RDC has specific procedures that must be followed and data requestors are charged for the services (http://www.cdc.gov/nchs/r&d/rdc.htm). Fees for remote on-line access are typically between $250 and $1000 per month depending on the number of records outlined in the request. Following approval of the request by the Director, Data User Agreements must be completed and signed by data requestors before customized data files are released or otherwise made available through the RDC. The Data User Agreement defines the specific requirements and restrictions on the use and disposal of the data by the requestor. DVS staff member will follow-up with data requestors, insuring that data user agreements are completed correctly and properly executed, creating the data files, and monitoring their disposition.

The United States implemented the latest (10th) revision of the International Classification of Diseases , Tenth Revision (ICD-10) starting with mortality data for 1999. However, deaths from 1999-2002 were coded in certain datasets according to the International Classification of Diseases, Ninth Revision (ICD-9), most probably due to the lag in ICD-10 usage by individual certifiers and/or registrars. Because of this issue, comparability of these datasets with later years and other data sources would need to be assessed. ICD9/ICD10 comparability ratio tables have been constructed by the Bureau of Health Information, Department of Healthcare Financing to assist in these analyses **(Appendix G).**

3.4.2 Pennsylvania Department of Health Bureau of Health Statistics and Research (other state health departments e.g. West Virginia, Ohio)

The Pennsylvania Department of Health may release confidential data to organizations or individuals only for specific medical research purposes. Academic researchers and other qualified entities can access Pennsylvania mortality data directly through the Pennsylvania Department of Health Bureau of Health Statistics and Research. Complete datasets are currently available from 1999-2005. Preliminary mortality data is available for 2006. Protected access datasets from the Bureau of Health Statistics and Research include street address and ZIP code of residence (Zip+ 4), as well as demographic variables such as age, race, education, marital status and occupation**.**

Institutional review board (IRB) approval of a project is mandatory prior to the requesting of data if a study or project requires the receipt of personal identifiers from Pennsylvania records. The process of obtaining confidential data is initiated with the submission of a completed “Application for Access to Protected Data”. Guidelines and procedures for these “follow-back” activities using Pennsylvania records are covered in detail in the Pennsylvania “User’s Guide for Access to Protected Data”. The application must be reviewed and approved by the Department prior to release of the information. Applications can be obtained by writing to the Director, Division of Health Statistics and Research. The application and user’s guide are available from the PADOH Bureau of Health Statistics and Research (contact: Raymond Powell, PADOH). Similar procedures are in place in West Virginian and Ohio.

3.4.3 Allegheny County (PA) Health Department (other local health departments)

The Allegheny County Health Department, as a large, local health department, also maintains a comprehensive mortality dataset derived from the Pennsylvania Department of Health registrar’s database. Other county health departments in the Southwestern PA region are less sophisticated in the maintenance of such databases.

The electronic database at the Allegheny County level has recorded deaths from the early 1990s. Variables available include age, gender, street address, ZIP code (5-digit), cause of death, date of death, and others. Geo-coding to street address and 5-digit (or potentially 9-digit) ZIP code for deaths that occur in the County is conducted as needed for specific projects. Death certificates are available at both the county and state levels; the county health department routinely conducts a verification analysis of the electronic database by comparison with actual death certificates. The error rate is approximately 2%. Allegheny County officials have de facto access to these documents. These data with identifiers can be accessed onsite at the ACHD by requesting entities. However, any database constructed for use off-site must have all individual level identifiers removed. If investigators partner with Allegheny County officials in the conduct of the retrospective epidemiological assessment, access to mortality data is enhanced.

3.4.4 Strengths and Weaknesses of Mortality as a Health Endpoint for Retrospective Air Quality Studies

Registration of deaths is mandatory in the United States. Therefore, case ascertainment of mortality as an endpoint is relatively straightforward and inclusive compared with other health outcomes (e.g. hospitalizations, ED visits, pharmaceutical usage, etc.) potentially associated with air pollution. Standardized mortality data are more readily available retrospectively for an expanded area of interest, such as the Pittsburgh core base statistical area or larger area spanning two or more states. Investigators can also select a “control disease” unrelated to air pollution (e.g. deaths from motor vehicle accidents, etc.) for model testing. Medical certification of the underlying cause of death constitutes a medico-legal opinion but is by no means absolute. Various errors and oversights may occur during certification due to lack of understanding as to how to complete the death certificate, including listing of causes in an incorrect order, listing more than one disease or condition on the same line, omitting the interval between onset and death, etc.

As noted previously, however, given the highly developed state of medical treatment for cardiopulmonary and other disease in the US today, mortality is most likely a relatively insensitive indicator of health effects related to the impact of short-term rise and fall of air pollutants, even in the most susceptible populations. These data are more likely of use in investigating, through existing large cohorts, the longer-term effects of pollutants assembled via HMOs and/ other healthcare providers and large-scale prospective or historical prospective epidemiological projects in the region.

Adequate power to detect a significant association (if one exists) is critical in the design of epidemiological studies. Mortality is a relatively rare outcome compared to hospitalizations and ED visits. Approximately 15,000 total (all-cause) deaths per year were reported in Allegheny County (PA) (population ~ 1.26 million) from 1999-2004 for a total of 90,664 deaths for the 6-year period (~ 41 deaths/day) (**Table 26)**. In the seven-county Pittsburgh MSA (population ~2.4 million); a total of 171,034 deaths were observed (~ 78 deaths/day); in the 10-county combined core base statistical area (population ~ 2.6 million), a total of 186,180 deaths were reported during the 1999-2004 time period (~ 85 deaths/day). Approximately 40% of deaths in the region were attributable to respiratory (influenza, pneumonia, emphysema) or cardiovascular causes. In comparison, total hospitalizations for the same time period in Allegheny County alone were approximately 1.2 million. Approximately 160 daily hospital admissions might be attributable to cardio-respiratory causes in Allegheny County hospitals. Daily ED visits are a full two to three fold higher.

Finally, it is unlikely that all pollution-related deaths are exclusively due to exposure to air pollutants shortly before death. Time-series models will likely underestimate overall mortality risk by failing to capture mortality associated with the influence of increased $PM_{2.5}$ or its components on the development over time of chronic diseases leading initially to frailty and subsequently death (Kunzli et al., 2001). Retrospective cohort as opposed to time series studies can capture this aspect of mortality but suffer from potential exposure misclassification and other biases. Retrospective cohorts are also often difficult to reconstruct in a comprehensive assessment.

Table 26: Total (All-cause) Deaths in Southwestern PA MSA and Combined Core Statistical Area by County of Residence 1999-2004 (all causes; ICD-10/ICD-9: A00-Z99/000-799, E800-E999).

County	Ave. Pop 1999-2004	Resident Deaths 2004	Resident Deaths 2003	Resident Deaths 2002	Resident Deaths 2001	Resident Deaths 2000	Resident Deaths 1999	Resident Deaths Total
Pittsburgh MSA								
Allegheny	1,260,000	14,507	15,104	15,100	15,478	15,072	15,403	**90,664**
Armstrong	72,000	835	842	912	864	866	878	**5,197**
Beaver	179,500	2,075	2,105	2,209	2,147	2,049	2,200	**12,785**
Butler	176.000	1,867	1,757	1,816	1,714	1,633	1,685	**10,472**
Fayette	146,000	1,746	1,813	1,809	1,861	1,813	1,869	**10,911**
Washington	203,000	2,483	2,354	2,449	2,568	2,445	2,443	**14,742**
Westmoreland	368,000	4,350	4,402	4,529	4,321	4,260	4,401	**26,263**
Sub Total	**2,404,500**	**27,863**	**28,377**	**28,824**	**28,953**	**28,138**	**28,879**	**171,034**
+Combined Core								
Greene	40,000	424	482	473	444	447	477	**2,747**
Indiana	89,000	851	894	913	903	892	922	**5,375**
Lawrence	94,000	1,173	1,171	1,224	1,167	1,150	1,139	**7,024**
Grand Total	**2,627,500**	**30,311**	**30,924**	**31,434**	**31,467**	**30,627**	**31,417**	**186,180**

3.5 Inventory and Assessment of Morbidity Datasets

3.5.1 Hospital Admissions

Pennsylvania Health Care Cost Containment Council Hospital Discharge Data Sets (1999-2004)

On June 6, 2005, the PITT-PM health outcomes subgroup interviewed officials from the Pennsylvania Health Care Cost Containment Council (PHC4) Special Requests Unit concerning statewide hospital admission/discharge data collected by the agency. The PHC4 (http://www.phc4.org/) is an independent state agency formed under Pennsylvania statute (Act 89, as amended by Act 14) in order to address rapidly growing health care costs. Act 89, as amended by Act 14, specifically assigns the Council three primary responsibilities:

- Collect, analyze and make available to the public data about the cost and quality of health care in Pennsylvania,
- Study, upon request, the issue of access to care for those Pennsylvanians who are uninsured,
- Review and make recommendations about proposed or existing mandated health insurance benefits upon request of the legislative or executive branches of the Commonwealth.

The Council collects over 3.8 million inpatient hospital discharge and ambulatory/outpatient procedure records each year on individuals of all ages from hospitals and freestanding ambulatory surgery centers in Pennsylvania. These data, which includes hospital charge and treatment information as well as other financial data, are collected on a quarterly basis and then manually verified currently by PHC4 staff. PHC4 edits the data and provides error reports to each data source. The health care facility will make error corrections and provide PHC4 with corrected information. The data are processed using a series of validation rules before being finalized and made available for further analysis and public release. Compliance across health care institutions in Pennsylvania approaches 100% (99% in recently released 2006 reports). The Council also collects data from managed care plans on a voluntary basis. The Council shares these data with the public through free public reports. These reports are widely distributed, and can be found on the Council's Web site, http://www.phc4.org.

The Council also produces standardized and customized reports and data sets through its *Special Requests* division for a wide variety of users including hospitals, policy-makers, researchers, physicians, insurers, and other group purchasers. The standardized data sets do not have individual identifiers (e.g. name, social secutirty numbers, street address, etc.) and do not contain date of admission or date of discharge in the individual records. Only year and quarter of admission are presented (see **Appendix H** for typical PHC4 dataset layout). Zip Codes are available only in the 5-digit rather than 9-digit format, limiting ability to geocode health effects to a specific location within a certain ZIP code. These standard datasets can be obtained on a regional basis for ~$625 but are not particularly useful for time series studies of air quality and health since date of admission is not provided.

Researchers can request customized data sets to include dates of admission and discharge for linkage of health effects to air pollutant levels on a specific day. ZIP code is provided; however, individual street addresses for more precise geocoding are not currently acquired during the PHC4 data collection process. These customized data sets are requested through the Special Requests Unit. The application is available online at the PHC4 website. A $75.00 non-refundable processing fee is required at the time of submission of the application. The programmer/ analyst time is billed at $75.00/hour and usual programming costs range from $350.00 to $450.00 for a custom dataset. Additionally, the PHC4 charges $0.0025 per record for each individual record included. For example, the total cost for a custom data set with 6 years of hospital admission data for Allegheny County (1.26 million total admissions for the period) would be approximately $3,500. Requests for custom datasets with identifiers (e.g. date of admission and /or discharge) require approval of both the Unit supervisor (s) and the Council Executive Committee. The Executive Committee meets every two months; therefore turn-around time for requests for customized data with identifiers can be 3-4 months. Application for protected access to a custom data set by non-commercial entities is made through the Special Requests Unit and is available at http://www.phc4.org/services/datarequests/docs/specialreq_otherdatarequest.pdf.

3.5.2 Ohio Department of Health and West Virginia Healthcare Authority Hospital Discharge Datasets

Contact was made with the Ohio Department of Health (ODH) and the West Virginia Healthcare Authority (WVHA) related to the availability of comprehensive hospital admission data for these states from 1999-2004. Both agencies maintain comprehensive hospital admission databases for the time period of interest comparable to that archived by the PHC4. For an appropriate research application with documentation of Institutional Review Board Approval, these data are accessible to study investigators.

Hospital discharge data collection in Ohio was initiated in 1986 by the Office of Health Policy and Planning, within the Ohio Department of Health. In 1987, legislation was passed establishing a hospital data collection system and making submission of hospital discharge data mandatory for all Ohio licensed hospitals. Data include aggregate hospital-level discharge data for all hospitals, acute and specialty that are licensed in the state of Ohio. ODH typically requires academic or industry investigators to partner with an Ohio hospital for data access to individual level data. Ohio University has obtained hospitalization data previously for its work on air quality modeling in the Upper Ohio River Valley.

The West Virginia Healthcare Authority (WVHCA) has collected patient level data for all licensed WVA hospitals since 1985. Specific data variables collected include: hospital, patient age, gender, type of admission, source of admission, length of stay, discharge status, ZIP code and county of residence, marital status, procedures performed, DRG code, charges, physician, physician specialty, payer category and up to five diagnosis codes. The database does not include patient social security numbers as unique patient identifiers; however, date of birth, gender, and ZIP code are used to match patient files. No data tapes are publicly available, though aggregate information is available to managed care companies, insurers, and consultants. Special runs also may be requested by researchers, similarly to PHC4. All data cells with fewer than ten cases are suppressed and requests for protected elements are generally limited to specific variables.

3.5.3 Additional Data Acquisition/Abstraction through Individual Hospitals

Although the PHC4 data are relatively complete, some data elements that would enhance a retrospective study of $PM_{2.5}$ and health are not actively collected by the agency. For example, street address is not provided; geocoding of the home residence of subjects admitted to local hospitals would be limited to 5-digit ZIP code as the sole address identifier. Individual level data, such as education, occupation, smoking status, etc. that might be important for control of confounding in long-term studies are not included in the PHC4 dataset. Certain data elements, including street address, could be acquired from retrospective electronic or hard copy medical records from individual hospitals, but the process would be long, tedious and costly. In addition, IRB approvals from separate hospitals or hospital systems would be required. HIPAA regulations would most probably necessitate the participation of a third party to act as the honest broker for individually identifiable patient data. Although this method of data gathering is physically possible, the study team does not recommend this approach to paper-copy data collection/abstraction for a retrospective study due to cost and time,

3.5.4 Estimation of "Hospitalization Density" for the Pittsburgh Region using PHC4 Data

The PHC4 datasets remain the most comprehensive and readily available source of hospital admissions in Pennsylvania. As such, we employed this dataset to investigate more completely the number of hospital admissions in a 16-county western Pennsylvania region by patient's county of residence and hospital of admission from 1999-2004 to estimate the "hospital admission density" (e.g., potential sample size with admissions as the health outcome of interest) for the region. These data were assembled from county profiles of inpatient utilization provided online at http://www.phc4.org/countyprofiles/ county-wide by PHC4.

Figure 52 shows the population density (count/sq. mile) in the counties in the Pittsburgh combined core base statistical area and surrounding counties in Pennsylvania, Ohio and West Virginia. Also shown are the locations of area hospitals and the respective hospital admissions (counts) for the 1999-2004 period of interest in relation to the $PM_{2.5}$ mass and speciation monitoring sites. It is not surprising that the majority

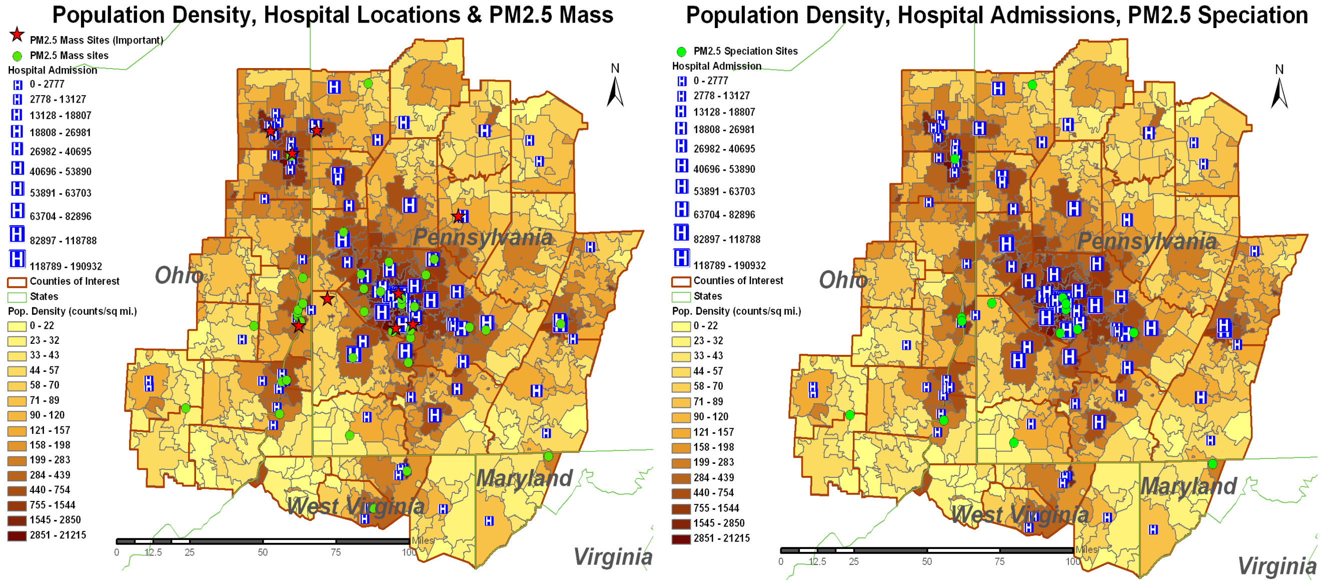


Figure 52: Population density and hospital admissions in the Pittsburgh Combined Core Statistical Area and in selected surrounding counties in Ohio and West Virginia (1999-2004). $PM_{2.5}$ mass sites (left) and speciation sites (right) are shown in green.

of hospitals as well as $PM_{2.5}$ monitoring sites are located within the most heavily populated regional urban area (Pittsburgh and its major traffic arteries). In addition, the hospitals with the highest number of admissions are located in the central, urban area of the Pittsburgh MSA. The rural areas (depicted with yellow to light brown shading) are often served by one or two community hospitals. Most hospitals, even those in more rural counties, are located within more heavily populated areas.

In **Table 27**, the PHC4 designated Region 1 is roughly representative of the Pittsburgh MSA including Greene County. Region 2 includes counties to the north of the Pittsburgh MSA and Region 3 represents counties to the east. From 1999-2004, a total of 3.06 million hospital admissions were recorded. A total of 78% of the hospital admissions (N = 2.38 million) occurred among residents in Region 1 of the PHC4 reporting areas (Allegheny, Armstrong, Beaver, Butler, Fayette, Green, Washington and Westmoreland counties). In Allegheny County alone, approximately 1.26 million admissions among county residents were observed (41% of the 16-county total).

Table 27: Total hospital admissions in 16 Western Pennsylvania counties 1999-2004.

County Name:	Total Hospital Admissions 2004	Total Hospital Admissions 1999	Total Hospital Admissions 1999-2004	Estimated Yearly Average by County
Region 1				
Allegheny	208,346	208,331	1,259,637	209,940
Armstrong	10,771	9,726	61,057	10,176
Beaver	27,547	27,305	162,866	27,144
Butler	26,775	22,882	148,145	24,691
Fayette	25,247	26,397	156,018	26,003
Greene	4,923	4,770	29,208	4,868
Washington	35,214	32,770	203,288	33,881
Westmoreland	61,465	59,987	362,373	60,396
Region 2				
Clarion	6,092	5,669	36,345	6,058
Jefferson	7,112	7,250	43,854	7,309
Lawrence	19,359	17,787	111,976	18,663
Mercer	20,186	20,107	121,870	20,312
Venango	9,191	9,031	54,926	9,154

County Name:	Total Hospital Admissions 2004	Total Hospital Admissions 1999	Total Hospital Admissions 1999-2004	Estimated Yearly Average by County
Region 3				
Cambria	26,700	25,805	159,330	26,555
Indiana	13,106	11,862	74,518	12,420
Somerset	11,709	11,755	71,184	11,864
Total Admissions	513,743	501,434	3,056,595	507,589

Ninety-four (94) hospitals were identified in the previously described 16-county area of western Pennsylvania (**Table 28**). A total of 27 institutions reported at least 50,000 admissions over the 6-year period and accounted for 75% of the total admissions (N = 2.26 million). Typically, air pollution studies focus on exacerbation of circulatory or respiratory disorders as the outcome of interest. Circulatory or respiratory admissions (ICD 9 codes 390-519) represented approximately 30% of all admissions. In the 8 county PHC4 Region 1, approximately 709,000 circulatory or respiratory admissions were reported from 1999-2004. In Allegheny County alone, a total of 357,000 circulatory or respiratory-related admissions were observed. These observations suggest that the density of hospital admissions in the Pittsburgh region from 1999-2004 will support an epidemiological study with circulatory and/or respiratory disease as the outcome of interest.

Table 28: Hospital admissions by hospital and patient county of residence in the 16 counties of the Western Pennsylvania region (1999-2004).

NAME	Allegheny	Armstrong	Beaver	Butler	Fayette	Green	Washington	Westmoreland	Clarion	Jefferson	Lawrence	Mercer	Venango	Cambria	Indiana	Somerset	TOTAL ADMIS-SIONS
ALIQUIPPA COMMUNITY HOSPITAL	890		17445				102				134						18571
ALLE KISKI MEDICAL CENTER	21363	9513						28156	74						164		59270
ALLEGHENY GENERAL HOSPITAL	91203	3344	7019	9553	3905	684	8011	7738	1769	1281	3999	3669	728	1103	3910	1946	149862
ALTOONA HOSPITAL	210			9523				52		171				11864	253	102	22175
ARMSTRONG COUNTY MEMORIAL HOSPITAL	303	27327		3966				1760	1849	197					862		36264
BON SECOURS HOLY FAMILY REGIONAL HEALTH CENTER														3457	79		3536
BROOKVILLE HOSPITAL	63	289							3118	9549			54		54		13127

NAME	Allegheny	Armstrong	Beaver	Butler	Fayette	Green	Washington	Westmoreland	Clarion	Jefferson	Lawrence	Mercer	Venango	Cambria	Indiana	Somerset	TOTAL ADMIS-SIONS
BROWNSVILLE GENERAL HOSPITAL INC					9484	702	2326	84									12596
BUTLER MEMORIAL HOSPITAL	1257	2390	528	60809			75	322	1401	67	1097	655	381		99		69081
CANONSBURG GENERAL HOSPITAL	1845				128	187	20550										22710
CHILDREN'S HOME OF PITTSBURGH	457						58	66									581
CHILDREN'S HOSPITAL OF PITTSBURGH	32736	1420	2565	3633	2260	425	3848	5972	478	556	1464	1286	520	1265	1084	763	60275
CHILDREN'S INSTITUTE OF PITTSBURGH	438			52			60	112									662
CITIZEN'S HOSPITAL	1492	763		103				6331									8689
CLARION HOSPITAL	64	873		120					17582	1845			1225	86			21795
CLARION PSYCHIATRIC CENTER	59	355		530				113	1335	380	105	177	936		478		4468
CLEARFIELD HOSPITAL										67							67
CONEMAUGH MEMORIAL MEDICAL CENTER	151				108			2233		64				77501	2469	18347	100873
CORRY MEMORIAL HOSPITAL													66				66
DUBOIS REGIONAL CENTER		53							641	14237			52		366		15349
ELLWOOD CITY HOSPITAL	59		2794	1960							12365						17178
ELK REGIONAL HOSPITAL										167							167
FORBES REGIONAL HOSPITAL	58721	578	139	270	343		200	25245						74	617	56	86243
FRICK HOSPITAL	95				13030		153	19767								73	33118
GEISENGER MEDICAL DANVILLE										190				103			293
GREENE COUNTY MEMORIAL HOSPITAL					334	11781	348										12463
HAMOT MEDICAL CENTER	140			53					246	120		558	848				1965
HEALTH SOUTH REHAB HOSPITAL ALTOONA														1192	59		1251
HEALTH SOUTH REHAB HOSPITAL ERIE												105	78				183
HEALTH SOUTH REHAB HOSPITAL HAMARVILLE	6577	1072	210	1559	182		271	2345	108	73	228	127	56	51	223		13082
HEALTH SOUTH REHAB HOSPITAL PITTSBURGH	5264				169		59	1950							111		7553
HEALTH SOUTH REHAB HOSPITAL SEWICKLEY	998		1922	83			157				153						3313
HIGHLANDS HOSPITAL					14451	80	88	832								74	15525

NAME	Allegheny	Armstrong	Beaver	Butler	Fayette	Green	Washington	Westmoreland	Clarion	Jefferson	Lawrence	Mercer	Venango	Cambria	Indiana	Somerset	TOTAL ADMIS-SIONS
INDIANA REGIONAL MEDICAL CENTER		2080						733	90	343				1598	39413		44257
JAMESON MEMORIAL HOSPITAL	69		476	954							55867	718					58084
JEFFERSON REGIONAL MEDICAL CENTER	78907		53	79	5077	177	7725	2979									94997
KINDRED HOSPITAL OF HERITAGE VALLEY			206														206
KINDRED HOSPITAL OF PITTSBURGH	882		208				161										1251
LAKEWOOD PSYCHIATRIC HOSPITAL	72						73										145
LATROBE AREA HOSPITAL	270	187			969		89	56385						116	10345	213	68574
LIFECARE HOSPITALS OF PITTSBURGH	5091	63		174	88		55	661				56					6188
MAGEE WOMENS HOSP OF THE UPMC HEALTH SYS	86205	865	1995	6639	2222	274	5094	8058	213	189	777	662	258	448	430	257	114586
MEADOWS PSYCHIATRIC CENTER														350		87	437
MEADVILLE MEDICAL CENTER	195								92			1144	1153				2584
MEDICAL CENTER BEAVER PA	2109		83191	698			186				4413	64					90661
MERCY HOSPITAL OF PITTSBURGH	94521	425	1540	2256	4756	543	5977	3974	171	181	1328	320	115	277	637	223	117244
MERCY JEANNETTE HOSPITAL	1210				795		122	33610							246		35983
MERCY PROVIDENCE (NORTH SHORE)	17024		155	67			354	113									17713
MEYERSDALE COMMUNITY HOSPITAL																2777	2777
MILTON HERSHEY MEDICAL CENTER	62													111			173
MINER'S HOSPITAL OF NORTHERN CAMBRIA														6606	1000		7606
MONONGAHELA VALLEY HOSPITAL	1548				15607	668	38580	10899									67302
MONSOUR MEDICAL CENTER	736	53			789	64	226	9051						108	178	104	11309
MOUNT NITTANY MEDICAL CENTER														112			112
NASON HOSPITAL														231			231
NORTHWEST MEDICAL CENTER UPMC	62	281		811					3154	115	52	779	35441				40695
OHIO VALLEY GENERAL HOSPITAL	24401		615	144	76		1218	96									26550

NAME	Allegheny	Armstrong	Beaver	Butler	Fayette	Green	Washington	Westmoreland	Clarion	Jefferson	Lawrence	Mercer	Venango	Cambria	Indiana	Somerset	TOTAL ADMISSIONS
PUNXSUTAWNEY AREA HOSPITAL		150							170	9135					2524		11979
SELECT SPECIALTY HOSPITAL GREENSBURG					154			872									1026
SELECT SPECIALTY HOSPITAL JOHNSTOWN														1324	62	490	1876
SELECT SPECIALTY HOSPITAL PITTSBURGH	1652			71	59		66	58									1906
SEMPER CARE HOSPITAL UPMC	53																53
SEWICKLEY VALLEY HOSPITAL	31046		26750	1138	52		852	146			363					124	60471
SHARON REGIONAL HEALTH SYSTEM	90		178	341							2936	49717	110				53372
SOMERSET COMMUNITY HOSPITAL	92				264			236						163		26226	26981
SOUTHWOOD PSYCHIATRIC HOSPITAL	1934		690	159	609	366	1420	184			63						5425
ST CLAIR MEMORIAL HOSPITAL	76340		185	167	288	218	12004	238									89440
ST FRANCIS HOSPITAL OF NEW CASTLE	92		152								14691	145					15080
ST FRANCIS MEDICAL CENTER	43323	478	1137	2228	580	107	882	3056	345	236	1123	81		111	134	69	53890
ST. FRANCIS CENTRAL HOSPITAL	4064		149	331	135		244	96			341				70		5430
ST. FRANCIS HOSPITAL CRANBERRY	616		259	1545													2420
ST. VINCENT HEALTH CENTER	92			93					302	109		1122	3091				4809
SUBURBAN GENERAL HOSPITAL	22787		228	359			56										23430
TITUSVILLE AREA HOSPITAL									75				4219				4294
TYRONE HOSPITAL														83			83
UNIONTOWN HOSPITAL	102				57498	1777	645	373								244	60639
UNITED COMMUNITY HOSPITAL	62	109		5306					164		976	11433	757				18807
UPMC BEDFORD																84	84
UPMC BRADDOCK	36643	66	313	58	192		181	937							108		38498
UPMC HORIZON	153		59	556					117	63	2477	41286	624				45335
UPMC LEE REGIONAL	51							3340						41092	1829	6619	52931
UPMC MCKEESPORT HOSPITAL	53178				123		261	2693									56255
UPMC PASSAVANT	42999	79	1435	15384	57		134	275			420	130					60913
UPMC PASSAVANT CRANBERRY	166		138	861													1165

NAME	Allegheny	Armstrong	Beaver	Butler	Fayette	Green	Washington	Westmoreland	Clarion	Jefferson	Lawrence	Mercer	Venango	Cambria	Indiana	Somerset	TOTAL ADMIS-SIONS
UPMC PRESBYTERIAN	115985	3019	5223	5928	8453	1039	8295	18433	1512	1976	3730	5582	2852	3696	3413	1796	190932
UPMC REHABILITATION HOSPITAL	8385	51	174	154	290		325	617	54	59	114	130	78	70	90		10591
UPMC SHADYSIDE	68906	982	821	1369	1965	300	2132	6811	249	911	572	483	258	601	705	313	87378
UPMC SOUTH SIDE	31856		108	105	102		279	258									32708
UPMC ST MARGARET	52093	1085	303	1876	279		451	6665	116	142	259	104		82	156	92	63703
WARREN GENERAL HOSPITAL													53				53
WASHINGTON HOSPITAL	1674		99	58	1900	9012	75260	292									88295
WESTERN PENNSYLVANIA HOSPITAL	93647	2054	1639	4354	2033	227	1992	9039	284	676	781	366	138	606	491	461	118788
WESTMORELAND REGIONAL HOSPITAL	1031	76			5085		237	75446						59	801	161	82896
WINDBER HOSPITAL								52						3554		8612	12218
Facilities with < 50 admissions	1590	702	791	689	569	512	588	1015	559	755	875	751	659	1061	908	779	12803
Total Admissions by County 1999-2004 (Column Total)	1228451	60782	161892	147166	155460	29143	202470	360769	36268	43854	111703	121650	54750	159155	74368	71092	3018973

Note #1: Speciality Hospitals with less than 50 admission per year and small rehab hospitals are not presented separately
Note #2: Circulatory plus respiratory admissions (ICD9 390-519) accounted for approximately 30% of all admissions
Note #3: A total of 78% of the hospital admissions (N= 2,346,133) in the 16 counties from 1999-2004 occurred in PHC4 Region I (8 counties)
Note #4: A total of 27 facilities reported admissions of at least 50,000 over the the 6 year period from 1999-2004 accounting for 2,256,955 total admissions (75%)
Note #5: A total of 65%% of the hospital admissions (N = 1.95 million) in the 16 county area occur in 4 counties (Allegheny, Beaver, Washington, Westmoreland)

3.5.5 Utilization of the PHC4 Hospitalization Data in Retrospective Studies

PITT-PM investigators consider the PHC4 data source a key low-cost, relatively comprehensive resource for investigating retrospectively the association between air quality and health. As a health outcome, hospital admissions are more sensitive to daily changes in air pollution than mortality. Typically, air pollution studies focus on exacerbation of circulatory or respiratory disorders as the outcome of interest. As such, in November 2005 we obtained as test data the Pennsylvania Health Care Cost Containment hospital admission files for l999 through 2004 for individuals residing in Allegheny County with a primary hospital discharge diagnosis of all circulatory (ICD-9 codes 390-459) or respiratory (ICD-9 codes 460-519) conditions. In addition, hospital admission data were obtained for two potential “control” conditions less likely to be related to daily air quality, namely hospitalizations for fractures (ICD-9 800-829) and hospitalization with E-codes denoting motor vehicle accidents (E810-819). A series of descriptive analyses were conducted and tables were generated to consider both the quantity of cardiopulmonary admission and control disease data available by age and gender as well as the distribution by time (month, year, day of week) and specific diagnosis. These analyses are described in the following tables.

3.6 Analysis of the PHC4 Allegheny County Cardiopulmonary Hospital Admissions Dataset (1999-2004)

A total of 346,424 admissions among Allegheny County residents for the period 1999 through 2004 with a primary discharge diagnosis of circulatory or respiratory disease were observed. Circulatory or respiratory admissions represented approximately 28% of all hospital admissions.

3.6.1 Circulatory and Respiratory Hospital Admissions by Year and Month

Tables 29 and 30 show the distribution of hospital admissions for discharge diagnoses of respiratory system disease (ICD 460 to 519) and for circulatory system disease (ICD 390-459) by year and month. From l999 to 2004, a total of 113,553 hospital admissions for respiratory system disease occurred among Allegheny County residents of all ages (**Table 29**). It can be readily observed that a higher proportion of respiratory system admissions occur during December, January, February and March of each year, reflecting a well-documented seasonal trend. During the l999 to 2004 time period, there were a total of 232,871 admissions for circulatory system disease (**Table 30**). A seasonal pattern was not observed for circulatory system diseases. For both respiratory and circulatory diseases, a decreased number of admissions was observed for December 2004, most probably related to under-reporting of the last period for which data were requested.

Table 29: Hospital admissions for respiratory system diseases by year and month of admission to Allegheny County residents, 1999-2004. First admissions and readmissions with primary discharge diagnosis ICD9 460-519.

	1999-2004		**1999**		**2000**		**2001**		**2002**		**2003**		**2004**	
	No.	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
January	12505	11.0	2268	11.1	2619	13.7	2031	10.7	2008	10.9	1581	8.3	1998	11.3
February	11835	10.4	2757	13.5	1694	8.9	1987	10.5	2158	11.8	1604	8.4	1635	9.2
March	11516	10.1	2401	11.8	1698	8.9	1892	10.0	1974	10.8	1820	9.6	1731	9.8
April	9447	8.3	1610	7.9	1554	8.1	1652	8.7	1610	8.8	1519	8.0	1502	8.5
May	8539	7.5	1407	6.9	1446	7.6	1568	8.3	1313	7.2	1492	7.9	1313	7.4
June	8042	7.1	1242	6.1	1399	7.3	1421	7.5	1198	6.5	1472	7.8	1310	7.4
July	7156	6.3	1068	5.2	1235	6.5	1234	6.5	1153	6.3	1192	6.3	1274	7.2
August	7191	6.3	1118	5.5	1228	6.4	1202	6.3	1209	6.6	1204	6.3	1230	7.0
September	8281	7.3	1379	6.8	1565	8.2	1316	6.9	1284	7.0	1402	7.4	1335	7.5
October	9072	8.0	1492	7.3	1521	8.0	1571	8.3	1450	7.9	1508	7.9	1530	8.7
November	8983	7.9	1441	7.1	1504	7.9	1547	8.1	1476	8.0	1533	8.1	1482	8.4
December	10986	9.7	2215	10.9	1664	8.7	1573	8.3	1522	8.3	2665	14.0	1347	7.6
All months	113553	100.0	20398	100.0	19127	100.0	18994	100.0	18355	100.0	18992	100.0	17687	100.0

Table 30: Hospital admissions for circulatory system diseases by year and month of admission. First admissions and readmissions with primary discharge diagnosis ICD9 390-459. Data Source: Pennsylvania Health Care Cost Containment Council.

	1999-2004		**1999**		**2000**		**2001**		**2002**		**2003**		**2004**	
	No.	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
January	20574	8.8	3452	8.7	3483	8.7	3795	9.4	3390	8.7	3318	8.9	3136	8.6
February	18936	8.1	3262	8.2	3370	8.4	3217	8.0	3111	8.0	2872	7.7	3104	8.5
March	20952	9.0	3753	9.4	3499	8.7	3651	9.1	3404	8.8	3241	8.7	3404	9.4
April	19836	8.5	3477	8.7	3319	8.2	3328	8.3	3470	8.9	3156	8.4	3086	8.5
May	19861	8.5	3303	8.3	3535	8.8	3467	8.6	3319	8.5	3284	8.8	2953	8.1
June	18969	8.1	3348	8.4	3209	8.0	3265	8.1	3154	8.1	3024	8.1	2969	8.2
July	19033	8.2	3226	8.1	3240	8.0	3184	7.9	3264	8.4	3115	8.3	3004	8.3
August	19057	8.2	3086	7.8	3316	8.2	3377	8.4	3208	8.2	2971	7.9	3099	8.5
September	18500	7.9	3066	7.7	3092	7.7	3135	7.8	3038	7.8	3105	8.3	3064	8.4
October	19897	8.5	3330	8.4	3478	8.6	3346	8.3	3316	8.5	3294	8.8	3133	8.6
November	18896	8.1	3187	8.0	3416	8.5	3294	8.2	3115	8.0	2948	7.9	2936	8.1
December	18360	7.9	3263	8.2	3298	8.2	3175	7.9	3099	8.0	3077	8.2	2448	6.7
All months	232871	100.0	39753	100.0	40255	100.0	40234	100.0	38888	100.0	37405	100.0	36336	100.0

3.6.2 Average Daily Totals of Circulatory and Respiratory Hospital Admissions for Allegheny County

In **Tables 31 and 32**, the daily totals are given by specific disease classification for circulatory and respiratory hospital admissions for Allegheny County from 1999-2004. Average daily admissions for all circulatory or respiratory diseases respectively were 53.8 (range 23-87) and 51.8 (range 20-167) per day. On average, the highest daily admissions for circulatory diseases were attributed to "other heart disease" (21.7/day), ischemic heart disease (13.3/day), and cerebrovascular disease (10.0/dy); for respiratory causes, the highest daily admissions were for pneumonia (17.6/day), chronic bronchitis (11.1/day), "other respiratory diseases" (7.5/day) and asthma (6.5/day).

Table 31: Daily hospital admissions for circulatory diseases (ICD 390-459) Allegheny County, 1999-2004: descriptive statistics.*

	ICD	Mean	Std. Deviation	Minimum	Maximum
Hypertensive Disease	401-405	2.4	1.6	0	10
Ischemic Heart Disease	**410-414**	**13.3**	**3.9**	**3**	**29**
Other Heart Disease	420-429	21.7	5.7	3	43
Cerebrovascular Disease	430-438	10.0	3.3	1	24
Atherosclerosis	440	0.4	0.6	0	4
Aortic Aneurysm	441	0.3	0.6	0	4
Other Cardiovascular	various	5.7	2.7	0	17
Total	390-459	53.8	9.7	23	87

*Data Source: Pennsylvania Health Care Containment Council

Table 32: Daily hospital admissions for respiratory diseases (ICD 460-519) Allegheny County, 1999-2004: descriptive statistics. Data Source: Pennsylvania Health Care Containment Council.

	ICD	Mean	Std. Deviation	Minimum	Maximum
Acute Bronchitis	466	2.0	2.4	0	20
Other Acute Respiratory	460-465	0.8	1.0	0	6
Other Upper Respiratory	470-478	0.9	1.0	0	6
Pneumonia	480-486	17.6	6.9	4	65
Influenza	487	0.2	0.8	0	10
Chronic Bronchitis	490-491	11.1	4.6	1	41
Emphysema	492	0.3	0.5	0	3
Asthma	**493**	**6.5**	**3.4**	**0**	**23**
Other COPD	494-496	0.6	0.8	0	6
Lung Disease-external agents	500-508	4.4	2.2	0	14
Other Respiratory	510-519	7.4	3.2	0	22
Total	460-519	51.8	16.5	20	167

3.6.3 Average Annual Circulatory and Respiratory Admissions/Admission Rates

Tables 33 and 34 show the average annual number of hospital admissions and admission rates for respiratory and circulatory system diseases by age and gender. Admission rates were calculated based on the age group-specific population numbers for Allegheny County from the 2000 census.

Table 33: Hospital admissions for respiratory system diseases for Allegheny County residents. 1999-2004 average annual number of admissions and admission rate. First admissions and readmissions with primary discharge diagnosis ICD9 460-519.*

	Total		**Male**		**Female**	
Age	**Avg. Ann. #**	**Rate***	**Avg. Ann. #**	**Rate**	**Avg. Ann. #**	**Rate**
<5	1170	164.5	717	197.1	453	130.5
5-24	1020	31.9	575	35.3	445	28.4
25-44	1431	39.4	574	32.3	857	46.1
45-64	3578	119.5	1529	108.1	2049	129.8
65-84	8927	445.7	3921	481.2	5006	421.4
85+	2800	995.0	984	1275.9	1817	888.9
Total	18926	147.7	8298	136.7	10627	157.5

*per 10,000 population 2000 census

Admission rates for respiratory diseases were 164.5/10,000 for children under 5 years, decrease through age 64 years, and increase to 445.7/10,000 in the 65-84 age group, and 995.0/10,000 in the 85+ group (**Table 33**). Admission rates for respiratory disease appear to be greater among males until age 25 when the rate among women increases. Starting at age 65 the male rate again is greater than the female rate. The influence of cigarette smoking in the various cohorts and occupational status may influence these sex ratios. Admission rates for circulatory system diseases (**Table 34**) are very low for children and young adults under age 24 (~5/10,000) and increase dramatically with age, the highest rates in the elderly with 1126.1/10,000 in the 65-84 year age group and 2128.7/10,000 in the 85+ group. Men have consistently higher admission rates for circulatory system diseases than females in all age categories.

Table 34: Hospital admissions for circulatory diseases for Allegheny County residents. Average annual number of admissions and admission rate. First admissions and readmissions with primary discharge diagnosis ICD9 390-459. Data Source: Pennsylvania Health Care Cost Containment Council.*

	Total		**Male**		**Female**	
Age	**Avg. Ann. #**	**Rate***	**Avg. Ann. #**	**Rate**	**Avg. Ann #**	**Rate**
<5	35	4.9	21	5.6	14	4.1
5-24	149	4.7	79	4.8	71	4.5
25-44	1605	44.2	953	53.7	651	35.0
45-64	8480	283.3	5095	360.4	3385	214.3

	Total		**Male**		**Female**	
Age	**Avg. Ann. #**	**Rate***	**Avg. Ann. #**	**Rate**	**Avg. Ann #**	**Rate**
65-84	22552	1126.1	10716	1315.0	11836	996.4
85+	5991	2128.7	1843	2391.2	4148	2029.6
Total	38811	302.8	18706	308.2	20105	298.0

*per 10,000 population 2000 census

3.6.4 Respiratory Disease Subgroups by Age

Table 35 presents the distribution of hospital admissions for respiratory diseases by subcategory and age group for Allegheny County residents for the period 1999 through 2004. Disease rubrics consisting of pneumonia (480-486), chronic bronchitis (490-491) and asthma (493) had the greatest number of admissions. The 65-84 year age group represented 48% of pneumonia hospitalizations and 64% of admissions for chronic bronchitis. Children under the age of five years accounted for 13% of all asthma related hospitalizations. **Table 36** shows the average annual age specific admission/re-admission rates (per 10,000 population based on the 2000 census for Allegheny County) by category of respiratory disease (primary discharge diagnosis). For children under the age of five, asthma and acute bronchitis had the highest rates of hospital admission followed by pneumonia. For the age group 5 to 24, asthma remained the leading disease condition for admission. Likewise, during the adult years of 25 through 64, asthma, pneumonia and chronic bronchitis had the highest rates of primary disease admission for this time period. In the 85 and over age category, pneumonia had the highest rate of respiratory disease hospital admission (462.6 per 10,000 population.)

Table 35: Hospital admissions for respiratory system diseases by disease category and age for Allegheny County residents, 1999-2004. Total number of admissions - 6 years. (Note: First admissions and readmissions with primary discharge diagnosis ICD9 460-519.)

	Age Group						
Disease Subgroup	**<5**	**5-24**	**25-44**	**45-64**	**65-84**	**85+**	**Total**
Acute bronchitis/bronchiolitis (466)	1959	62	306	481	1005	479	4292
Other acute respiratory (460-465)	604	362	301	233	249	57	1806
Other upper respiratory (470-478)	361	657	391	281	240	31	1961
Pneumonia (480-486)	1764	1419	2501	6364	18730	7812	38590
Influenza 487	54	40	62	87	175	73	491
Bronchitis, non acute (490-491)	10	39	540	5511	15607	2547	24254
Emphysema (492)	0	8	28	178	377	46	637
Asthma (493)	1903	2835	2779	3454	2764	562	14297
Other COPD (494-496)	0	4	23	218	830	155	1230
Lung disease/other respiratory	362	693	1652	4663	13586	5039	25995
All respiratory	7017	6119	8583	21470	53563	16801	113553

Table 36: Hospital admissions for respiratory system diseases by disease category and age to Allegheny County residents, 1999-2004. Average annual admission rate (per 10,000 population). First admissions and readmissions with primary discharge diagnosis ICD9 460-519.

	Age Group					
Disease Subgroup	**<5**	**5-24**	**25-44**	**45-64**	**65-84**	**85+**
Acute bronchitis/bronchiolitis (466)	45.9	0.3	1.4	2.7	8.4	28.4
Other acute respiratory (460-465)	14.2	1.9	1.4	1.3	2.1	3.4
Other upper respiratory (470-478)	8.5	3.4	1.8	1.6	2.0	1.8
Pneumonia (480-486)	41.4	7.4	11.5	35.4	155.9	462.6
Influenza 487	1.3	0.2	0.3	0.5	1.5	4.3
Bronchitis, non-acute (490-491)	0.2	0.2	2.5	30.7	129.9	150.8
Emphysema (492)	0.0	0.0	0.1	1.0	3.1	2.7
Asthma (493)	44.6	14.8	12.8	19.2	23.0	33.3
Other COPD (494-496)	0.0	0.0	0.1	1.2	6.9	9.2
Lung disease/other respiratory	8.5	3.6	7.6	26.0	113.1	298.4
All respiratory	164.5	31.9	39.4	119.5	445.7	995.0

Data Source: Pennsylvania Health Care Cost Containment Council

3.6.5 Circulatory Disease Subgroups by Age

Table 37 presents first admissions and readmissions with primary discharge diagnosis for diseases of the circulatory system (ICD9 390 - 459) among Allegheny County residents by age. A total of 232,871 admissions/readmissions occurred during this six year period. The most common discharge diagnoses were 1) other heart disease (35%), principally heart failure and cardiac dysrhythmias; 2) ischemic heart disease (29%), nearly all of which were acute myocardial infarction; and 3) cerebrovascular disease (17%). Nearly all (95.4%) of these admissions occurred among those over the age of 45 with 73.5% among those aged 65 and over. **Table 38** presents the average annual first admissions and readmission rates for circulatory conditions by age for the period of l999 through 2004 for Allegheny County. Not unexpectedly, although the numbers of admissions are greatest in the 65-84 age group, the rates (with a denominator attached) of circulatory disease admission are greatest in the 85 and older age group.

Table 37: Hospital admissions for circulatory system diseases by disease category and age to Allegheny County residents, 1999-2004. Total number of admissions - 6 years. First admissions and readmissions with primary discharge diagnosis ICD9 390-459.

	Age Group						
Disease Subgroup	**<5**	**5-24**	**25-44**	**45-64**	**65-84**	**85+**	**Total**
Hypertensive disease (401-405)	13	73	1043	2346	4299	1282	9056
Ischemic heart disease (410-414)	0	24	2335	20163	38656	6818	67996
Other heart disease (420-429)	84	326	2858	13374	48622	15976	81240
Cerebrovascular disease (430-438)	15	102	1093	7241	24151	6951	39553
Atherosclerosis (440)	1	3	73	1168	3432	642	5319

	Age Group						
Disease Subgroup	**<5**	**5-24**	**25-44**	**45-64**	**65-84**	**85+**	**Total**
Aortic aneurysm (441)	0	5	31	467	2256	279	3038
All other circulatory	96	363	2195	6120	13897	3998	26669
All Circulatory	209	896	9628	50879	135313	35946	232871

Table 38: Hospital admissions for circulatory system diseases by disease category and age to Allegheny county residents, 1999-2004. Average annual admission rate per 10,000 population. First admissions and readmissions with primary discharge diagnosis ICD9 390 - 459.

	Age Group						
Disease Subgroup	**<5**	**5-24**	**25-44**	**45-64**	**65-84**	**85+**	**Total**
Hypertensive disease (401-405)	0.3	0.4	4.8	13.1	35.8	75.9	11.8
Ischemic heart disease (410-414)	0.0	0.1	10.7	112.3	321.7	403.8	88.4
Other heart disease (420-429)	2.0	1.7	13.1	74.5	404.6	946.1	105.6
Cerebrovascular disease (430-438)	0.4	0.5	5.0	40.3	201.0	411.6	51.4
Atherosclerosis (440)	0.0	0.0	0.3	6.5	28.6	38.0	6.9
Aortic aneurysm (441)	0.0	0.0	0.1	2.6	18.8	16.5	4.0
All other circulatory	2.3	1.9	10.1	34.1	115.7	236.8	34.7
All Circulatory	4.9	4.7	44.2	283.3	1126.1	2128.8	302.8

Data Source: Pennsylvania Health Care Cost Containment Council

3.6.6 Day of the Week Effects and Source of Admission

We further examined day of the week effects in hospital admissions in relation to admission source (e.g,, emergent vs. non-emergent) for Allegheny County residents with a primary discharge diagnosis of either circulatory disease (ICD-9 390-459) or respiratory disease (ICD-9 460-519) for the time period July 2001-June 2002, a period for which the most complete data for air pollutants is available. During this period there were 39,359 hospital admissions for cardiovascular disease and 18,704 for respiratory disease **(Table 39**).

Table 39: Day-to-day variation in hospital admissions for circulatory system and respiratory disease among Allegheny County residents for July 2001-June 2002.

	Circulatory		**Respiratory**	
	All Admissions	**Admissions from ER**	**All Admissions**	**Admissions from ER**
Sunday	3686 (9.4%)	2743 (13.2%)	2279 (12.2%)	1792 (13.9%)
Monday	6724 (17.1%)	3176 (15.3%)	3044 (16.3%)	2033 (15.8%)
Tuesday	6701 (17.0%)	3144 (15.1%)	2863 (15.3%)	1863 (14.4%)
Wednesday	6382 (16.2%)	3042 (14.6%)	2735 (14.6%)	1805 (14.0%)
Thursday	6037 (15.3%)	2909 (14.0%)	2850 (15.2%)	1903 (14.8%)
Friday	6002 (15.2%)	3035 (14.6%)	2728 (14.6%)	1802 (14.0%)

	Circulatory		Respiratory	
	All Admissions	Admissions from ER	All Admissions	Admissions from ER
Saturday	3827 (9.7%)	2768 (13.3%)	2205 (11.8%)	1696 (13.2%)
TOTAL	39359 (100.0%)	20817 (100.0%)	18704 (100.0%)	12894 (100.0%)

A day of the week effect was apparent, with a higher number of admissions on weekdays, especially for circulatory diseases, suggesting that at least some of these admissions might be procedure driven and not necessarily mediated by the events of the day. Since a number of these admissions could be previously scheduled, we chose to limit this further analyses to those admitted from the Emergency Department.

Shown in **Figure 53** is a graph of circulatory and respiratory hospital admissions admitted from the ED by day for the year July 1, 2001-June 30, 2002. During this period, the mean daily number of admissions from the ED for cardiovascular diseases was 57 with a minimum of 35 and a maximum of 84. For respiratory diseases, the mean number of daily admissions was 35 with a minimum of 13 and a maximum of 70. A clear seasonal pattern with a higher number of admissions from January through March was evident for respiratory disease and suggested but much less prominent for circulatory disease. These exploratory analyses demonstrate that daily variability in hospital admissions for both circulatory and respiratory diseases is evident in the Pittsburgh region.

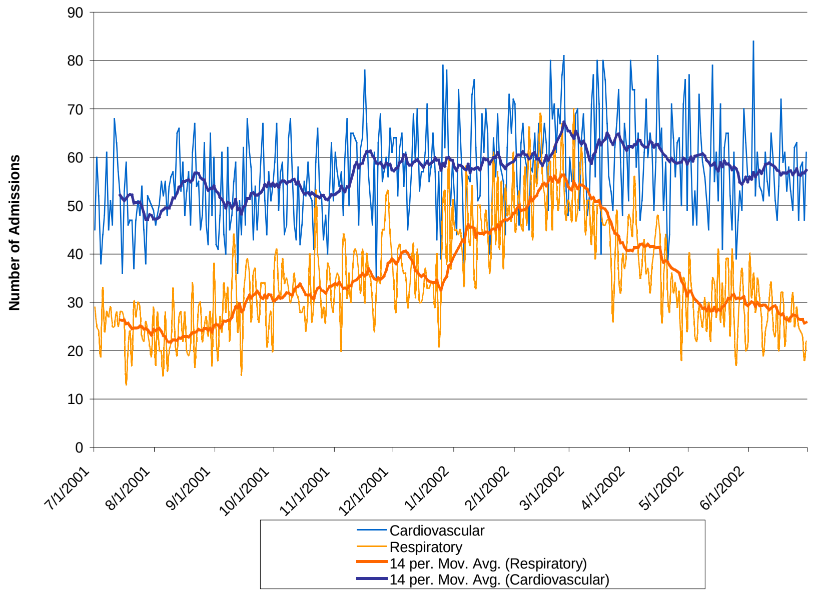


Figure 53: Hospital admissions from the emergency department for circulatory and respiratory disease among Allegheny County residents (July 2001-June 2002). Data source: Pennsylvania Cost Care Containment Council.

3.6.7 Correlation of Daily Hospital Admissions for Circulatory and Respiratory from the Emergency Department across Allegheny County Hospitals

To determine if the "ups-and-downs" of daily admissions from the ED for respiratory and circulatory diseases might track across all county hospitals, we conducted a correlation analysis using the Pearson correlation coefficient to assess the strength of the association.

In **Table 40** the correlation of daily respiratory hospital admissions across 11 of the largest area hospitals is presented. The correlation coefficients ranged from -0.035 for UPMC Presby/Shadyside and Children's Hospital to 0.280 for UPMC Presby/Shadyside and Jefferson Hospitals. Most of the coefficients were in the 0.100-0.130 range and were statistically significant, suggesting some tracking of daily respiratory admissions from EDs across area hospitals and a possible common source exposure (e.g. meteorological influences, particulates, other).

Table 40: Correlation matrix of daily hospital admissions from the ED for respiratory diseases (1999-2004).

Hospital	**Children's**	**Jefferson**	**St. Margaret**	**UPMC Braddock**	**UPMC Passavant**	**UPMC McKeesport**	**UPMC Presby/Shady**	**AGH**	**St. Clair**	**Mercy**
Forbes	.135**	.128**	.087**	.098**	.101**	.135**	.129**	.109**	.131**	.119**
Children's	**1**	.041	.070**	.106**	.092**	.151**	-.035	.195**	.084**	.192**
Jefferson		**1**	.162**	.120**	.151**	.118**	.280**	.084**	.146**	.222**
St. Margaret			**1**	.134**	.095**	.126**	.173**	.099**	.102**	.151**
UPMC Braddock				**1**	.065**	.161**	.145**	.120**	.081**	.154**
UPMC Passavant					**1**	.132**	.206**	.083**	.140**	.107**
UPMC McKeesport						**1**	.078**	.146**	.072**	.173**
UPMC Presby/Shady							**1**	.027	.223**	.114**
AGH								**1**	.079**	.124**
St. Clair									**1**	.115**

*** Correlation is significant at the 0.01 level*

Conversely, as seen in **Table 41**, daily circulatory admissions from the ED are much less correlated than respiratory admissions, with coefficients typically in the range of .03-.100 with fewer statistically significant associations. Although UPMC Presby/Shadyside circulatory admissions were significantly correlated with Jefferson Hospital (.281), St. Clair Hospital (0.139) and St. Margaret (0.111) admissions, most other coefficients were < 0.100. These results suggest that daily circulatory admissions for the ED across the area are less likely mediated by regional events than perhaps respiratory admissions.

Table 41: Correlation matrix of daily hospital admissions from the ED for circulatory diseases (hospitals with > 5000 circulatory emergency department admissions) (1999-2004).

Hospital	**Jefferson**	**St. Margaret**	**UPMC Passavant**	**UPMC McKeesport**	**UPMC Presby/ Shady**	**AGH**	**St. Clair**	**Mercy**
Forbes	.041	.075**	,035	.032	.063**	.054*	.062**	.041
Jefferson	**1**	.053*	.031	-.038	.281**	-.025	.061**	.102**
St. Margaret		**1**	.054*	.010	.111**	.055**	.020	.090**
UPMC Passavant			**1**	.009	.073**	.033	.104**	.039
UPMC McKeesport				**1**	-.023	.019	.003	.016
UPMC Presby/Shady					**1**	.051*	.139**	.089**
AGH						**1**	.034	.041
St. Clair							**1**	.010

*** Correlation is significant at the .01 level * correlation is significant at the .05 level*

3.7 Descriptive Analysis of Control Disease Hospitalizations (Fractures) for Allegheny County from 1999-2004

As noted previously, we also requested sample hospitalization data for two "control" disorders considered to be unrelated to air pollution, specifically fractures and motor vehicle injuries, for exploratory descriptive analyses. A summary of the analyses for fractures follows. A total of 34,447 fracture hospitalizations were reported in Allegheny County from 1999-2004 with an average of 15.7 per day for the time period of interest. Unlike respiratory admissions and in some respect circulatory admissions, fractures admissions from all sources (ED and non-ED) were relatively constant by year, month and day of the week **(Table 42)**.

Table 42: Hospital admissions for fractures (ICD-9 800-829) Allegheny County, January 1999–December 2004: distribution by year of admission, month of admission, day of week, age, and gender.*

	Number	Percentage
Year of Admission		
1999	5886	17.1
2000	5786	16.8
2001	5837	16.9
2002	5715	16.6
2003	5639	16.4
2004	5584	16.2
Month of Admission		
January	3096	9
February	2723	7.9
March	2655	7.7
April	2858	8.3
May	2962	8.6
June	2815	8.2
July	2910	8.4
August	2899	8.4
September	2842	8.3
October	2944	8.5
November	2743	8
December	3000	8.7

	Number	Percentage
Day of Week		
Sunday	4464	13
Monday	5105	14.8
Tuesday	5044	14.6
Wednesday	4970	14.4
Thursday	5046	14.6
Friday	5140	14.9
Saturday	4678	13.6
Age		
<5	321	0.9
5-24	3003	8.7
25-44	3959	11.5
45-64	5131	14.9
65-84	13945	40.5
85	8088	23.5
Gender		
Male	12478	36.2
Female	21969	63.8

**Data Source: Pennsylvania Health Care Containment Council*

3.7.1 Correlation of Daily Hospital Admissions for Fractures from the Emergency Department across Allegheny County Hospitals

In **Table 43**, the Pearson correlation coefficients are shown for 10 major hospitals in the Pittsburgh area for daily fracture admissions through the ED. The correlation coefficients ranged from .000 to .066. The highest correlation coefficient for fracture ED admissions (0.066) was for UPMC McKeesport and Mercy Hospital. The magnitude of these coefficients suggests that the number of daily fractures admissions through the ED is uncorrelated at Pittsburgh area hospitals. Fracture admissions show promise as potential control admissions for the retrospective study. Admissions for injuries and as well as gall bladder surgeries and appendectomies might also be explored as control admissions for modeling.

Table 43: Correlation matrix of daily hospital admissions from the ED for fractures (1999-2004).

Hospital	**Children's**	**Jefferson**	**St. Margaret**	**UPMC Passavant**	**UPMC McKeesport**	**UPMC Presby/ Shady**	**AGH**	**St. Clair**	**Mercy**
Forbes	-.017	-.031	.012	..012	.062**	.000	.020	-.021	.062**
Children's	**1**	.009	-.007	-.014	.027	.011	0.016	-.014	.026
Jefferson		**1**	.011	.004	.030	.020	.011	.014	.010
St. Margaret			**1**	-.003	.019	.038	.035	.011	.000
UPMC Passavant				**1**	.038	.023	-.003	.024	.023
UPMC McKeesport					**1**	.063**	.023	.038	.066**
UPMC Presby/Shady						**1**	.027	.025	.020
AGH							**1**	.001	.001
St. Clair								**1**	.053*

** Correlation is significant at the 0.01 level * Correlation is significant at the 0.05 level

3.7.2 Summary of Exploratory Analysis of Hospitalizations in Allegheny County, Pennsylvania and the Pittsburgh MSA

1) The "hospitalization density" for circulatory and respiratory disease in Allegheny County and the surrounding region for the 1999-2004 time-period of interest will support a retrospective epidemiological study of the relationship between $PM_{2.5}$ and hospital admissions.

2) As reported in the literature, seasonal and day of the week patterns are evident for respiratory disease admissions and less apparent for circulatory disease admissions in Allegheny County, Pennsylvania.

3) Highest average annual number of admissions for both respiratory and circulatory diseases occurs in the 65-85 age group. Children less than 5 years of age account for 13% of asthma-related admissions.

4) Daily respiratory admissions from EDs across Allegheny County hospitals appear to be correlated, suggesting a possible common source exposure (e.g. meteorological influences, particulates, etc).

5) Daily circulatory admissions for the ED across Allegheny County hospitals demonstrate little correlation and are less likely to be influenced by regional factors than respiratory admissions; however, some correlation for specific hospital systems is evident.

6) Fractures admissions demonstrated consistency in daily averages with little associations with year, month, or day of the week.

7) Fracture admissions are uncorrelated across most hospitals in the Pittsburgh area, are unlikely to be influenced by regional effect and represent a possible control hospitalization for the retrospective study.

3.7.3 Strengths and Weaknesses of Hospitalizations as a Health Endpoint for Retrospective Air Quality Studies

PHC4 data can be very useful in ecological, case-crossover and other studies for the evaluation of health effects related to air pollutants in the Pittsburgh region. For utilization and cost containment reasons, hospitals are required to submit reports of hospital admissions to the agency. Compliance in reporting, although not 100%, is usually 98% or higher in the state of Pennsylvania. In addition, hospitalization data, like mortality data, are more readily available retrospectively for an expanded area of interest, such as the Pittsburgh core base statistical area or larger area.

Inclusion of certain demographic information in the PHC4 dataset, including age, gender, race and 5-digit ZIP code of residence, allows for assessment of some potential confounders (long term studies) and crude geocoding for spatial resolution of outcomes. However, the PHC4 agency does not currently require reporting by hospital of ZIP+ 4 (9-digit ZIP code) for the home residence of admitted patients. ICD-9/10 coding of both primary and secondary diagnoses (a total of 8) can help to identify subpopulations with underlying cardiopulmonary co-morbidities. Also an assigned system ID allows linkage of initial and subsequent admissions on a specific subject for possible assessment of susceptible subpopulations. Ability to select with relative ease a "control disease" unrelated to air pollution (e.g motor vehicle accidents, certain fractures, etc) for the time period of interest is an advantage of this dataset. Use of PHC4 data requires contact and agreement with one agency versus multiple health care systems or hospitals for data access.

Limitations of these data include lack of availability of 9-digit ZIP code, lack of specific street address, and limited individual level data (e.g. no information on smoking, occupation, etc) for cohort studies.

3.8 Inventory and Assessment of Emergency Department Data

3.8.1 Emergency Room Visit Data

Given the trends in improved treatments for disease, we are not convinced that mortality or even hospital admissions are alone sensitive enough indicators to capture the all potential effects of daily changes in air pollution on health. To that end, emergency department (ED) visits, physician office visits and medication usage arguably have the most upside potential to capture health effects related to short-term variations in pollutants. In the 1999-2004 time-period to the present, the most well characterized one of these less defined health outcomes (ED visits, physician office visits and medication usage) is ED visits.

At the inception of this feasibility assessment, little was known about the availability of electronic ED visit data from hospitals in the total Pittsburgh region from 1999-2004. It is estimated that nationwide only 30% of all hospitals store ED visit information in electronic format. In general, hospitals associated with large health systems are currently storing, and have archived for a number of years, ED data in an electronic format. However, many of the smaller, independent hospitals have been “online” only recently (2002-present). Obviously, manual data abstraction of ED visit information from paper records is a tedious and costly task and would be prohibitive if funding for a retrospective project is limited. Therefore it was necessary to characterize the electronic data collection and archival capabilities of regional hospitals for ED data from 1999-2004.

3.8.2 Pittsburgh Region Emergency Department Visit Data Archival Survey

In conjunction with the Allegheny County Health Department, the PITT-PM health outcomes project team developed a paper-based survey **(Appendix I)** in order to query regional hospitals concerning their past, current and future ED data collection methods and plans. In addition, variables collected and archived were assessed. A total of forty-five (45) questionnaires were sent in a sampling survey to area hospitals in eleven (11) counties in southwestern PA. Of the 45 hospitals, a total of 41 (91%) responded to the survey; 37 of the 41 hospitals (90%) had active ED departments and provided complete survey information.

Descriptive analyses were completed on the survey responses. As shown in **Table 44** below, a total of 26 (70.3%) of the 37 hospitals reported electronic archival of emergency room visit data for at least some portion of the 1999-2004 time period. However, only 15 of the 37 hospitals (40.5%) reported electronic archival of their data for the entire time period of interest (1999-2004). As expected, hospitals affiliated with the larger health care systems, such as the University of Pittsburgh Medical Center, West Penn-Allegheny Health System and the Pittsburgh Mercy Health System were more likely to report electronic archival than the smaller county or city-based independent hospitals. ED records, particularly for smaller facilities, are less likely than hospital admission records to have ICD-9 codes recorded for primary and multiple secondary diagnoses. Therefore the ability to electronically search the ED notes for chief complaints or keywords associated with a specific condition might be important. In addition, depending on the final study area and health outcome of interest, manual data entry of retrospective ED records for certain crucial hospitals might be desirable for adequate coverage of the study area.

Table 44: Archival of emergency room visit data 1999-2004.

Data Type		Number	Percentage
Hard copy only		11	30
Electronic data		26	70
First full year of electronic data	**1999**	15	41
	2000	1	3
	2001	4	11
	2002	1	3
	2003	2	5
	2004	1	3
	2005	1	3
	Missing	1	3

In **Table 45**, we have outlined the availability of electronic records from 1999-2004 for emergency department visits and number of visits by hospital for the Pittsburgh MSA. Based on data from the Pennsylvania Dept of Health, Bureau of Health Statistics and Research Annual Hospital Questionnaire (2003-2004), a total of 6,557,784 ED visits to hospitals within the Pittsburgh MSA are estimated for the 1999-2004 time period. Allegheny, Armstrong and Butler Counties have the most complete electronic ED data coverage since 1999. Washington and Westmoreland Counties also have considerable available electronic ED data for the period of interest. Conversely, Fayette has limited electronic data until 2003. Hospitals in counties in the expanded Pittsburgh region, such as Greene, Indiana, and Lawrence, also have limited ED data until the latter years of the study period but are not reflected in the table below.

It is also notable that of the 662,292 ED visits to Allegheny County hospitals in 2003-2004, approximately 46% of those visits are to hospitals associated with the University of Pittsburgh Medical Center (UPMC). Data from UPMC-associated hospitals are well characterized, available in electronic format and readily accessible to University investigators with the appropriate data agreement. Representativeness of the UPMC ED visit data related to all of Allegheny County and Pittsburgh MSA visits as a whole would be further explored in the retrospective assessment.

Table 45: Availability of electronic records for emergency room visits and number of visits by hospital for Pittsburgh MSA hospitals, 1999-2004.

	Availability of Electronic Data for Emergency Room Visits*						Number of ER Visits**	Estimated ER Visits 6 yr 1999-
	1999	2000	2001	2002	2003	2004	7/03-7/04	2004
	– none; p - partial(6-11 mo); X - full year							
Allegheny County								
Alle Kiski Medical Center	–	–	–	–	–	–	30,603	183,618
Allegheny General Hospital	X	X	X	X	X	X	44,856	269,136
Childrens Hospital of Pittsburgh (UPMC)	–	–	–	X	X	X	56,183	337,098
Forbes Regional Hospital	X	X	X	X	X	X	39,323	235,938
Jefferson Regional Medical Center	–	–	–	–	–	–	48,784	292,704
Magee-Womens Hospital of UPMC	–	–	–	–	–	–	9,621	57,726
Mercy Hospital of Pittsburgh	–	–	–	–	–	X	40,178	241,068
Mercy Providence Hospital	merged with Mercy Hospital 1/2004						4,526	27,156
Ohio Valley General Hospital	–	–	–	–	–	–	21,534	129,204
Sewickley Valley Hospital	–	–	X	X	X	X	37,169	223,014
St. Clair Memorial Hospital	X	X	X	X	X	X	48,578	291,468
Suburban General Hospital	–	–	–	–	–	–	13,986	83,916
UPMC Braddock	X	X	X	X	X	X	24,292	145,752
UPMC McKeesport	X	X	X	X	X	X	30,259	181,554
UPMC Passavant	X	X	X	X	X	X	31,916	191,496
UPMC Presbyterian/Shadyside	X	X	X	X	X	X	97,395	584,370
UPMC South Side	–	p	X	X	X	X	20,148	120,888
UPMC St. Margaret	–	–	X	X	X	X	32,911	197,466
Western Pennsylvania Hospital	X	X	X	X	X	X	30,030	180,180
VA Healthcare (Federal)	X	X	X	X	X	X		
Allegheny County subtotal							***662,292***	***3,973,752***
Armstrong County								
Armstrong County Memorial Hosp.	X	X	X	X	X	X	25,877	155,262
Armstrong County subtotal							***25,877***	***155,262***
Beaver County								
Aliquippa Community Hospital	–	–	–	–	X	X	12,227	73,362
Medical Center Beaver	–	–	X	X	X	X	46,882	281,292
Beaver County subtotal							***59,109***	***354,654***
Butler County								
Butler Memorial Hospital	X	X	X	X	X	X	37,163	222,978
UPMC Passavant Cranberry	not obtained						16,054	96,324
Butler County subtotal							***53,217***	***319,302***
Fayette County								
Brownsville General Hospital	–	–	–	–	–	–	14,917	89,502
Highlands Hospital	–	–	–	–	–	p	14,475	86,850

	Availability of Electronic Data for Emergency Room Visits*						**Number of ER Visits****	**Estimated ER Visits 6 yr 1999-**
	1999	**2000**	**2001**	**2002**	**2003**	**2004**	**7/03-7/04**	**2004**
	– none; p - partial(6-11 mo); X - full year							
Uniontown Hospital	–	–	–	p	X	X	49,432	296,592
Fayette County subtotal							***78,824***	***472,944***
Washington County								
Canonsburg General Hospital	–	–	–	–	–	–	18,958	113,748
Monongahela Valley Hospital	X	X	X	X	X	X	32,553	195,318
Washington Hospital	X	X	X	X	X	X	39,820	238,920
Washington County subtotal							***91,331***	***547,986***
Westmoreland County								
Frick Hospital	–	–	–	p	X	X	22,420	134,520
Latrobe Area Hospital	–	X	X	X	X	X	36,631	219,786
Mercy Jeannette Hospital	X	X	X	X	X	X	20,298	121,788
Monsour Medical Center	–	–	–	–	–	–	3,419	20,514
Westmoreland Regional Hospital	X	X	X	X	X	X	39,546	237,276
Westmoreland County subtotal							***122,314***	**733,884**
Pittsburgh MSA subtotal							***1,092,964***	***6,557,784***

* Source: Survey by Allegheny County Health Department and University of Pittsburgh GSPH

** Source: Pennsylvania Dept of Health, Bureau of Health Statistics and Research - The Annual Hospital Questionnaire

3.9 Evaluation of Additional Secondary Sources for ED Visit Data Retrieval

The availability of other supplemental electronic data sources for ED visits might improve health outcome coverage for the 1999-2004 time period. In addition to the 10+ county hospital survey of the availability of electronic emergency room information carried out through the auspices of the ACHD, other secondary sources of emergency room information were explored, including ED data collections systems that are unique to the Pittsburgh region as noted below.

3.9.1 UPMC Medical Archival Retrieval System (MARS) database

The University of Pittsburgh Medical Center currently consists of more than 10 hospitals in the region and as such has developed a valuable integrated system of secondary data retrieval. A meeting was held on September 12, 2005 with Ms Melissa Saul, MPH, Director of the Clinical Research Informatics Service, University of Pittsburgh School of the Health Sciences and project staff to discuss the MARS system (Medical Archival Retrieval System) and its utility in providing de-identified electronic information on emergency department visits for the DOE/NETL PITT $PM_{2.5}$ project.

MARS was developed at the University of Pittsburgh in 1986 to improve health care by integrating the computer systems that supported medical care at the departmental level. The concept was to create a complete electronic medical record that would increase the efficiency of patient care and provide the basis for rational decisions about resource allocation. The initial focus of the program was on inpatient hospital care but is now extended to all patients seen at the University of Pittsburgh Medical Center (UPMC) nineteen hospitals, physician offices, and outpatient clinics. The current MARS repository houses 97 million clinical reports and 307 million financial transactions. Complete listings of the clinical, financial and auxiliary databases that are integrated with MARS are presented in **Appendix J** (Melissa Saul, Director, Clinical Research Informatics Services, personal communication).

MARS is implemented in a UNIX-based, distributed parallel-processing environment which is organized around three fundamental concepts. These concepts are (1) MARS accepts all machine-readable data without requiring structure at the point of data entry. Data is transformed into a simple canonical internal format after capture. This eliminates the need for controlled vocabularies and structured entry programs; (2) MARS is indexed on every word and every number in the database with parallel use of a proximity operator. This makes it possible to recognize individual terms, as well as multi-word terms in structured or unstructured data. It also provides the basis for imposing structure on data after collection, through the use of statistics.

In addition to availability of medical notes, the medical record discharge abstract with ICD-9 codes (International Classification of Disease, 9th revision) are provided as a result of each visit as well as demographic information including birth date, gender, race, ZIP code, and county of residence of patient thus making the MARS database extremely attractive for environmental health tracking and disease surveillance. The medical discharge abstract is available for all emergency room visits in all but two of the UPMC hospitals with plans to include these remaining hospitals within the next year.

To supplement the medical record discharge abstract, there are over one million transcribed ED notes available for study since the system's inception for emergency room notes capture in l995. These notes provide specific details of the ED visit including chief complaint, past medical history, physical examination findings and discharge diagnosis. ED notes are available for all but four of the UPMC hospitals. There are plans by MARS officials to include transcribed ED notes from these remaining hospitals.

With regard to the PITT-PM project, the availability of electronic records through MARS for the specific time period January, 1999 through December 2004 was discussed. A request was made by the study team to the Director of Medical Informatics Systems (MS) to provide information on the distribution of available electronic records by year of visit/ county/ZIP code of residence of the patient and hospital site within the UPMC system for the time 1999-2004 time period of interest. A total of 540 Zip Codes are located within 50 miles of Pittsburgh, Pennsylvania, including 487 in Pennsylvania, 17 in West Virginia and 36 in Ohio. Ms. Saul provided a test data set with information from seven of the hospitals within the UPMC Health System as an overview of the catchment area for these UPMC-related healthcare entities. A total of 1,387,025 emergency department visits were recorded at hospitals in the MARS system for the 1999-2004. Shown in **Table 46** are the local area ZIP Codes that contributed at least 5000 ED visits (~ 850 per year) to these seven hospitals for the 1999-2004 period (Total = 1, 086, 834). Highlighted in yellow is the primary hospital (s) that residents from a given ZIP code use for ED services. These data suggest that the majority of patients seen in the larger UPMC hospital emergency rooms are residents of Allegheny County and those residents tend to present to the ED in the hospital closest to their homes. Most Zip Codes in Allegheny County are represented in the MARS database. Residents from outlying counties tend to use the ED services at their county/community hospital. These ED visit data will need to be retrieved in electronic format from these individual community hospitals.

Table 46: ZIP codes contributing >= 5000 ED visits to seven UPMC MARS hospitals (1999-2004).

ZIP Code	BRH	BVH	HHG	HHS	MCH	PUH	SHY	SMH	SSH	Total
15205 Pittsburgh	201	14	54	3	78	2564	699	281	1122	5016
15084 Tarentum	61	4	7	1	17	619	207	4653	26	5595
15144 Springdale	35	2	7	0	9	254	125	5127	38	5597
15135 McKeesport	155	0	5	0	5229	384	93	28	22	5916
16154 Transfer	3	0	4570	1400	1	103	3	3	5	6088
15037 Elizabeth	149	0	16	2	5388	789	126	54	91	6615
15025 Clairton	502	9	29	1	4626	1602	251	72	330	7422
16159 West Middlesex	4	1	178	7291	0	165	3	0	2	7644
15226 Pittsburgh	209	6	24	1	60	2518	686	250	4179	7933
16150 Sharpsville	1	2	1078	6686	3	154	3	7	8	7942
15045 Glassport	282	0	9	0	7239	410	74	20	72	8106
15239 Pittsburgh	464	6	18	5	170	1718	1501	4327	79	8288
15101 Allison Park	75	13	28	6	18	1266	597	6325	79	8407
15112 East Pittsburgh	6614	1	16	0	479	832	438	121	64	8565
15232 Pittsburgh	109	4	7	3	53	2979	5256	299	132	8842
15236 Pittsburgh	494	11	27	6	631	3778	899	257	3138	9241
15116 Glenshaw	80	3	12	0	21	935	496	7816	73	9436
16137 Mercer	2	2	4230	4970	2	303	13	21	4	9547
15212 Pittsburgh	592	39	58	9	158	4047	1414	1804	1450	9571
15145 Turtle Creek	6489	2	12	3	658	1182	987	274	70	9677
15216 Pittsburgh	320	16	36	8	103	3719	995	288	4199	9684
15224 Pittsburgh	292	4	19	3	80	2782	5040	1323	291	9834
15139 Oakmont	91	1	1	0	24	658	700	8365	25	9865
15146 Monroeville	1989	8	15	6	859	3485	2440	1113	209	10124
15642 Irwin	1149	4	52	2	5585	2157	925	342	115	10331
16134 Jamestown	10	3	10001	169	4	138	8	3	5	10341
15133 McKeesport	196	0	6	0	9882	506	113	43	60	10806
15024 Cheswick	48	0	8	0	12	598	259	10215	30	11170
15209 Pittsburgh	187	8	28	1	38	1107	811	9460	227	11867
15223 Pittsburgh	109	4	14	3	19	831	438	10420	81	11919
15131 McKeesport	470	0	16	1	10836	771	311	106	33	12544
15211 Pittsburgh	212	14	15	5	38	2398	651	245	9279	12857
16121 Farrell	1	6	180	14349	0	224	8	5	2	14775
15201 Pittsburgh	357	8	16	7	77	3296	4799	7136	351	16047
15208 Pittsburgh	792	1	3	3	429	6301	8083	897	287	16796
16148 Hermitage	3	0	893	15738	1	417	18	11	4	17085
15137 North Versailles	6031	5	14	0	8905	1254	794	188	88	17279

ZIP Code	BRH	BVH	HHG	HHS	MCH	PUH	SHY	SMH	SSH	Total
15207 Pittsburgh	1838	4	34	2	536	9144	3641	303	1908	17410
15068 New Kensington	181	4	30	5	91	2171	924	14149	96	17651
15238 Pittsburgh	67	1	16	3	21	1206	821	15501	62	17698
15227 Pittsburgh	465	17	64	4	273	4045	1018	282	11812	17980
15219 Pittsburgh	568	19	5	4	269	12751	4185	469	942	19212
15203 Pittsburgh	199	8	7	2	87	2586	601	202	17743	21435
16146 Sharon	7	5	787	20489	3	340	18	18	18	21685
15122 West Mifflin	9007	6	52	0	7645	3023	1056	204	781	21774
15215 Pittsburgh	132	3	9	2	18	1300	741	19633	114	21952
15218 Pittsburgh	12361	10	15	0	363	4473	4606	550	296	22674
15001 Aliquippa	67	20734	34	21	28	1491	159	64	79	22677
15213 Pittsburgh	327	13	8	8	129	16390	5673	437	585	23570
15110 Duquesne	4980	3	4	3	16303	1776	630	62	255	24016
15217 Pittsburgh	553	4	9	3	175	11309	10753	720	491	24017
15147 Verona	761	7	30	0	145	2683	3776	17992	160	25554
15120 Homestead	17434	13	61	4	2134	5530	2553	289	913	28931
15235 Pittsburgh	2153	12	31	3	520	7604	8983	10395	402	30103
15104 Braddock	35766	13	14	4	804	2615	1307	179	190	40892
15206 Pittsburgh	1103	12	24	11	419	12767	19889	6423	896	41544
15221 Pittsburgh	11305	24	19	3	708	12876	16428	1820	580	43763
16125 Greenville	8	6	51237	2095	3	607	34	20	14	54024
15210 Pittsburgh	877	23	50	3	331	8440	1827	468	47065	59084
15132 McKeesport	2904	14	27	5	92185	3910	841	177	353	100416
Total										**1086834**

Abbreviations: BRH= UPMC Braddock; BVH = UPMC Beaver Valley; HHG= UPMC Horizon Greenville; HHS = UPMC Horizon Shenango; MCH= UPMC McKeesport; PUH= UPMC Presbyterian; SHY= UPMC Shadyside; SMH= UPMC St. Margaret's; SSH =UPMC South Side

3.9.2 Availability of MARS Data

Data is continuously fed into MARS over a network from several hundred clinical and financial domains. It is estimated that almost 500,000 new clinical reports and 450,000 financial transactions are received each week. Approximately 15,000 - 20,000 reports are retrieved daily for the support of clinical activity, and there are approximately 5100 logins each day. MARS offers three types of user interface: (1) an intelligent terminal interface, which formulates Boolean queries automatically for users; (2) a World Wide Web browser interface; and (3) a batch command and editing interface, which supports customized retrieval strategies for activities such as medical rounds or commonly used queries.

3.9.3 Integration of MARS Data

All records obtained on a single patient at any given time are linked via a unique patient identifier. Patients who cross institutional (hospital) domains are linked through a Master Patient Index maintained in an Oracle^{TM} database. In addition, a minimum of three demographic items are stored with each record. This strengthens linkages and facilitates searching for common patient characteristics within clinical and financial records.

3.9.4 Research Applications using MARS Technology

CRIS (Clinical Research Informatics Service) is a jointly sponsored service of the Office of Clinical Research and the Center for Biomedical Informatics (Ms. Melissa Saul, Director). CRIS is available for use by faculty in the Schools of the Health Sciences, University of Pittsburgh and for UPMC special projects requiring de-identified datasets. CRIS is a certified honest broker with the University of Pittsburgh IRB and has a business associate agreement with UPMC. The polices and procedures of CRIS are posted on the Office of Clinical Research website (http://www.clinicalresearch.pitt.edu)

CRIS uses the De-ID application developed by the Center for Biomedical Informatics at the University of Pittsburgh and licensed by the University to De-ID Data Corp, Philadelphia, PA. The De-ID application is used by the National Cancer Institute and other academic medical centers for various research applications. De-ID has as its main usage, the ability to consider medical information housed with a person's record with safeguards to that individual's identity. The PITT-PM project will be limited to obtaining diagnostic code and limited demographic information on individual emergency visits. However, the ability to cross identify cardio-pulmonary outcomes, by ICD codes, date of birth, gender, race and place of residence without names and addresses through the honest broker system will be an invaluable tool in the conduct of this project.

De-ID automatically creates a linkage file when a dataset is processed. The linkage file is stored in an encrypted format and only available for viewing with the password given at the time of processing. The study identifier is a two-part code; part one is the number of the report for that patient; and part two is a unique 12 alphanumeric code for that patient. This is done so that the study id remains consistent across data sets but different admissions and/or multiple reports can be easily identified.

The Center for Biomedical Informatics (CBMI) performs formal evaluations of the De-ID© software. Five physicians are doing a current evaluation at UPMC Presbyterian. Also, the Center for Pathology Informatics performed an independent evaluation of the De-ID software last year (Gupta et al., 2004)

3.9.5 Real-Time Outbreak and Disease Surveillance (RODS) Data

On June 17, 2005 the PITT-PM health outcomes sub group met with Dr. Michael Wagner, Director the of the RODS Laboratory at the University of Pittsburgh Center for Biomedical Informatics. The Real-time Outbreak and Disease Surveillance (RODS) Laboratory is a collaboration between Dr. Wagner and colleagues at the University of Pittsburgh Center for Bioinformatics and the Auton Lab at the Carnegie Mellon University School of Computer Science. The laboratory was founded in 1999 to investigate methods for real-time detection and assessment of disease outbreaks. The objectives of the project include algorithm development, assessment of novel types of surveillance data, natural language processing and analyses of syndrome detectability. The laboratory is home to four large projects that work with health departments to create surveillance systems: RODS software development, the Public Health Data Center, the National Retail Data Monitor (NRDM) and the BioWatch Support Program.

The primary focus of this work is the use of real time streaming of ED visit information related to eight syndromic clusters including: respiratory, nausea, rash, neurologic, constitutional, gastrointestinal and other disorders that might be environmentally driven (e.g. infectious agents). Currently 110 of 190 hospitals throughout the Pennsylvania feed information into the RODS server through Health Level 7 (HL7) formatted messaging. RODS includes over 80% of ED visits in Allegheny County. For the 1999-2004 time- period, several area hospitals collected ED data in hard copy format and did not maintain on-site electronic ED databases. Since RODS has been acquiring ER information for the period 2000-2004 for the majority of these hospitals, it may be possible to retrieve information on these ED visits from the RODS Public Health Data Center for application in research studies.

RODS software can process and display the data in the form of graphs and maps via a secure web interface. The data can also be run against other surveillance software or algorithms other than what is included in the RODS software. RODS is capable of receiving and analyzing several types of data such as emergency department registrations, chest x-rays, orders for cultures, culture results, etc. These records are captured as HL-7 messages that are transmitted directly to RODS from a hospital or health system's HL-7 message router in real time. Using the application does not require any client software as the processed data is viewed through a web browser. Hospitals send the data via an HL-7 interface that is created to automatically send existing messages via a VPN connection, SSL or SFTP directly from the hospital or health system's message router to the Data Center in real time. Historical data provided by a hospital is used to build an accurate baseline for alerting. Automatic detection algorithms run on the data and search for anomalies that may indicate an outbreak; alerts are issued automatically to health department officials.

Currently, the data elements collected include age (without date of birth), gender, home ZIP code and free-text chief compliant (e.g. shortness of breath, chest pain, etc). These data have been collected and warehoused since 2000. RODS software aggregates the data into daily counts by syndrome (natural language processing of the chief complaint) and residential ZIP code for analysis. In this feasibility assessment, the PITT-PM investigators are evaluating the possibility of utilizing these retrospectively collected archived data. At the time of initial presentation in the ER, a patient has not yet been assigned an ICD9/10 code nor has the physician of record diagnostically verified the chief complaint. However, this syndromic clustering reported at initial presentation might nonetheless have utility as a sensitive endpoint and real time indicator of any association between point source air pollution exceedances and respiratory (or cardiovascular) health outcomes. An elevated average or maximum $PM_{2.5}$ concentration (or certain speciated components) within a certain area or ZIP code could be correlated with spikes in the "respiratory syndrome" pattern over time. This type of pattern analysis might identify a more direct link between a "pollutant upset" in a certain area than hospital admissions over a one to two day period. These syndromes can be validated by later comparison with hospital admission data or ER discharge data at a 24- or 48-hour lag.

The DOE/NETL cooperative agreement involves collaboration of GSPH with the ACHD, which is currently intimately involved in the RODS program of syndromic data evaluation and analysis. Our plans are to continue to explore, in conjunction with ACHD, the use of these emergency room data collected in real-time in a retrospective epidemiological study.

3.9.6 Physicians' Office Visits

As noted previously, capturing physician office visits retrospectively for respiratory and/ or cardiovascular disease exacerbations is a difficult task for a city, county or regional research study. Unfortunately, no central agency or organization in Pennsylvania is responsible for the collection of data on such visits. However, retrospective physician office data can potentially be accessed through several of the patient care provider networks in the area. As discussed in the previous progress reports, the Medical Archival Retrieval System (MARS) at UPMC aggregates office visit data for physicians affiliated with the UPMC health care system. In addition, several health maintenance organizations (HMOs), preferred provider organizations (PPOs) and point of service (POS) plans in the area collect office visit data on their subscribers. Most plans have been operational in the Pittsburgh region since the early 1980s. The largest regional HMOs and/or health plans are described briefly in **Table 47**.

Regional HMOs offer perhaps the best opportunity for capturing retrospectively office visit information for the 1999-2004 period. The four largest southwestern Pennsylvania HMOs covering the Pittsburgh SMA include UPMC Health Plan (~134,000 HMO enrollees), Keystone Health Plan West (~110,000 HMO enrollees), Aetna (~27,000 HMO enrollees) and Health America (PA total ~200,000; Pittsburgh SMA total not available). Also Gateway Health Plan, established in1992 as a managed care alternative to the Department of Public Welfare's Medical Assistance Program in Pennsylvania, serves a number of medical assistance recipients throughout Pennsylvania, particularly in Allegheny County. Five Pittsburgh POS plans cover an additional ~300,000 local enrollees; ten PPOs (including Highmark, UPMC, Aetna and Health Assurance) have approximately 900,000 subscribers. Certain information might be available through claims data, although not as readily as in the better characterized participation within an HMO.

In the past, most of these organizations have indicated a willingness to assist public health agencies, including the Allegheny County Health Department and the University of Pittsburgh Graduate School of Public Health, in disease surveillance and research efforts. However, since the passage of HIPAA, many health care organizations are reluctant to share individual enrollee data with outside research institutions. Several of the health plans above, including Keystone Health Plan West and Health America, voiced these concerns during recent conference calls. De-identified aggregated data (e.g., number of physician office visits on a given day) might be accessed more readily. Although aggregated daily counts of office visits might be accessible, individual level data such as age, city, ZIP code, etc. is likely restricted. The UPMC Health Plan, as an example, does have procedures and provisions for data sharing of a HIPAA-compliant limited data set with academic and other research institutions. This type of data set will contain certain individual level information including date of birth, date of visit (admission), and geographic data to the level of county, city and ZIP code. Limited data sets can be used and disclosed only for research, public health or health care operations. However, UPMC officials noted that, more recently, data sharing in which data files are moved off-site for analysis has been minimal. Requests for data sharing are entertained on a study-by-study basis and are handled based on the organizational trust built between the data steward and the external investigator. Data usage agreements must be executed with each individual agency prior to the partnering of these organizations with any non-covered entity. Third-party “honest broker systems” are a requirement for access to individual level data that can potentially be traced back to individual patients. Access to these types of data sets might involve a fee payment ($100 to $3000) to the health plan for database assembly costs.

Table 47: Pittsburgh MSA (regional) health plans and HMOs.

Name	Date Licensed (Pittsburgh)	~ Total Local Enrollees[a]	Participating Local Hospitals	Geographic Coverage[b]	HMO	PPO	POS
Highmark Blue Cross/Blue Shield (Keystone Health Plan West)	1986	400,000	34	Pittsburgh MSA + 21	√	√	√
University of Pittsburgh (UPMC Health Plan)	1998	225,000	38	Pittsburgh MSA + 19	√	√	√
Aetna	1986	80,000	25	Pittsburgh MSA + 42	√	√	√
Health America of PA (Health Assurance)	1975	200,000[c]	38	Pittsburgh MSA + 54	√	√	√
Intergroup Services	~1986	520,000	38	Pittsburgh MSA		√	
American Health Care Group	1980s	130,000	25	Pittsburgh MSA		√	
Cigna Healthcare	1980s	90,000	37	Pittsburgh MSA		√	√

[a] Includes HMO, PPO and POS plans. [b] Pgh SMA + NUM = Pittsburgh Statistical Metropolitan Area (includes Allegheny, Armstrong, Beaver, Butler, Fayette Washington, and Westmoreland counties + additional number of counties covered in PA.
[c] Total in PA

A significant issue in the use of office visits or symptomatology at office visits as an outcome is the retrospective differentiation of office visits scheduled for regular exams vs. unscheduled visits associated with exacerbation of circulatory or respiratory disease potentially attributable to air pollutants. For the majority of the HMOs, this information is not recorded or available as a part of the medical record. As an additional caveat to use of retrospective health plan data, HMOs have noted declining enrollment in the last 5 years. This recent enrollment volatility might preclude an accurate assessment of the relationship between air quality and unscheduled office visits thought HMOs from 1999-2004, specifically if a health or poor health bias is associated with the shift from HMOs to other types of health plans. Partnering with these plans for a prospective (longitudinal) study in which data collection can be tailored to the needs of the study, might be more practical and meaningful but considerably more expensive in terms of staff time and commitment.

3.9.7 Pharmaceutical Usage Databases (Prescription Medications)

Prescription requests and medication use are potentially more sensitive indicators for exacerbation of certain circulatory and respiratory diseases related to ambient and/or indoor $PM_{2.5}$. Typically, studies assessing medication use in relation to changes in ambient air pollutants involve costly prospective panel studies of high-risk adults or children. However, certain pharmaceutical usage databases might be available for retrospective data analysis. For instance, Verispan (formerly Scott-Levin Pharmaceutical Company, Yardley. PA) and IMS Health (Fairfield, CT) collect comprehensive data on prescription drug usage in multiple metropolitan centers across the US. These databases have traditionally been used as market research tools for the pharmaceutical industry, but have broad applications to health research. Both Verispan and IMS Health officials indicated a keen interest in partnering with academic and industry groups to collect data in a manner that will facilitate health outcomes research. Currently, however, the type of data collected and the timeliness (or lack thereof) of reporting are issues that make these data unsuitable for a retrospective analysis.

3.9.8 Verispan Datasets

Verispan has secured rights to data for nearly half of all U.S. prescriptions and nearly one-quarter of all U.S. electronic medical transactions annually. Verispan captures more than 25% of all prescriptions from 98% of all 3-digit zip codes and 45% of all prescriptions from approximately 80% of all zip codes (Source: http://www.verispan.com/). Verispan can provide insight into prescription and medical activity at the national, regional and individual prescriber level. Verispan’s Source Prescription Audit (SPA) and Physician Drug and Diagnosis Audit (PDDA) represent novel potential sources of health data for retrospective analyses. A limitation of the use of these Verispan datasets in health outcomes research is the temporal and spatial resolution of the data. Coverage in defined geographic areas is limited; data are currently sent only monthly from participating providers. Finer resolution is possible but will most likely require partnering with Verispan to construct an appropriate data pass-through scheme for prospective studies.

3.9.9 IMS Health Datasets

Similarly, IMS Health (http://www.imshealth.com/) receives pharmaceutical usage data from more than 29,000 data suppliers covering 225,000 data sites worldwide. Data sources include drug manufacturers, wholesalers, retailers, pharmacies, mail order, long-term care facilities and hospitals. IMS also captures consumer purchase information from pharmacies equipped with Electronic-Point-of-Sale (EPOS) systems (PharmaTrend).

IMS does collect ZIP code level data but does not typically sell these data to non-PharmaTrend clients. A specific agreement would need to be negotiated to obtain the data. Normal one time, one market ZIP code level reports can cost between $75,000 -$100,000. ZIP code level reports capture sales of products into the channels of trade; they do not track dispensed Rx's. ZIP code sales data are captured monthly at its most granular level. Daily prescirptions are potentially available through Early Insight, an IMS Health web-based application that tracks daily prescription volume and market share. IMS has MSA level data that tracks prescriptions at much more reasonable costs. Data are available from 1999-present. Unfortunately, Pittsburgh MSA level data is currently captured only monthly.

3.9.10 National Retail Data Monitor (Over-the-Counter Sales)

The National Retail Data Monitor (NRDM) is a public health surveillance tool that collects and analyzes daily sales data for over-the-counter (OTC) health-care products. NRDM grew out of the Pennsylvania Retail Data Monitor, a system developed by the Commonwealth of Pennsylvania and the Real-time Outbreak and Disease Surveillance (RODS) Laboratory (http://www.health.pitt.edu/rods) at the University of Pittsburgh. The Pennsylvania system began receiving data from retailers in December 2002 and was expanded in scope to a nationwide initiative soon after its introduction. The current coverage of retail data nationwide is approximately 20%, but much higher in many large urban areas, particularly in Pennsylvania.

NRDM collects sales data for selected OTC health-care products in near real time from >15,000 retail stores and makes the information available to public health officials. NRDM is one of the first examples of a national data utility for public health surveillance that collects, redistributes, and analyzes daily sales-volume data of selected health-care products, thereby reducing the effort for both data providers and health departments.

After decades of investment into developing Universal Product Codes (UPCs), optical check-out scanners, and analytic data warehouses, the retail industry has in effect constructed 95% of a surveillance-system pyramid onto which a capstone of data integration and analytic capability can be added to produce NRDM. NRDM's objectives are to 1) enlist participation of retailers to achieve 70% coverage of OTC sales nationally; 2) influence the industry toward real-time data collection; 3) obtain supplemental information needed for spatial analysis, adjustment for promotional effects, and maintenance of UPC analytic categories (e.g., liquid cough medications); 4) promote and develop this type of surveillance practice; 5) achieve fault and load tolerance; and, 6) develop detection algorithms for the data (http://rods.health.pitt.edu/NRDM.htm).

Although not specifically collected for air pollution research, data previously assembled on OTC drugs purchased for respiratory illnesses from 2003-2004 might be useful in evaluating for little additional cost the overall health effects of variations in $PM_{2.5}$ concentrations and components in a retrospective study. We will continue to explore the availability and usefulness of these data for retrospective and prospective studies in air quality research.

3.9.11 Implantable Cardioverter Defibrillators

Individuals at risk for sudden cardiac death with implanted cardioverter defibrillators (ICDs) represent a unique group of subjects particularly sensitive to changes in the levels of fine particulates. No population-based registry for ICDs was identified in the Pittsburgh metropolitan region that covers the 1999-2004 time period. Since January 2005, however, the Centers for Medicare and Medicaid Services have mandated a registry for all Medicare patients undergoing implantation of ICDs (http://www.cms.hhs.gov/CoverageGenInfo/07_ICDregistry.asp - TopOfPage) for primary prevention of sudden cardiac arrest. The registry also includes a longitudinal component to capture follow-up data on ICD patients. Prospective studies will potentially be able to tap into this relatively new database for research studies.

From 1998-2003, the University of Pittsburgh Medical Center and the VA Pittsburgh Healthcare System were principal sites in the multi-site *Defibrillators in Non-Ischemic Cardiomyopathy Treatment Evaluation (DEFINITE) Trial and Registry* that followed 458 subjects with non-ischemic dilated cardiomyopathy. A total of 229 subjects in the trial underwent prophylactic defibrillator implantation in one of the two study treatment arms. This select group might also represent a potential panel for a longitudinal analysis.

3.10 Strengths and Weaknesses of ED Visits, Physicians Office Visits and Pharmaceutical Data as a Health Endpoint for Retrospective Studies

Health outcomes such as ED visits, physicians' office visits, prescriptions and medication use are most probably more sensitive to short-term fluctuations in $PM_{2.5}$ than either mortality or hospital admissions. These data are, however, much less readily available retrospectively in a central repository and/or in electronic format. ED visits are the better characterized of these datasets, but are not available in a standardized format from a central collection agency in Pennsylvania for all hospitals. ED visit information would be acquired in electronic format from major hospital systems and/or individual hospitals through separate protected access agreements. Data from the MARS and RODS systems can help to supplement the ED visit information. From an outcomes perspectives, ED data could be explored as health outcome of interest in a more limited geographic area, in specific counties such as Allegheny, Armstrong, Butler, Washington and Westmoreland. These data might help to fill gaps in health outcomes information related to $PM_{2.5}$ in the Pittsburgh region.

It is more costly and impractical to acquire and assemble a comprehensive retrospective dataset for 1999-2004 for physicians' office visits and/or pharmaceutical data. Nonetheless, for future prospective studies on $PM_{2.5}$ and health, the richness of these datasets and the potential for partnering with the data owners, particularly the pharmaceutical market research industry, is intriguing.

3.11 Long-term Effects of $PM_{2.5}$ on Health Outcomes in the Pittsburgh Region (Retrospective Cohort Studies)

Time series studies are designed to estimate the short term effect of $PM_{2.5}$ on mortality and morbidity as health endpoints of interest. For example, in a time series analysis, death is considered to be a "once only event" with no dimension in time (Kunzli et al., 2001). Therefore, time series studies assume that the event is influenced by factors that act shortly before the event (death), such as acute weather changes, day-to-day variation in air pollution, etc. Illness and death are, however, likely influenced by multiple exposures over time, potentially years or even decades earlier. These long term effects of $PM_{2.5}$ on health outcomes are more appropriately captured in cohort (either prospective or retrospective) studies rather than time series analyses.

For the assessment of the long term effects of $PM_{2.5}$ on residents of the Pittsburgh region, the incidence of disease (e.g. respiratory, cardiovascular) and/or death would be assessed in a defined population over a specified period of time. Prospective cohort studies require the observation of persons from a point in time into the future. These studies require a large sample size and a long period of follow-up and are, therefore, both expensive and time-consuming and inappropriate for a study capturing effects from 1999-2004. Retrospective cohort studies use study populations that were defined in the past by exposure and can be located and evaluated today for health outcomes of interest. Retrospective cohort designs are generally more cost effective and less labor intensive, although identification and assembly of a retrospective cohort can be difficult.

The American Cancer Society's Cancer Prevention Study II (CPS-II) is an ongoing prospective mortality study of over 1.2 million adults recruited in 1982 by ACS volunteers. A recent ancillary study included about 500,000 of the CPS-II participants who resided in over 100 metropolitan areas for which data on air pollutants were available. This cohort study included participants from the Pittsburgh metropolitan area. Overall mortality rates, as well as mortality rates for cardiopulmonary diseases, lung cancer, and other causes were determined from 1982 to 1998, longer follow-up than for any previous study. The results of the study suggested that long term exposure to combustion-related $PM_{2.5}$ was a risk factor for cardiopulmonary and lung cancer mortality. No analyses were performed on a regional basis. These data (and subjects) are potentially available for further evaluation (retrospectively and prospectively) of the long term effects of speciated components of $PM_{2.5}$.

The University of Pittsburgh Graduate School of Public Health has also recruited several local cohorts, both independently and as a site for various national multicenter studies, such as the Cardiovascular Health Study (CHS), the Study of Osteoporotic Fractures (SOF), Women's Health Initiative (WHI), Health, Aging and Body Composition (Health ABC) Study and others. Participants from several of these mostly middle-aged and elderly local cohorts would potentially be available for a study of the long term effects of $PM_{2.5}$ in the Pittsburgh region. For example, the Health ABC Study enrolled locally approximately 1500 subjects and the SOF and WHI studies approximately 2000-2500 individuals each. The individual level data collected for these studies is a potentially rich resource for retrospective cohort analyses.

In addition, the Pittsburgh regional HMOs, initiated in the 1980s and 1990s, remain a potential secondary source for retrospective (and prospective cohorts). As noted previously, however, these healthcare entities are less receptive to partnering with outside institutions for research since the passage of the HIPAA regulations related to patient confidentiality and privacy. The University of Pittsburgh Graduate School of Public Health does have an excellent relationship with these health plans and the expectation is that these groups will be willing to share information with acceptable "honest broker" agreements in place.

The emerging fields of genomics and proteomics make the assessment of biomarkers for exposure to air pollutants in cohorts an attractive area for future air pollution research. PAH-DNA adducts and protein adducts such as benzopyrene-hemoglobin and 4-ABP-hemoglobin have been evaluated in human populations exposed to differing levels of air pollutants primarily in Europe (Vineis and Husgafvel-Pursiainen, 2005). As these fields mature, it will potentially be possible to assess various biomarkers for exposure to $PM_{2.5}$ from specific sources. Such a cohort study is currently, however, outside the scope of the PITT-PM study group's retrospective study proposal.

3.12 Key Health Outcome Issues to Consider in the Design of a Retrospective Study of Speciated $PM_{2.5}$ and Health Effects

There are several key issues are of importance in the design of a retrospective study to assess the health effects of $PM_{2.5}$ and its components. These points are outlined in this section and are addressed in the proposed retrospective study design. These include but are not limited to:

- Sensitivity of the health outcome of interest to short (or long term) effects of $PM_{2.5}$ or its components (exploration of hospitalizations, ED visits, physicians office visits as outcomes for $PM_{2.5}$ retrospective research efforts)

- Selection of specific respiratory (asthma, chronic bronchitis, pneumonia) and circulatory (all ischemic heart disease, myocardial infarction) admission disease categories for analysis; considerations for stratification of admission by disease subcategories for time series analyses if power analysis permits

- Specific evaluation of vulnerable populations such as the elderly and the very young
- Use of a single health outcome of interest (mortality, hospitalizations, emergency room admissions or other, separately) or validation of some composite health outcome variable to assess short term effects
- Possibility of "dilution of effects" by inclusion of health outcomes that are not related to $PM_{2.5}$ or populations that are not exposed to $PM_{2.5}$
- Consideration of separate analyses for respiratory and circulatory disease, given the difference in lags suggested in literature for $PM_{2.5}$ effects
- Power issues related to conducting separate analyses for specific sub-disease category health outcomes (e.g., asthma, ischemic heart disease, etc)
- Sub-region analyses: given the regional nature of many pollutants, is it possible to effectively evaluate a smaller geographical entity – e.g. ZIP code level data – in the time series analysis? Number of daily admissions of ED visits required to ensure enough statistical power to detect a significant small area effect if it exists; influence of variability in daily counts vs. number of total available days of interest on overall statistical power of the study
- Evaluation of mortality and morbidity in existing cohorts to determine the long-term impact of exposure to $PM_{2.5}$ and its component species

3.13 Development of Comprehensive Health Outcomes Datasets for a Retrospective Epidemiological Study

Although the health outcomes datasets that are available retrospectively for the Pittsburgh region have been identified, these datasets have not been physically acquired by the study team as a part of this feasibility assessment since acquisition would require significant time and cost that were beyond the scope of this project. We have determined, however, that both mortality and hospitalization standardized data are available from 1999-2004 (and to the present 2006) for the region. We have also demonstrated that ED data are available from 40% of the area hospitals from 1999-2004 and that the hospitals with the most complete data are associated with the large healthcare systems (UPMC Health System, West-Penn Allegheny Health System, Mercy Health System) that are more likely to partner with university and industry-based research groups. We have obtained and used test data for Allegheny County hospitalization descriptive analyses and for the preliminary assessment of statistical models. The complete health outcomes datasets will be obtained as a task in the proposed retrospective epidemiology study plan and a comprehensive health outcomes database constructed for statistical analyses.

3.14 Technical and Cost Analysis for the Health Outcomes

3.14.1 Technical Analysis

From this feasibility assessment, it has been determined that retrospective health datasets for mortality and hospitalizations from 1999-2004 or later can be constructed for the Pittsburgh region. An ED visit dataset can also be constructed but will likely be more limited in geographic coverage. All data to be used for the retrospective epidemiological assessment of the health effects of $PM_{2.5}$ particulates will be obtained exclusively from existing secondary data sources, primarily at the onset of the project period. $PM_{2.5}$ data will be obtained from various federal, state and local air quality monitoring networks, including the U.S. EPA Air Quality Monitoring System, U.S. Department of Energy National Energy Technology Laboratory and the Pittsburgh Supersite. Mortality data will be obtained from the Pennsylvania Department of Health Bureau of Health Statistics and Research and verified using National Center for Health Statistics (NCHS) Division of Vital Statistics. Recent quality analysis comparing these electronic datasets to death certificates suggests that the error rate is 2% or less. Hospitalization data is collected by the Pennsylvania Health Care Cost Containment Council (PHC4). The data are processed using a series of validation rules before being finalized and made available for further analysis and public release. PHC4 edits the data and provides error reports to each data source. The health care facility will make error corrections and provide PHC4 with corrected information. Compliance across health care institutions in Pennsylvania approaches 100% (99% in recently released 2006 reports). Emergency department (ED) data will be acquired from individual hospitals/hospital systems through directed agreements. If necessary, the investigators will utilize an "honest broker" system to acquire identified ED data from hospitals for use in the study. Verification of the accuracy and integrity of the ED and other data will be conducted by the data research associate and will include ID verification, data range, and type verification, and duplicate entry checks. Additional data editing and report generation will be performed quarterly to assure data integrity and completeness. Meteorological parameters will be obtained from the National Oceanic and Atmospheric Administration (NOAA). These data are governed by strict quality guidelines as described online (http://www.cio.noaa.gov/itmanagement/IQ_Guidelines_110606.htm).

3.14.2 Cost Analysis

For the health outcomes aspect of the PITT-PM project, the greatest staff time commitment is associated with the acquisition and quality validation of the various health datasets, including mortality, hospitalizations and ED visits. Since the mortality and hospitalization data are collected and validated at centralized agencies by regulation or statute, this task is less time consuming. For emergency department data, however, the PITT-PM health outcomes group will collect and assemble data from various health care entities and will be responsible for the cleaning and quality assurance of the data and the validity of the final dataset. Descriptive statistical analyses of all individual and aggregated data will be performed.

4 Statistical Methodology Assessment

4.1 Time Series Power Analysis

4.1.1 Introduction

A power analysis was conducted to help to determine the length of a time series needed to adequately determine the relationship between exposure factors and health effects. A data set was assembled covering the period from 10/9/98 to 12/31/00 for Pittsburgh using NMMAPS (http://www.ihapss.jhsph.edu/software/NMMAPS/NMMAPS.htm) and ACAPS data sets. The admissions data for people over 65 years of age was used from the ACAPS data set while $PM_{2.5}$ concentrations and other co-pollutants were taken from the NMMAPS data set. Summary statistics are shown in **Tables 48 and 49.** The R language and environment for statistical computing (R Development Core Team, 2006) was used for almost all computations. (R is freely available for download at http://www.r-project.org/.)

Table 48: Descriptive statistics for NMMAPS/ACAPS time series data.

	Elderly Hospital Admisisons	$PM_{2.5}$	SO_2	NO_2	NO_x	NO	Ozone	Temper-ature	Relative Humidity
Mean	113.861	16.220	0.011	0.023	0.043	0.019	0.025	50.870	48.500
Standard Deviation	28.077	10.240	0.006	0.007	0.030	0.025	0.014	17.186	15.921

Table 49: Correlation matrix of exposure and health variables from the NMMAPS/ACAPS dataset. The two highest correlations are highlighted in yellow.

	Admiss.old	tp	cs	sn	pm25	so2	no2	nox	no	ozone	mntp	mnrh
Admiss.old	1.00	-0.25	0.15	0.26	-0.06	0.00	-0.05	0.01	0.03	-0.09	-0.16	-0.04
temp	-0.25	1.00	0.14	-0.41	0.00	0.06	-0.10	-0.05	-0.03	-0.09	-0.03	0.32
cs	0.15	0.14	1.00	0.00	-0.16	0.37	0.21	0.40	0.40	-0.80	-0.81	0.14
sn	0.26	-0.41	0.00	1.00	-0.10	-0.06	0.03	0.00	-0.01	-0.03	-0.33	-0.25
pm25	-0.06	0.00	-0.16	-0.10	1.00	-0.03	0.07	0.00	-0.02	0.13	0.36	-0.04
so2	0.00	0.06	0.37	-0.06	-0.03	1.00	0.49	0.53	0.50	-0.27	-0.26	0.01
no2	-0.05	-0.10	0.21	0.03	0.07	0.49	1.00	0.81	0.68	-0.16	-0.17	0.00
nox	0.01	-0.05	0.40	0.00	0.00	0.53	0.81	1.00	0.98	-0.39	-0.32	0.03
no	0.03	-0.03	0.43	-0.01	-0.02	0.50	0.68	0.98	1.00	-0.44	-0.34	0.04

	Admiss.old	tp	cs	sn	pm25	so2	no2	nox	no	ozone	mntp	mnrh
ozone	-0.09	-0.09	-0.80	-0.03	0.13	-0.27	-0.16	-0.39	-0.44	1.00	0.66	-0.11
mntp	-0.16	-0.03	-0.81	-0.33	0.36	-0.26	-0.17	-0.32	-0.34	0.66	1.00	-0.07
mnrh	-0.04	0.32	0.14	-0.25	-0.04	0.01	0.00	0.03	0.04	-0.11	-0.07	1.00

Note: tp (time point) is an index for the day starting at 10,508 and ending at 11,322. cs and sn are cosine and sine functions of the time to handle seasonality. mntp is the mean daily temperature and mnrh is the mean relative humidity.

It should be noted that this first simple regression analysis completely ignores autocorrelation in the dependent and independent variables. **Figure 54** shows the autocorrelation function (ACF) and partial autocorrelation function (PACF) for the residuals for the model of admissions as a function of $PM_{2.5}$, weather, and co-pollutants. Although the day-of-week, trend, and yearly seasonality effects have been removed, there is still a substantial and statistically significant correlation for several lags remaining. **Listing 1** shows the estimated coefficients for the model. The standardized coefficient estimates are shown in **Listing 2.**

To determine the variance inflation factor (the increase in standard errors due to the intercorrelation among the independent variables), $PM_{2.5}$ was regressed on the remaining independent variables and the results are shown in **Listing 3** The R^2 was 0.2265 and the variance inflation factor was estimated to be 1.29 .

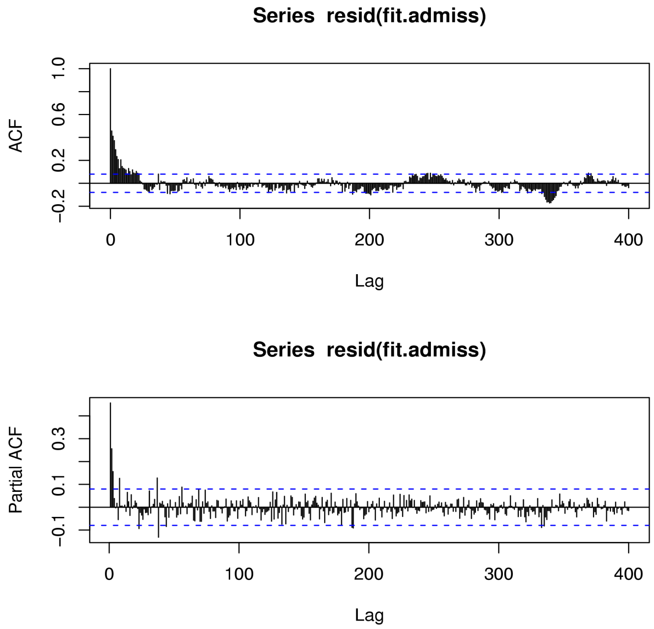


Figure 54: Autocorrelation function (ACF) and partial autocorrelation function (PACF) for residuals.

Analysis of the residuals from the fitted linear regression (**Listing 4**) shows an approximate autoregressive process of order 3 (ϕ_1=0.30, ϕ_2=0.22, ϕ_3=0.20) with σ=14, approximately. (The ACF of the residuals to the fitted AR model showed no statistically significant autocorrelation confirming the adequacy of the model.)

Listing 1: Model for 65 and over hospital admissions (admiss.old).

```
> summary(fit.admiss <-
lm(admiss.old~dow+tp+cos(2*pi*tp/365)+sin(2*pi*tp/365)+pm25mean+so2+no2+no+ozone+mntp+mnrh,
+   data=complete.copy[!is.na(complete.copy$pm25mean),]))

Call:
lm(formula = admiss.old ~ dow + tp + cos(2 * pi * tp/365) + sin(2 *
    pi * tp/365) + pm25mean + so2 + no2 + no + ozone + mntp +
    mnrh, data = complete.copy[!is.na(complete.copy$pm25mean),
    ])

Residuals:
     Min       1Q   Median       3Q      Max
-90.8484  -8.6963   0.7477   9.6822  61.7141

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          253.533196  44.472980   5.701 1.89e-08 ***
dowSun                -2.868334   2.732245  -1.050 0.294236
dowMon                49.741030   2.643650  18.815  < 2e-16 ***
dowTue                45.139919   2.668532  16.916  < 2e-16 ***
dowWed                40.609712   2.649529  15.327  < 2e-16 ***
dowThu                37.551810   2.607584  14.401  < 2e-16 ***
dowFri                33.057982   2.754037  12.003  < 2e-16 ***
tp                    -0.017222   0.003961  -4.348 1.62e-05 ***
cos(2 * pi * tp/365)  14.528318   2.497498   5.817 9.82e-09 ***
sin(2 * pi * tp/365)  10.747829   1.300285   8.266 9.21e-16 ***
pm25mean              -0.060734   0.076630  -0.793 0.428350
so2                  -97.525069 164.904218  -0.591 0.554476
no2                   76.023885 142.156409   0.535 0.592996
no                     7.497306  48.425366   0.155 0.877015
ozone                 33.437176  86.359749   0.387 0.698759
mntp                   0.332376   0.091141   3.647 0.000289 ***
mnrh                   0.022920   0.046236   0.496 0.620277
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.99 on 590 degrees of freedom
Multiple R-Squared: 0.6251,     Adjusted R-squared: 0.6149
F-statistic: 61.47 on 16 and 590 DF,  p-value: < 2.2e-16
```

Listing 2: Model for 65 and over hospital admissions (admiss.old) - standardized coefficients.

```
> summary(fit.admiss <-
lm(admiss.old~dow+scale(tp)+scale(cos(2*pi*tp/365))+scale(sin(2*pi*tp/365))+scale(pm25mean)+
scale(so2)+scale(no2)+scale(no)+scale(ozone)+scale(mntp)+scale(mnrh),
+   data=complete.copy[!is.na(complete.copy$pm25mean),]))
```

```
Call:
lm(formula = admiss.old ~ dow + scale(tp) + scale(cos(2 * pi *
    tp/365)) + scale(sin(2 * pi * tp/365)) + scale(pm25mean) +
    scale(so2) + scale(no2) + scale(no) + scale(ozone) + scale(mntp) +
    scale(mnrh), data = complete.copy[!is.na(complete.copy$pm25mean),
    ])

Residuals:
     Min       1Q   Median       3Q      Max
-90.8484  -8.6963   0.7477   9.6822  61.7141

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  84.1106     1.9374  43.414  < 2e-16 ***
dowSun                       -2.8683     2.7322  -1.050 0.294236
dowMon                       49.7410     2.6437  18.815  < 2e-16 ***
dowTue                       45.1399     2.6685  16.916  < 2e-16 ***
dowWed                       40.6097     2.6495  15.327  < 2e-16 ***
dowThu                       37.5518     2.6076  14.401  < 2e-16 ***
dowFri                       33.0580     2.7540  12.003  < 2e-16 ***
scale(tp)                    -3.5279     0.8114  -4.348 1.62e-05 ***
scale(cos(2 * pi * tp/365))  10.0913     1.7348   5.817 9.82e-09 ***
scale(sin(2 * pi * tp/365))   7.7418     0.9366   8.266 9.21e-16 ***
scale(pm25mean)              -0.6219     0.7846  -0.793 0.428350
scale(so2)                   -0.5116     0.8650  -0.591 0.554476
scale(no2)                    0.5533     1.0347   0.535 0.592996
scale(no)                     0.1686     1.0893   0.155 0.877015
scale(ozone)                  0.4691     1.2116   0.387 0.698759
scale(mntp)                   5.8054     1.5919   3.647 0.000289 ***
scale(mnrh)                   0.3709     0.7481   0.496 0.620277
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.99 on 590 degrees of freedom
Multiple R-Squared: 0.6251,     Adjusted R-squared: 0.6149
F-statistic: 61.47 on 16 and 590 DF,  p-value: < 2.2e-16
```

Listing 3: PM$_{2.5}$ as a function of the other covariates.

```
> summary(fit.pm25 <-
lm(pm25mean~dow+tp+cos(2*pi*tp/365)+sin(2*pi*tp/365)+so2+no2+no+ozone+mntp+mnrh,data=complet
e.copy[!is.na(complete.copy$pm25mean),]))

Call:
lm(formula = pm25mean ~ dow + tp + cos(2 * pi * tp/365) + sin(2 *
    pi * tp/365) + so2 + no2 + no + ozone + mntp + mnrh, data =
complete.copy[!is.na(complete.copy$pm25mean),
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-17.645  -5.804  -1.418   4.290  47.671

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -7.482e+01  2.367e+01  -3.161 0.001655 **
dowSun               -2.059e+00  1.464e+00  -1.406 0.160161
dowMon               -1.289e+00  1.418e+00  -0.909 0.363826
```

```
dowTue                -2.804e+00  1.428e+00  -1.964 0.049974 *
dowWed                -9.798e-01  1.422e+00  -0.689 0.490966
dowThu                -1.192e+00  1.399e+00  -0.852 0.394612
dowFri                 4.626e-01  1.478e+00   0.313 0.754459
tp                     4.409e-03  2.119e-03   2.081 0.037862 *
cos(2 * pi * tp/365)  7.919e+00  1.300e+00   6.089 2.04e-09 ***
sin(2 * pi * tp/365)  2.615e+00  6.896e-01   3.792 0.000165 ***
so2                   -1.466e+02  8.831e+01  -1.660 0.097386 .
no2                    1.912e+02  7.590e+01   2.519 0.012019 *
no                    -6.076e+00  2.599e+01  -0.234 0.815262
ozone                  5.154e+00  4.636e+01   0.111 0.911510
mntp                   4.974e-01  4.444e-02  11.193  < 2e-16 ***
mnrh                  -2.002e-02  2.481e-02  -0.807 0.420013
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.119 on 591 degrees of freedom
Multiple R-Squared: 0.2265,     Adjusted R-squared: 0.2069
F-statistic: 11.54 on 15 and 591 DF,  p-value: < 2.2e-16

            Variance inflation factor = 1/(1-R-Squared) = 1/(1-0.2265) = 1.292825
```

Listing 4: Analysis of the residuals from the fitted linear regression model.

```
> sd(fit.admiss$resid)
[1] 16.76217
> arima(fit.admiss$resid,order=c(3,0,0))

Call:
arima(x = fit.admiss$resid, order = c(3, 0, 0))

Coefficients:
         ar1     ar2     ar3  intercept
      0.2976  0.225  0.2011    -0.6937
s.e.  0.0399  0.041  0.0412     2.0427

sigma^2 estimated as 195.1:  log likelihood = -2462.17,  aic = 4934.34
> sqrt(195.1)
[1] 13.96782
```

4.1.2 Theoretical Power Analysis Based on Regression

The theoretical power computations were performed using an on-line Java applet developed by Lenth and available at http://www.stat.uiowa.edu/~rlenth/Power/. The significance level α was chosen to be 0.05. A variance inflation factor of 1.3 was used assuming 16 independent variables along with 10 μg cm^{-3} for the PM$_{2.5}$ standard deviation and 17 μg cm^{-3} for the residual standard deviation. The power for a one-tailed test was computed for three time series lengths: 1,095 days (3 years), 1,460 days (4 years), and 1,825 days (5 years) for a PM$_{2.5}$ coefficient that ranges from 0.05 to 0.3 in steps of 0.005. **Listing 5** shows a typical set of values used in the power analysis. **Listing 6** shows the output from the power analysis for a three-yearrs series of daily values. **Table 50** shows the coefficient β and power values for each of the three time series lengths. **Figure 55** shows how the power increases with the increase in the beta coefficient for each of the three time series lengths. This figure shows that for *n* = 1,095, 80% power occurs for β = 0.15; for n = 1,460, 80% power occurs for β = 0.13; and for *n* = 1,825, 80% power occurs for β = 0.115. The R code for these estimates is shown in **Listings 7, 8, and .9**

Listing 5: Example input for power analysis.

```
# Assumptions:

# Dependent variable = admissions
# alpha = 0.05
# upper-tailed test
# beta for pm2.5 = 0.06
# no. of independent variables = p = 16
# R^2 between pm2.5 and 16 other covariates =  0.2265
# Variance inflation factor = 1/(1-R^2) = 1.29
# Length of admissions series = n = 4*365 = 1460
# sd of pm 2.5 = 10
# error sd = 16.93

> mean(pm25mean,na.rm=TRUE)
[1] 0.1130049
> mean(pm2.5,na.rm=TRUE)
[1] 16.22452
> sd(pm2.5,na.rm=TRUE)
[1] 10.24053
> sd(pm25mean,na.rm=TRUE)
[1] 10.23935


On-line Power Analysis Software:

http://www.stat.uiowa.edu/~rlenth/Power/
```

Listing 6: Power as a function of beta for 1,095 days (3 years).

```
# Power vs. Detectable beta[j]
#   Two-tailed: false
#   Solve for: Sample size
#   No. of predictors = 16
#   SD of x[j] = 10
#   VIF[j] = 1.3
#   Alpha = .05
#   Error SD = 17
#   Sample size = 1095
beta   power
.05    .21424
.055   .23995
.06    .26724
.065   .29603
.07    .32615
.075   .35746
.08    .38976
.085   .42284
```

```
.09     .45647
.095    .49042
.1      .52443
.105    .55827
.11     .59169
.115    .62446
.12     .65635
.125    .68716
.13     .71672
.135    .74487
.14     .77148
.145    .79646
.15     .81973
.155    .84126
.16     .86103
.165    .87905
.17     .89537
.175    .91002
.18     .92309
.185    .93467
.19     .94485
.195    .95373
.2      .96142
.205    .96804
.21     .97369
.215    .97848
.22     .98251
.225    .98588
.23     .98867
.235    .99097
.24     .99285
.245    .99438
.25     .99561
.255    .99659
.26     .99737
.265    .99799
.27     .99847
.275    .99884
.28     .99913
.285    .99935
.29     .99952
.295    .99965
.3      .99974
```

Table 50: Power as a function of true β coefficient and time series length based on multiple regression model.

β	**N=1,095**	**N=1,460**	**N=1,825**	β	**N=1,095**	**N=1,460**	**N=1,825**
0.050	0.21	0.25	0.29	0.155	0.84	0.92	0.96
0.055	0.24	0.29	0.33	0.160	0.86	0.93	0.97
0.060	0.27	0.32	0.37	0.165	0.88	0.95	0.98
0.065	0.30	0.36	0.42	0.170	0.90	0.96	0.98
0.070	0.33	0.40	0.46	0.175	0.91	0.96	0.99
0.075	0.36	0.43	0.50	0.180	0.92	0.97	0.99
0.080	0.39	0.47	0.55	0.185	0.93	0.98	0.99
0.085	0.42	0.51	0.59	0.190	0.94	0.98	0.99
0.090	0.46	0.55	0.63	0.195	0.95	0.99	1.00
0.095	0.49	0.59	0.67	0.200	0.96	0.99	1.00
0.100	0.52	0.63	0.71	0.205	0.97	0.99	1.00
0.105	0.56	0.66	0.75	0.210	0.97	0.99	1.00
0.110	0.59	0.70	0.78	0.215	0.98	1.00	1.00
0.115	0.62	0.73	0.81	0.220	0.98	1.00	1.00
0.120	0.66	0.76	0.84	0.225	0.99	1.00	1.00
0.125	0.69	0.79	0.87	0.230	0.99	1.00	1.00
0.130	0.72	0.82	0.89	0.235	0.99	1.00	1.00
0.135	0.74	0.84	0.91	0.240	0.99	1.00	1.00
0.140	0.77	0.87	0.93	0.245	0.99	1.00	1.00
0.145	0.80	0.89	0.94	0.250	1.00	1.00	1.00
0.150	0.82	0.91	0.95	0.255	1.00	1.00	1.00

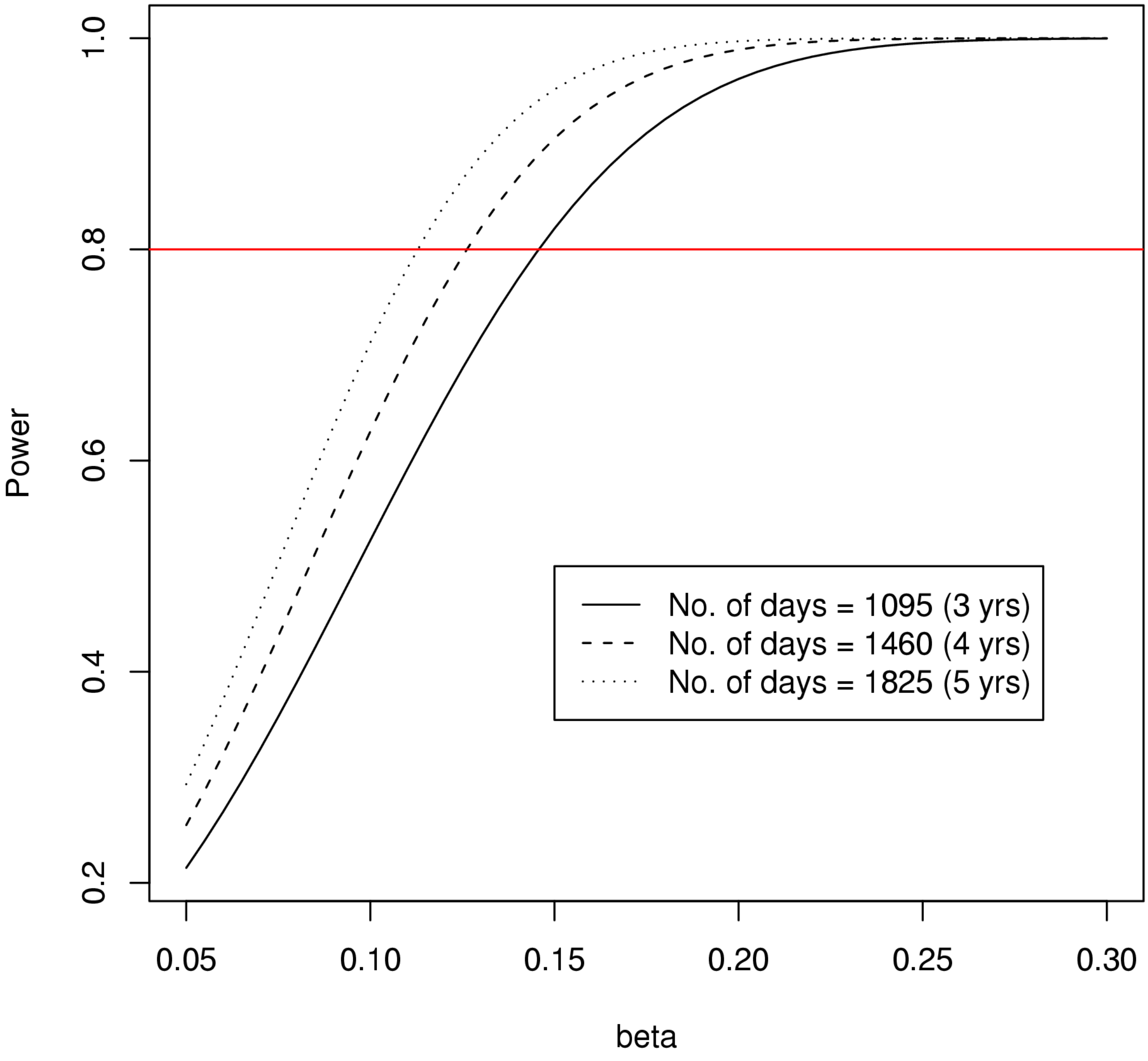


Figure 55: Power as a function of the number of days and the size of the true coefficient value.

4.1.3 Empirical Power Analysis Using Simulation

As an alternative to the approximate theoretical power computation discussed above, a simulation analysis was also performed. The simulation approach can easily include an autocorrelated error component which was ignored in the theoretical analysis. The simulation analysis requires modeling hospital admissions as a function of the independent variables and the residual error component. The modeling was based on the same data assembled from NMMAPS and ACAPS as discussed above. The intercorrelation of the independent variables and the error autocorrelation was modeled on the relationships observed in the assembled data set. The ACF and PACF graphs show that the error process can be represented as an autoregressive process of order 3 (AR(3)). An ARIMA model of order (p=3, d=0, q=0) with included independent variables was fitted to the error series and the estimated autoregressive coefficients for lags 1, 2, and 3 were approximately 0.3, 0.2, 0.2, respectively. The residual error standard deviation for this model was about 14. For each hypothesized value of beta (the coefficient for $PM_{2.5}$) from 0 to 0.15 in steps of 0.01, one thousand (1,000) statistically independent draws were made from the correlated error distribution resulting in a total of 48,000 estimated ARIMA models. The standard error of the estimated power $\hat{\pi}$ is:

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi \times (1-\pi)}{1{,}000}}.$$

For $\pi = 0.05$, the standard error is about 0.0069. For $\pi = 0.5$, the standard error is about 0.0158 and for $\pi =$ 0.8, the standard error is about 0.0126. Power as a function of β is tabulated in **Table 51** and graphed in **Figure 56.** As a check on the simulation, $\beta = 0$ was included. For $\beta = 0$, the power must equal the significance level, in this case 0.05. The simulation produced reasonably good estimates for $\beta = 0$. The average standard error for estimating β is shown in **Table 52**. Doubling the number of observations from 1,095 to 2,190 decreases the standard error by about 29%.The simulated power curves show the same general pattern as those based on theoretical considerations but show higher power. For N = 1,095 (three years), the power reaches 0.8 before $\beta = 0.11$ (compared to 0.15 for the theoretical analysis). For N = 1,460 (four years), the power reaches 0.8 just beyond $\beta = 0.09$ (compared to 0.13). Finally, for N = 1,825 (five years), the power reaches 0.8 before $\beta = 0.08$ (compared to 0.115). The increased power is due to the reduced residual error achieved by accounting for the information in the autocorrelated errors. This supports the idea that properly handling the autocorrelation in the residual errors, e.g., by using GLARMA modeling, can make a significant contribution to detecting smaller effects of explanatory factors.

Table 51. *Empirical power estimates for AR(3) error. Each estimate is based on 1,000 simulations.*

β	**Time Series Length (N)**		
	1,095	**1,460**	**1,825**
0.00	0.03	0.05	0.05
0.01	0.09	0.10	0.08
0.02	0.10	0.14	0.15
0.03	0.18	0.23	0.25
0.04	0.22	0.28	0.35
0.05	0.32	0.40	0.47
0.06	0.41	0.49	0.61
0.07	0.50	0.57	0.72
0.08	0.60	0.70	0.81
0.09	0.69	0.78	0.89
0.10	0.74	0.86	0.93
0.11	0.83	0.92	0.96
0.12	0.89	0.95	0.98
0.13	0.92	0.97	0.99
0.14	0.94	0.98	1.00
0.15	0.97	0.99	1.00

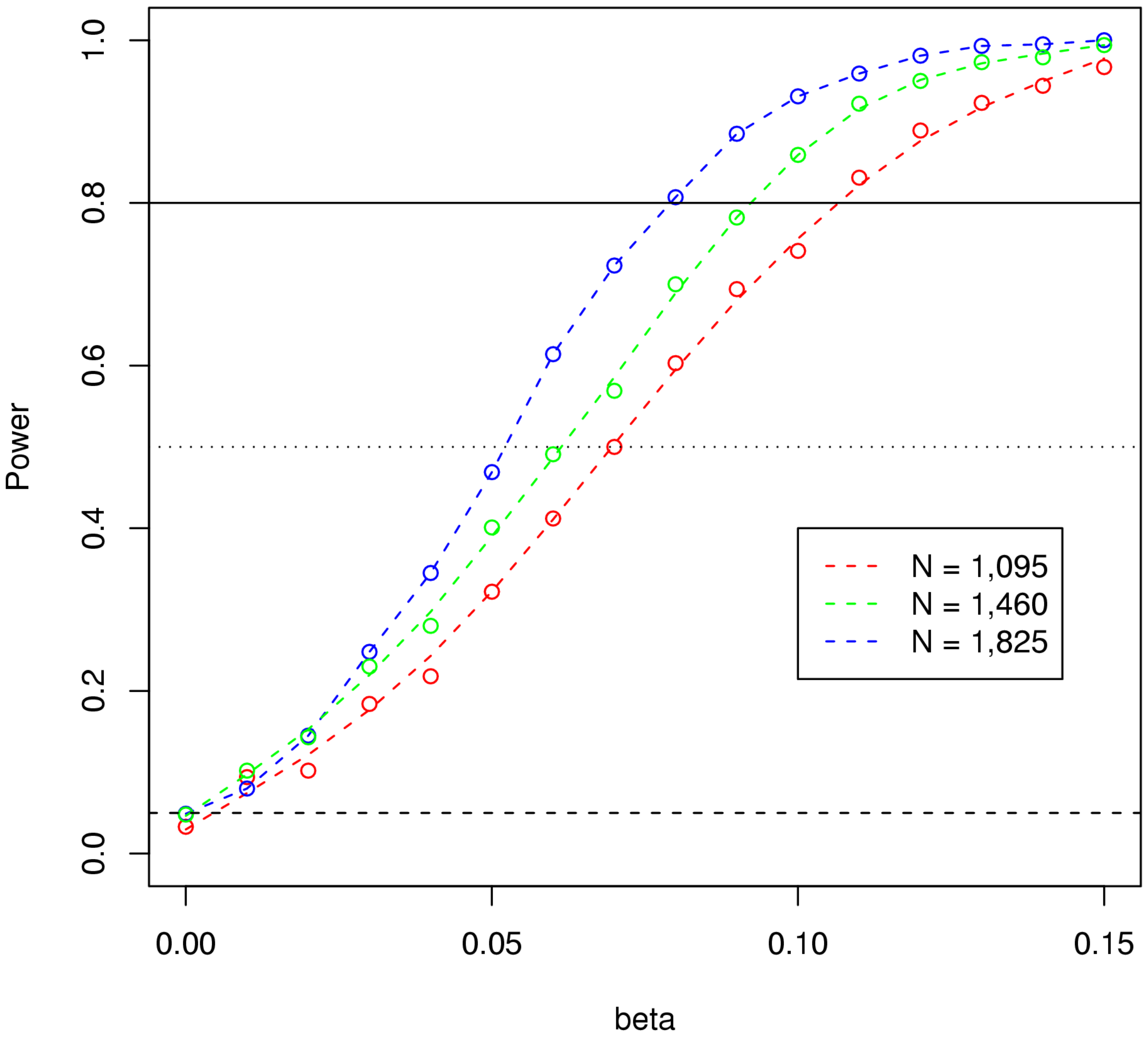


Figure 56. *Empirical power estimates based on 48,000 ARIMA simulations.*

Table 52. Approximate standard errors for β. Each estimate is based on 1,000 simulations.

Years	N	Standard Error for β	% Drop	Cumulative % Drop
2	730	0.0498	NA	NA
3	1,095	0.0425	14.6	14.6
4	1,460	0.0356	16.3	28.5
5	1,825	0.0325	8.7	34.7
6	2,190	0.0301	7.4	39.5

Listing 7: R function to sample from a multivariate normal distribution. Based on code from: ***http://maven.smith.edu/~nhorton/R/.***

```
# Simulated power for autocorrelated time series

rmultnorm <- function(n, mu, vmat, tol = 1e-07)
 # a function to generate random multivariate Gaussians
 {
    p <- ncol(vmat)
    if (length(mu)!=p)
       stop("mu vector is the wrong length")
    if (max(abs(vmat - t(vmat))) > tol)
       stop("vmat not symmetric")
    vs <- svd(vmat)
    vsqrt <- t(vs$v %*% (t(vs$u) * sqrt(vs$d)))
    ans <- matrix(rnorm(n * p), nrow = n) %*% vsqrt
    ans <- sweep(ans, 2, mu, "+")
    dimnames(ans) <- list(NULL, dimnames(vmat)[[2]])
    return(ans)
```

Listing 8: R code for plotting theoretical power estimates.

```
pwr <- read.table("/projects/PITT-PM/power/Data/power_beta.txt",
  sep="",header=TRUE)

postscript("/projects/PITT-PM/power/PS/power_beta2.eps",height=6.5,width=6.5,
  onefile=FALSE,horizontal=FALSE,paper="special")
plot(beta,n1095,type="l",lty=1,ylab="Power")
lines(beta,n1460,lty=2)
lines(beta,n1825,lty=3)
abline(h=0.8,col="red")
legend(0.15,0.5,c("No. of days = 1095 (3 yrs)",
  "No. of days = 1460 (4 yrs)","No. of days = 1825 (5 yrs)"),
  lty=c(1,2,3))
```

```
dev.off()
system("evince /projects/PITT-PM/power/PS/power_beta2.eps &")

data.frame(beta,"n=1095"=n1095,"n=1460"=n1460,"n=1825"=n1825)
```

Listing 9: R code to simulate time series model and compute power. Based on code from: http://maven.smith.edu/~nhorton/R/.

```
# Simulate and compute power

#attach(complete.copy)
vmat <- var(data.frame(pm25=pm25trend+pm25mean,so2,no2,no,ozone,mntp,mnrh)
  ,use="complete")

#N <- 1095
#N <- 1460
#N <- 1825

tpt <- seq(10508,10508+N-1,1)
cse <- cos(2*pi*tpt/365)
sne <- sin(2*pi*tpt/365)
dowt <- factor(as.character(weekdays(dates(tpt))),c("Sat","Sun","Mon","Tue","Wed","Thu","Fri"))

xmat <- rmultnorm(N, c(mean(pm25trend+pm25mean,na.rm=TRUE),mean(so2),mean(no2),mean(no),mean(ozone),
  mean(mntp),mean(mnrh)), vmat)
x1 <- xmat[,1]
x2 <- xmat[,2]
x3 <- xmat[,3]
x4 <- xmat[,4]
x5 <- xmat[,5]
x6 <- xmat[,6]
x7 <- xmat[,7]

numsim <- 1000
pow <- rep(NA,6)
beta <- 0.06

for(j in 1:6)
{

beta <- beta + 0.01

power <- rep(0,numsim)

cat("\n\nN =",N,"\nbeta =",beta,"\n")

for (i in 1:numsim) {
    cat("\r",i)
    cor.error <- arima.sim(model=list(order=c(3,0,0),ar=c(0.3,0.2,0.1)),n=N,sd=17)
    y <- 253.5+beta*x1-0.017222*tpt+14.528318*cse+10.747829*sne-2.868334*ifelse(dowt=="Sun",1,0)+
      49.741030*ifelse(dowt=="Mon",1,0)+45.139919*ifelse(dowt=="Tue",1,0)+40.609712*ifelse(dowt=="Wed",1,0
) +
      37.551810*ifelse(dowt=="Thu",1,0)+33.057982*ifelse(dowt=="Fri",1,0) -
      97.525069*x2 + 76.023885*x3 + 7.497306*x4 + 33.437176*x5 + 0.332376*x6 + 0.022920*x7

    y <- y+cor.error

# Estimate ARIMA model
    res <- arima(y,order=c(1,0,0),xreg=data.frame(x1,tpt,cse,sne,
      sun=ifelse(dowt=="Sun",1,0),mon=ifelse(dowt=="Mon",1,0),
      tue=ifelse(dowt=="Tue",1,0),wed=ifelse(dowt=="Wed",1,0),
      thu=ifelse(dowt=="Thu",1,0),fri=ifelse(dowt=="Fri",1,0),x2,x3,x4,x5,x6,x7))

    se.x <- sqrt(res$var.coef[3,3])
    obs.t.x <- res$coef[3]/se.x
    pval <- 1-pnorm((obs.t.x))
    power[i] <- pval<=0.05
}
pow[j] <- sum(power)/numsim

cat("\n\nempirical power for beta of ",beta,
    " is ",round(pow[j],3),".\n",sep="")
}

# Print power

# n = 1825
# ar=c(0.3,0.2,0.1)
data.frame(beta=seq(0.05,0.1,0.01),pow)
```

4.2 Calibration of Measurements Made by Different Methods

4.2.1 Introduction

It will be necessary to combine measurements made at multiple monitoring sites distributed in space and time in order to estimate daily values. This will be handled in an optimal manner using geostatistical techniques discussed below. When the same parameter is measured by different types of monitors using different analytical techniques it will be necessary calibrate the values before combining. The calibration needs to adjust the measurements for relative bias (i.e., systematic differences). The bias parameters are indirectly related to the imprecision (i.e., the amount of random measurement error) of each method.

Although the appropriate techniques for determining imprecision and bias of measurement methods are well-developed and easily available, they do not appear to be used routinely in air pollution research. This situation is not unique to air pollution research but also occurs in other "hard" sciences. Many of the techniques were developed for psychology, sociology, and economics where severe measurement problems are the rule. In reality, the measurement problems in the "hard" sciences are not that different although the conceptual problems and problems of identification are somewhat less severe. The most general formulation of the measurement error problem involves using latent variable (structural equation) modeling. In cases where there are only two measurements for each item in a set of items (like air parcels), the methods often attributed to Bland and Altman can be useful so long as the two methods have similar imprecision.

4.2.2 Measurement Error Model

Measurement error can often be represented by a linear model that relates the *n* true (theoretical) values to the observed *n m* measurements from *m* methods:

$$x_{ij}=\alpha_i+\beta_i\mu_j+\epsilon_{ij}$$

where x_{ij} denotes the measurement from the i^{th} method and the j^{th} item, μ_j is the true value for the j^{th} item, α_i and β_i are parameters that describe the (assumed linear) systematic error for the i^{th} method and ε_{ij} is a random error for the i^{th} method and j^{th} item and is assumed to be Normally distributed with mean zero and standard deviation σ_i. The parameter σ_i characterizes the imprecision of method i. In addition, the true values μ_j have mean $\bar{\mu}$ and standard deviation σ. In order to be able to estimate the parameters (denoted by the Greek letters), the number of methods (devices) m must be 2 or greater. The various parameters can be estimated using the method of moments and the method of maximum likelihood. The method of maximum likelihood typically provides better estimators. It should be emphasized that regressing x_{ij} on $x_{i'j}$ using ordinary linear regression provides biased estimators for the parameters. This bias can be very severe when $x_{i'j}$ has measurement error with $\sigma_{i'} > 0$. In particular, using linear regression will provide a calibration line that distorts the true systematic error (Ripley and Thompson, 1987).

4.2.3 Ideal Method for Determining Measurement Imprecision

The most efficient way to estimate the imprecision (the standard deviation of the random error component) would be to repeatedly measure the same item using the method of interest. Depending on what the item is and what the method is, this may be easy or it may be virtually impossible under routine conditions and/or very difficult even in laboratory settings. For air pollution research the item would be a (relatively small) parcel of air. (What is "small" would depend on the research.) In this special case:

$$\mu_j = \mu \; for \; j = 1 \, to \, n$$

In general, the total variance of the observed measurement x is:

$$\sigma_x^2 = \beta \, \sigma_\mu^2 + \sigma_\epsilon^2$$

because α and β. Because μ is now a constant, $\sigma_\mu^2 = 0$ so that the imprecision σ_ϵ is simply equal to σ_x which can be estimated by the sample standard deviation of the observed measurements s_x. Unfortunately, for air pollution research it would be very difficult or impossible to measure the same air parcel in the field over and over so that it will be necessary to measure a number of different air parcels taken at various points in time using two or more methods. Regardless of how many methods are used, fewer computational problems will be encountered if the μ_i are very similar so that σ_μ is small compared to the imprecision σ_{ϵ_i} for each method i.

4.2.4 Determining Bias for Two Methods of Approximately Equal Imprecision

If there are only two methods and the methods are of approximately the same imprecision, then an easy way to determine the relative bias and the common imprecision is to regress the differences of the paired measurements to their averages. This method, while not invented by Bland and Altman, has been championed by Bland and Altman as the proper way to determine how well two methods agree. (Bland and Altman and a number of other statisticians have warned repeatedly that using ordinary regression to regress one method on the other leads to a distorted view of the relative bias and should in practice never be used. This will be discussed in detail in a following section.) An advantage of the Bland-Altman approach is that it can be implemented using regression analysis programs. The disadvantage is that it cannot be used if the method imprecisions substantially differ and/or there are more than two methods. (In addition, as discussed in more detail below, the assumption of equal imprecision standard deviations is a relatively strong assumptions which can affect the estimates of the bias parameters.) Before using the Bland-Altman approach, the more general approach of latent variable modeling should be used to determine separate imprecision standard deviations and to check the affect of assuming equality on the bias parameter estimates.

4.2.5 Latent Variable Model for Measurement Error

Latent variable modeling provides a more general method for estimating the parameters in the measurement error model shown above. In this approach it is assumed that the measurements *x* are observed or manifest variables and are driven by the true values referred to as a latent (hidden) variable. The relationships are illustrated in.**Figure 57**.

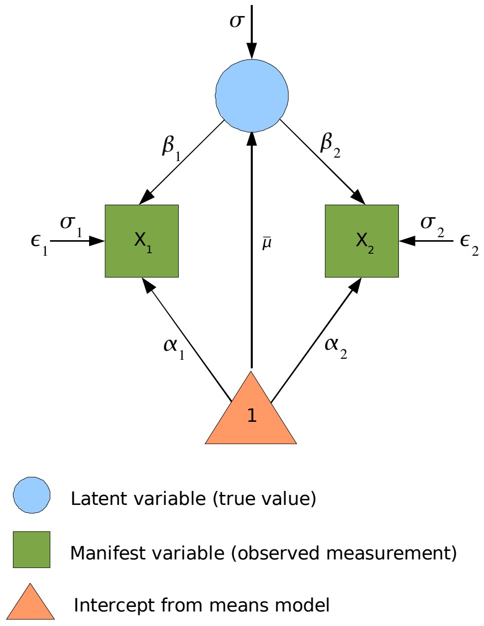


Figure 57: The (unobserved) latent variable α explains the observed measurements x_1 and x_2 via slope coefficients (β). The intercepts (α and $\bar{\mu}$) are represented by the latent variable denoted by the constant 1. The variability in the true values (free from measurement error) is characterized by σ while σ_1 and σ_2 characterize the method imprecisions.

4.2.6 Why Naive Regression Distorts Evaluation of Bias

The use of regression analysis for calibration when both methods have measurement error is known to distort the characterization of the actual bias and should be avoided. Three simulations are used to illustrate the problem. In the first simulation, there is a constant bias between the two methods. In the second simulation, there is no bias. Finally, in the third simulation, one of the methods has no measurement error.

4.2.6.1 Constant Bias Simulation Example

In this case: α_2=-2,α_1=2, β_1=β_2=1, σ_1=σ_2=1. X_1 and X_2 were simulated 100 times as multivariate Normal with means 10 and 12, respectively, and variances of 1 with a correlation of 0.77. The bias was thus constant and equal to 12-10 = 2. The true common imprecision would be the square root of (1-0.77) or about 0.48. In this case, the true relationship is:

$$X_2 = 2 + X_1,$$

or equivalently:

$$X_1 = -2 + X_2$$

An estimated calibration line based on observed data should give a result similar to the above relationship and not systematically differ. From the simulated data, the observed sample had means of 10.11 and 12.09 and standard deviations 1.08 and 1.11, respectively. The observed correlation was 0.80. Measurement error causes the magnitude of the correlation to be less than one. The separately estimated imprecisions for X_1 and X_2 were 0.46 and 0.53 and estimated common imprecision is 0.49. The estimated bias (X_2 - X_1) is 12.09 - 10.11 = 1.98.

Based on the Bland-Altman analysis shown in **Figure 58**, the estimated calibration line is:

$$X_2 = 1.65125 + 1.03273\ X_1,$$

or, equivalently

$$X_1 = -1.59892 + 0.96831\ X_2.$$

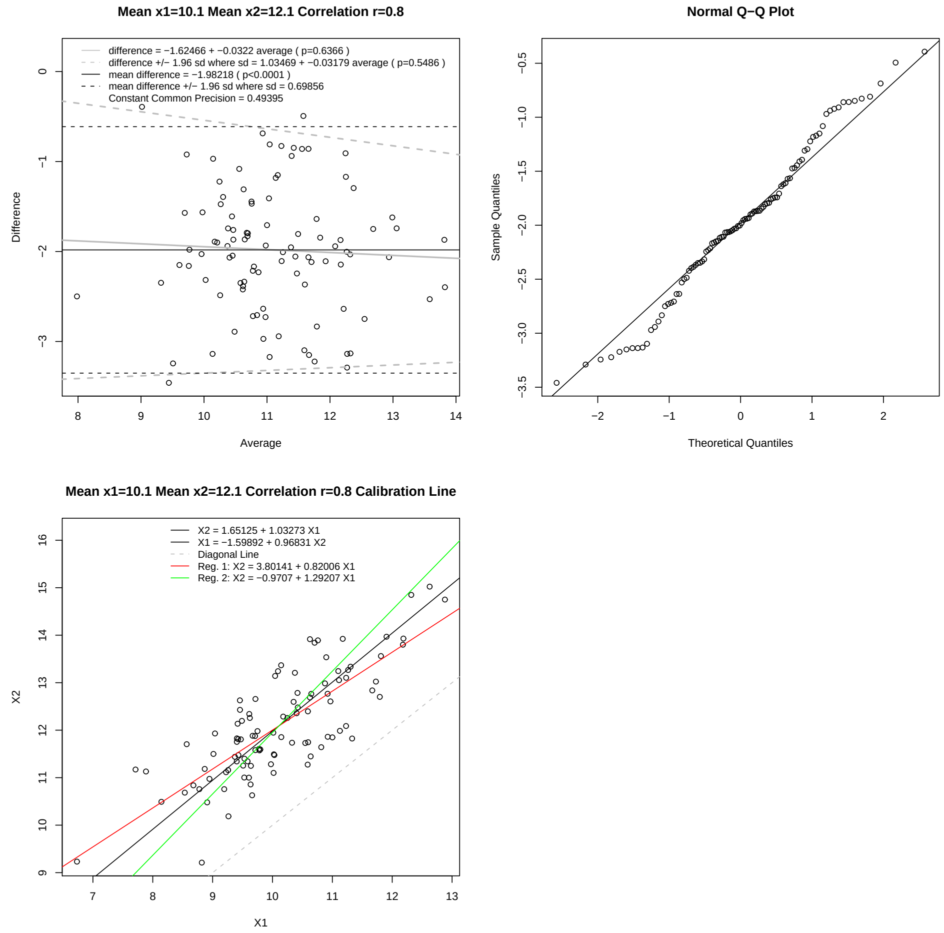


Figure 58: Simulated data for equal imprecision and constant bias case (α_2- α_1=2, β_1=β_2=1, σ_1=σ_2=1).

For example, if $X_1 = 10$, then $X_2 = 11.97855$. If $X_2 = 10$, then $X_1 = 8.08418$.

The relationship between the differences and the averages is:

$$X_1 - X_2 = -1.62466 - 0.0322\ [(X_1+X_2)/2]$$

The slope of 0.0322, however, is not statistically significantly different from zero (p=0.6366).

If we adopt the simpler model (constant bias, implies slope = 1) based on the observed mean difference (12.09 – 10.11 = 1.98) , then

$$X_2 = 1.98218 + X_1,$$

or, equivalently

$$X_1 = -1.98218 + X_2.$$

Over the range of the observed data (8 to 14), the simplified approximate calibration line:

$$X_2 = 1.98218 + X_1$$

and the original calibration line:

$$X_2 = 1.65125 + 1.03273\ X_1$$

are very similar. For $X_1 = 8$, the first equation yields $X_2 = 9.98$ while the second equation yields $X_2 = 9.91$.

The calibration plot Figure 58 includes the two different regression lines that relate X_1 to X_2: X_1 as a function of X_2 and X_2 as a function of X_1. Unlike the single calibration line (with two equivalent representations), there are two different regression lines whenever the absolute value of the correlation is less than one (the random measurement error will force the magnitude of the true correlation to be less than 1). Because the bias is constant with level, the true calibration line is parallel to the diagonal line. This is not true of the regression lines which are definitely not parallel to the diagonal line. Using the regression lines, we would incorrectly conclude that the bias was non-constant. Worse still, the nature of the bias would depend on which regression line was used. Thus, using regression naively for calibration data completely distorts the nature of the bias whenever there is measurement error.

When the correlation is near one, the two regression lines will still differ but will be very similar to each other and the correct calibration line. In general, unless the correlation is perfect and/or there is no measurement error (neither of which are realistic cases), direct regression should not be used for calibration.

Whenever the bias is constant (as it was in this example), the calibration line slope should be 1.0. Yet whenever the correlation is imperfect, the true regression slope must always be less than 1.0. The reason for this is that regression minimizes the sum of the squared errors so as to obtain the best prediction when the measurements are contaminated by measurement error. This, however, is not appropriate for calibration.

The regression of X_2 on X_1 yields:

$$\text{predicted } X_2 = 3.80141 + 0.82006\ X_1.$$

For $X_1 = 12$, X_2 is estimated to be 13.64213. The difference is 1.64213 which is substantially too small.

For $X_1 = 8$, X_2 is estimated to be 10.36189 and the difference is 2.36189 and is substantially too large. Thus, the bias appears to be non-constant which is a gross distortion of the nature of the actual bias in this case. Furthermore, if you use the other regression equation, the distortion is reversed. Thus, the use of naive regression here leads to complete nonsense.

4.2.6.2 No Bias Simulation Example

In this case: $\alpha_1=\alpha_2=0$, $\beta_1=\beta_2=1$, and $\sigma_1=\sigma_2=1.$ When the methods have the same amount of imprecision and there is no relative bias, we expect the points to fall along the diagonal line $X_2 = X_1$. In this case:

$$X_1=\mu+\epsilon_1$$

$$X_2=\mu+\epsilon_2$$

The two measurement methods are interchangeable. **Figure 60** illustrates the latent variable model. **Listing 12** shows the simulation results for a case where there is no bias. The true values ranged from 11, 12, ... , 20 and were identical for X_1 and X_2. The true standard deviation for X_1 and X_2 was about 2.886751. The imprecision SD's were both equal to one. The expected variances for X_1 and X_2 would be $2.886751^2 + 1^2 = 9.333333$. The expected means for X_1 and X_2 would be 15.5. The observed sample means were very similar and close to 15.5. The observed variances for X_1 and X_2 were about 8.295321 and 9.636726, respectively, and not too far from the expected 9.333333. The observed covariance was about 7.562288. The Grubbs type estimator for the imprecision variances would be 8.295321-7.562288 = 0.733033 for X_1 and 9.636726 – 7.562288 = 2.074438 for X_2. The resulting estimated imprecision sd's would then be about 0.856 and 1.440, respectively. The imprecision estimates illustrate the difficulty in precisely determining the imprecision. The agreement with the true imprecision sd's is not great even for a relatively large sample size.

The naive regression analysis is shown in **Listing 10** and the theoretical reason why the intercept estimate and slope estimate are wrong is illustrated in **Figure 59**. Note that estimated $\beta'=0.91163$. It is much lower than the true value of 1. (It is not quite statistically significantly different than 1, but with a larger sample size it would be.) The estimated intercept is 1.29011 but should be near zero. The sample correlation between X_1 and X_2 was about 0.85. The lower the correlation, the lower * will be.

Simultaneous estimates of all the parameters using maximum likelihood estimation for the latent variable model are shown in **Listing 13**. The model with constraints $\beta_1\beta_2=1$, $\sigma_1\geq 0$ and $\sigma_2\geq 0$ is shown in **Figure 60**. The Mx program output provides confidence intervals on all the parameters. The estimated relationships between the measured values and the true values are:

$$X_1 = -0.1918 + 1.0149$$

$$X_2 = 0.1884 + 0.9853$$

The intercepts are near zero (the 95% confidence intervals include zero) and the slopes are near one (the 95% confidence intervals include one). The calibration line relating X_1 to X_2 is:

$$X_1 = \alpha_1 - \alpha_2 \beta_1 / \beta_2 + \beta_1 / \beta_2 X_2$$

or equivalently

$$X_2 = \alpha_2 - \alpha_1 \beta_2 / \beta_1 + \beta_2 / \beta_1 X_1 .$$

For this data:

$$X_1 = -0.3859 + 1.0300\ X_2$$

$$X_2 = 0.3746 + 0.9708\ X_1.$$

The calibration line intercepts are much nearer zero and the slopes are nearer one than for the naive regression lines.

The estimated imprecision sd's are 0.7113 for X_1 and 1.5149 for X_2. The lower bound for the 95% confidence interval for both imprecision sd's is zero. The upper bound for X_1 is about 1.778 and for X_2 is about 1.917. These intervals are fairly wide but they do include the true imprecision sd values (1.0) and they largely overlap. Note that the estimate of the true sd for the measurements is 2.7500 which is close to the true sd of 2.8868.

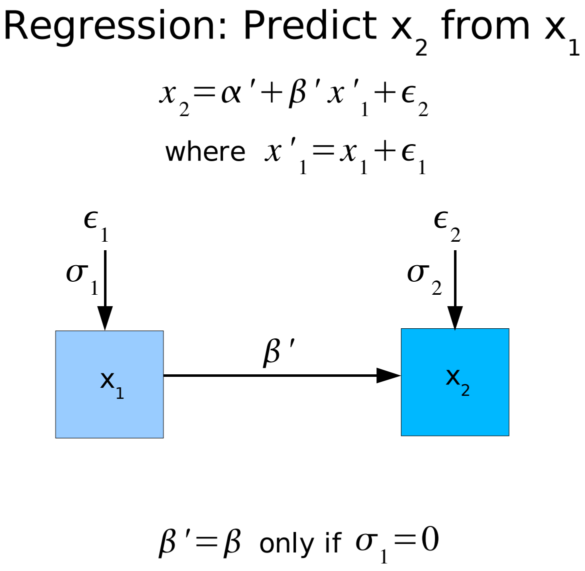

Regression: Predict x_2 from x_1
$x_2 = \alpha' + \beta' x'_1 + \epsilon_2$
where $x'_1 = x_1 + \epsilon_1$
ϵ_1
σ_1
ϵ_2
σ_2
β'
x_1
x_2
$\beta' = \beta$ only if $\sigma_1 = 0$

Figure 59: Regression model for predicting X_2 from X'_1 but ignoring the measurement error ϵ_1 in X'_1 .

shows the Bland-Altman plots and the resulting calibration line. As with the calibration line derived from the latent variable model, the Bland-Altman calibration line is much closer to the diagonal line than the regression lines. The advantages of the latent variable model compared to using the Bland-Altman method are 1) it does not need to assume equal imprecision sd's, 2) it can handle more than two methods, and 3) negative variance estimates can be avoided by using constrained optimization.

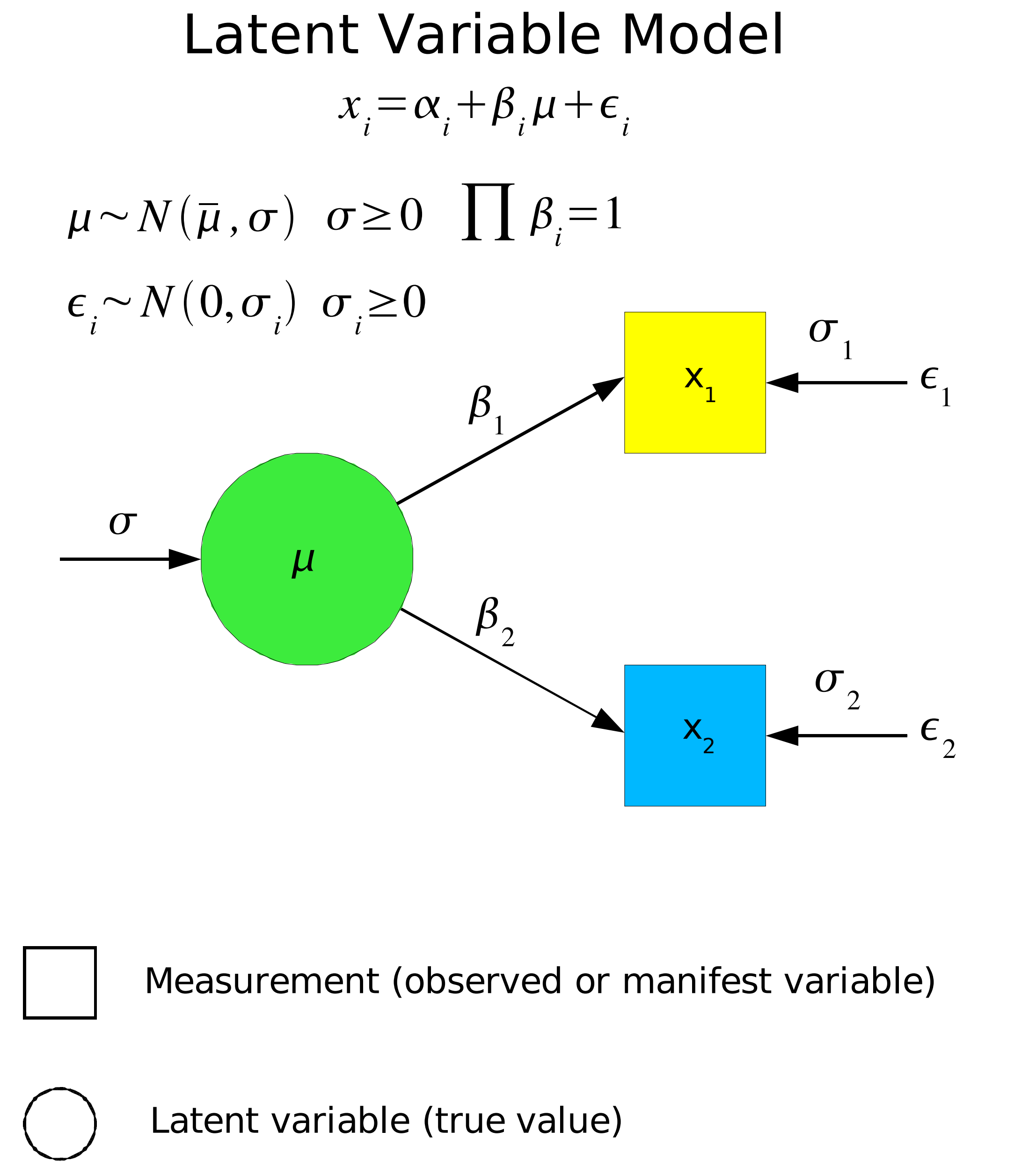

Latent Variable Model
$x_i = \alpha_i + \beta_i \mu + \epsilon_i$
$\mu \sim N(\bar{\mu}, \sigma) \quad \sigma \geq 0 \quad \prod \beta_i = 1$
$\epsilon_i \sim N(0, \sigma_i) \quad \sigma_i \geq 0$
σ
μ
β_1
β_2
x_1
x_2
σ_1
σ_2
ϵ_1
ϵ_2
Measurement (observed or manifest variable)
Latent variable (true value)

Figure 60: Latent variable model for two measurements with random error (diagram omits the means structure for simplicity).

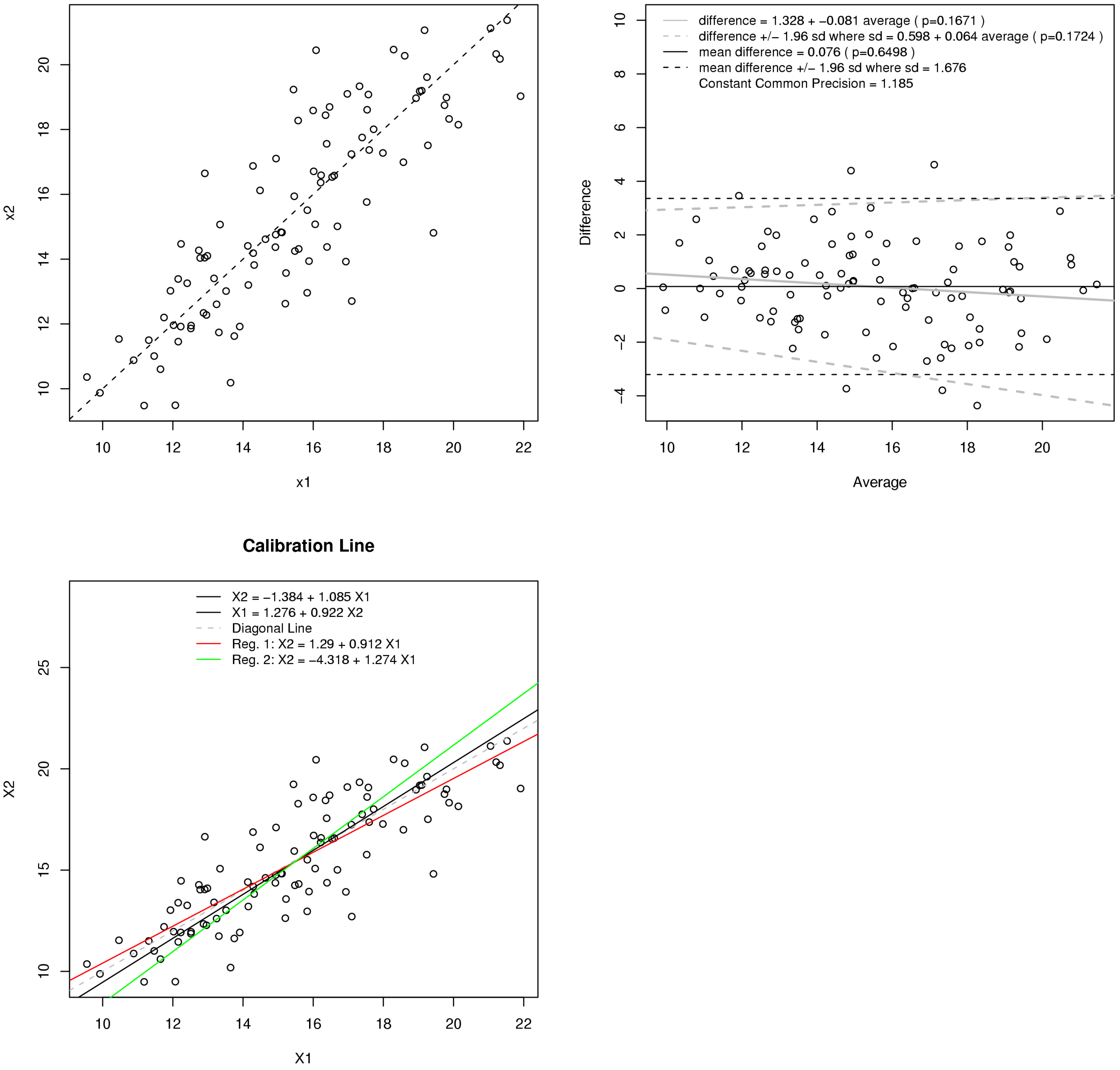


Figure 61: Scatter plot, Bland-Altman plot, and calibration for no relative bias example α=0, β=1 for both X_1 and X_2 and equal imprecision).

Listing 10: Annotated R code for simulation of no bias (α_1=α_2=0, β_1=β_2=1) and equal imprecision (σ_1=σ_2) case with sample size n = 100.

True SD = 2.886751

True imprecision for both methods = 1.0

True means for both methods = 15.5

True values = 11, 12, ... , 20 replicated 10 times

```
x <- 11:20
x1 <- x + rnorm(100)
hist(x1)
x2 <- x + rnorm(100)
df.no.bias <- data.frame(x1,x2)

> head(df.no.bias)
        x1       x2
1 12.15470 11.45415
2 12.95705 12.27728
3 12.74219 14.27082
4 12.02028 11.96076
5 15.83335 15.51070
6 14.28467 16.87499

# Variance-Covariance Matrix
> var(df.no.bias)
         x1       x2
x1 8.295321 7.562288 <- covariance
x2 7.562288 9.636726
# Note: sqrt(7.562288) = estimate of   (true sd = 2.89)
> sqrt(7.562288)
[1] 2.749961  # Very close to 2.89

# Imprecision sd for method 1 (using Grubbs method):
> sqrt(8.295321-7.562288)
[1] 0.8561735  # A little smaller than true value of 1

# Imprecision sd for method 2 (using Grubbs method):
> sqrt(9.636726-7.562288)
[1] 1.440291  # A little larger than true value of 1

> mean(df.no.bias)
      x1       x2
15.46293 15.38663 <-- Means almost exactly equal (little or no bias)
```

Listing 11: Regression of X_2 on X_1 for no bias example.

```
> summary(lm(x2~x1,data=df.no.bias))

Call:
lm(formula = x2 ~ x1, data = df.no.bias)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1895 -0.9350 -0.1195  0.8849  4.4947

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.29011    0.91345   1.412    0.161
x1         * 0.91163    0.05808  15.695   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.665 on 98 degrees of freedom
Multiple R-Squared: 0.7154,     Adjusted R-squared: 0.7125
F-statistic: 246.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

Listing 12: Latent variable model using the Mx program. The code constrains the imprecision sd's to being non-negative and the product of the β's equal to 1. The output is annotated.

```
  ** Mx startup successful **

   **MX-Linux version 1.64a**


 The following MX script lines were read for group    1

 #NGROUPS 3
  Note: #NGroup set number of groups to 3

 MEASUREMENT ERROR EXAMPLE - NO BIAS EXAMPLE - X1, X2
 ! TO RUN: > MXT < EXAMPLE_NO_BIAS.MX > EXAMPLE_NO_BIAS_OUTPUT.TXT
 DATA NOBSERVATIONS=100 NINPUT_VARIABLES=2
 CMATRIX FULL
 8.295321 7.562288
 7.562288 9.636726
 MEANS
 15.46293 15.38663
 BEGIN MATRICES;
 A FULL 2 2
 D DIAG 2 2
 X FULL 2 2
 K FULL 1 1
 V FULL 2 1
 L FULL 2 1
 END MATRICES;
```

```
SPECIFICATION A
1 2 3 4
SPECIFICATION D
6 7
SPECIFICATION X
0 0 0 11
MATRIX K 15.42478  ! THIS IS ARBITRARY BUT NECESSARY!
SPECIFICATION V
1 3
SPECIFICATION L
2 4
START .5 ALL
MEANS_MODEL V + L*K ;
COVARIANCE_MODEL A*X*A' + D ;
INTERVAL A 1 1 1
INTERVAL A 1 2 1
INTERVAL A 1 1 2
INTERVAL A 1 2 2
INTERVAL D 1 1 1
INTERVAL D 1 2 2
INTERVAL X 1 2 2
OPTIONS RSIDUALS
END


The following MX script lines were read for group    2

CONSTRAIN PRODUCT OF BETAS TO EQUAL 1
CONSTRAINT NINPUT_VARS=2
BEGIN MATRICES;
A FULL 2 2 = A1
Z STAN 1 1
B FULL 4 1
O ZERO 2 1
D DIAG 2 2 = D1
END MATRICES;
MATRIX B 1 2 2 2
CONSTRAINT \PROD(\PART(A,B)) = Z ; ! PRODUCT OF BETAS = 1
END


The following MX script lines were read for group    3

CONSTRAIN IMPRECISION VARIANCES > 0
CONSTRAINT NINPUT_VARS=2
BEGIN MATRICES;
A FULL 2 2 = A1
Z STAN 1 1
B FULL 4 1
O ZERO 2 1
D DIAG 2 2 = D1
END MATRICES;
MATRIX B 1 2 2 2
CONSTRAINT \D2V(D)' > O ;           ! IMPRECISION VARIANCES > 0
END

 PARAMETER SPECIFICATIONS
```

```
  GROUP NUMBER: 1

Measurement Error Example - No Bias Example - x1, x2

  MATRIX A
 This is a FULL matrix of order    2 by    2
    1 2
 1  1 2
 2  3 4

  MATRIX D
 This is a DIAGONAL matrix of order    2 by    2
    1 2
 1  6
 2  0 7

  MATRIX K
 This is a FULL matrix of order    1 by    1
  It has no free parameters specified

  MATRIX L
 This is a FULL matrix of order    2 by    1
    1
 1  2
 2  4

  MATRIX V
 This is a FULL matrix of order    2 by    1
    1
 1  1
 2  3

  MATRIX X
 This is a FULL matrix of order    2 by    2
     1  2
 1   0  0
 2   0 11

  GROUP NUMBER: 2

Constrain product of betas to equal 1

  MATRIX A
 This is a FULL matrix of order    2 by    2
    1 2
 1  1 2
 2  3 4

  MATRIX B
 This is a FULL matrix of order    4 by    1
  It has no free parameters specified

  MATRIX D
 This is a DIAGONAL matrix of order    2 by    2
    1 2
 1  6
 2  0 7

  MATRIX O
 This is a NULL matrix of order    2 by    1

  MATRIX Z
 This is a STANDARDISED matrix of order    1 by    1
  It has no free parameters specified

  GROUP NUMBER: 3

Constrain imprecision variances > 0

  MATRIX A
 This is a FULL matrix of order    2 by    2
    1 2
 1  1 2
 2  3 4

  MATRIX B
 This is a FULL matrix of order    4 by    1
  It has no free parameters specified

  MATRIX D
 This is a DIAGONAL matrix of order    2 by    2
    1 2
 1  6
 2  0 7

  MATRIX O
 This is a NULL matrix of order    2 by    1

  MATRIX Z
 This is a STANDARDISED matrix of order    1 by    1
  It has no free parameters specified

  Mx starting optimization; number of parameters =  7


  MX PARAMETER ESTIMATES

  GROUP NUMBER: 1

Measurement Error Example - No Bias Example - x1, x2

  MATRIX A
 This is a FULL matrix of order    2 by    2
            1          2
 1     -0.1918     1.0149
 2      0.1884     0.9853

  MATRIX D
 This is a DIAGONAL matrix of order    2 by    2
            1          2
 1      0.5059
 2      0.0000     2.2950
```

```
  MATRIX K
 This is a FULL matrix of order    1 by    1
             1
 1     15.4248

  MATRIX L
 This is a FULL matrix of order    2 by    1
             1
 1      1.0149
 2      0.9853

  MATRIX V
 This is a FULL matrix of order    2 by    1
          1
 1  -0.1918
 2   0.1884

  MATRIX X
 This is a FULL matrix of order    2 by    2
             1          2
 1      0.0000     0.0000
 2      0.0000     7.5623

  Vector of OBSERVED means
                1          2
 Mean     15.4629    15.3866

  Vector of EXPECTED means
                1          2
 Mean     15.4629    15.3866

  OBSERVED COVARIANCE MATRIX
             1          2
 1      8.2953
 2      7.5623     9.6367


  EXPECTED COVARIANCE MATRIX
             1          2
 1      8.2953
 2      7.5623     9.6367

  RESIDUAL MATRIX
               1            2
 1   -4.1456E-06
 2   -2.3494E-07   2.9528E-06

 Function value of this group:    7.2463E-11
  Where the fit function is Maximum Likelihood

  GROUP NUMBER: 2

Constrain product of betas to equal 1

  MATRIX A
 This is a FULL matrix of order    2 by    2
             1          2
 1     -0.1918     1.0149
 2      0.1884     0.9853

  MATRIX B
 This is a FULL matrix of order    4 by    1
             1
 1      1.0000
 2      2.0000
 3      2.0000
 4      2.0000

  MATRIX D
 This is a DIAGONAL matrix of order    2 by    2
             1          2
 1      0.5059
 2      0.0000     2.2950

  MATRIX O
 This is a NULL matrix of order    2 by    1

  MATRIX Z
 This is a STANDARDISED matrix of order    1 by    1
          1
 1   1.0000

  GROUP NUMBER: 3

Constrain imprecision variances > 0

  MATRIX A
 This is a FULL matrix of order    2 by    2
             1          2
 1     -0.1918     1.0149
 2      0.1884     0.9853

  MATRIX B
 This is a FULL matrix of order    4 by    1
             1
 1      1.0000
 2      2.0000
 3      2.0000
 4      2.0000

  MATRIX D
 This is a DIAGONAL matrix of order    2 by    2
             1          2
 1      0.5059
 2      0.0000     2.2950

  MATRIX O
 This is a NULL matrix of order    2 by    1

  MATRIX Z
 This is a STANDARDISED matrix of order    1 by    1
          1
 1   1.0000
 Your model has    7 estimated parameters and      8 Observed statistics
 Observed statistics include   3 constraints.

 Chi-squared fit of model >>>>>>>     0.000
 Degrees of freedom >>>>>>>>>>>>>         1
 Probability >>>>>>>>>>>>>>>>>>>>     1.000
 Akaike's Information Criterion >    -2.000
 RMSEA >>>>>>>>>>>>>>>>>>>>>>>>>>     0.000


 7  Confidence intervals requested in group  1

  Matrix Element Int.       Estimate          Lower          Upper  Lfail Ufail


  A    1    1    1  95.0    $\alpha_1$  -0.1918        -1.9344          2.8231 0 1    0 1

  A    1    2    1  95.0    $\alpha_2$   0.1884        -3.3647          1.7631 0 1    0 1

  A    1    1    2  95.0    $\beta_1$   1.0149          0.8284          1.1200 0 0    0 0

  A    1    2    2  95.0    $\beta_2$   0.9853          0.8929          1.2072 0 0    0 0

  D    1    1    1  95.0    $\sigma_1^2$  0.5059          0.0000          3.1620 1 1    0 1

  D    1    2    2  95.0    $\sigma_2^2$  2.2950          0.0000          3.6734 1 0    0 1

  X    1    2    2  95.0    $\sigma^2$    7.5623          5.6215         10.3682 0 1    0 1
 This problem used  0.0% of my workspace

 Task                      Time elapsed (DD:HH:MM:SS)
 Reading script & data       0: 0: 0: 0.01
 Execution                   0: 0: 0: 0.62
 TOTAL                       0: 0: 0: 0.63

 Total number of warnings issued: 0
 ______________________________________________________________________________
  Expand =  0

 ______________________________________________________________________________

Based on the SEM results:

Estimated imprecision SD for method 1: sqrt(0.5059) = 0.7112665

Estimated imprecision SD for method 2: sqrt(2.2950) = 1.514926

Estimated true SD of measurements: sqrt(7.5623) = 2.749964

Based on the estimates of  and , the calibration line using SEM is:


Slope:  0.9853/1.0149 = 0.9708346
```

```
Intercept:  0.1884 - (-0.1918)*0.9853/1.0149 = 0.3746061

Calibration equation:

x2 = 0.3746061 + 0.9708346*x1

This is similar to the Bland-Altman result and even closer to the true calibration
line:

x2 = 0 + 1*x1

The "calibration line" estimated by the linear regression is much farther away from
the true calibration line.

Note: Although the confidence intervals for the imprecisions substantially overlap
(leaving open the possibility that the imprecisions are equal or at least similar),
they are very wide even with a sample of size n = 100.
```

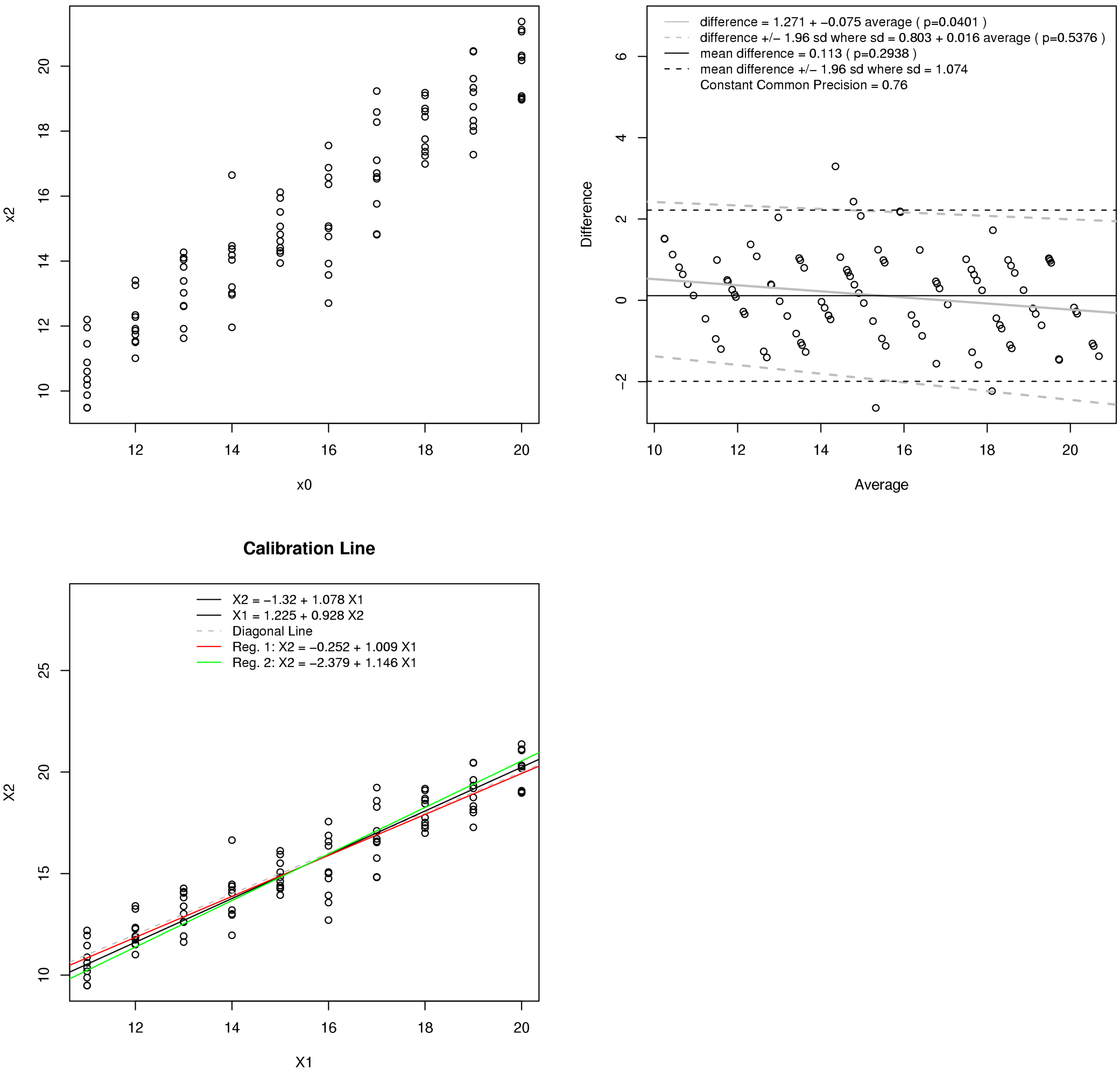


Figure 62: Scatter plot, Bland-Altman plot, and calibration for no measurement error in the independent variable X_1 (σ_1=0) (and no bias: α=0 and β=1 for both methods).

4.2.6.3 No-Measurement-Error Simulation Example

Figure 62 illustrates the case where one variable (X_1) has no measurement error. The vertical banding of the points in the scatter plots is due to the lack of measurement error. The dashed line indicating the diagonal (slope equals one) is difficult to see but the slopes of both regression lines are greater than one. The red line (regression 1 with estimated slope equal to 1.009) is the regression of X_2 on X_1 and would be the correct estimated calibration line because X_1 has no measurement error. The estimated slope is very close to one. The Bland-Altman line for X_2 as a function of X_1 (slope equals 1.078) is too steep. Represented as X_1 as a function of X_2, the slope (equal to 0.928) is too shallow.

The "no measurement error" case is rare to nonexistent. Theoretically, if X_1 had no measurement error, then the correct calibration line would be equal to the simple linear regression line and the Bland-Altman analysis would be incorrect and misleading. The Bland-Altman analysis assumes that the two methods have (at least roughly) the same imprecision. You can see the downward tilt in the upper right graph of **Figure 62**. Under the assumption of equal imprecision SD's, the Bland-Altman plot shows a bias. Given that the imprecision SD's are very different ($_1$=0, $_2$=1), the Bland-Altman analysis should not be used. Whenever the imprecisions are greatly different, the Bland-Altman analysis should not be used. These problems can be avoided by using an appropriate latent variable model. To summarize:

- For two or more methods measuring the same theoretical quantity to be in agreement, there must be little or no bias (systematic error) and the amount of imprecision (random error) should be small. When there is a substantial bias, the methods need to be calibrated. Bias is often not constant but may depend on the level of the quantity being measured. Imprecision is measured as a variance or standard deviation (SD) and also may change with level (typically increasing with level).

- For real world data, simple regression analysis (one method regressed on another method) should ***never*** be used for calibration. ***The resulting regression coefficients and correlation coefficients are completely useless and worse misleading when used in calibration.*** Not withstanding this, you will still see many recent examples of incorrect calibration in the literature.

- If the method imprecisions are at least roughly equal and you are comparing only two methods, then a Bland-Altman analysis could be used. The correct calibration line can be derived from the Bland-Altman analysis. Caveat: You may need to use the latent variable model first to estimate the method imprecisions before you have evidence that they are similar. Also, the assumption of exactly equal imprecision SDs is a fairly strong assumption and could distort estimation of the bias parameters.

- If the method imprecisions differ, or if there are more than two methods, then a latent variable model (a.k.a. structural equation model) should be used to determine the calibration line. The analysis will also determine the method imprecisions.

- When designing a study to measure method imprecision, the experimental units should vary as little as possible.

- Method imprecision can be computed from the variance-covariance matrix or by using a latent variable model. A satisfactory analysis may be impossible if the experimental units vary a great deal. When experimental units vary over a large range (compared to the size of the imprecision), negative variance estimates are likely along with extremely wide confidence intervals for the imprecision. If there are only two paired measurements and the number of paired measurements is small, the negative imprecision variance could be set to zero (and not changing the other imprecision variance) or you could use a method of constrained estimation (which will move the negative imprecision variance to zero or above and increase the other imprecision variance accordingly). Neither is wholly satisfactory in most cases. If there are enough paired samples, the data set could be broken into smaller subsets based on the averages for each pair (NOT on the value of just one method which would lead to a substantial bias). If the subsets are homogeneous enough, the covariances in each variance-covariance matrix will be smaller than both variances avoiding a negative imprecision variance. The separate imprecision variance estimates can then be plotted as a function of the average level. If the imprecision variance is roughly constant across levels, the separate variance-covariance matrices could be averaged (even if a few lead to a negative variance) and the overall imprecision variances computed.

4.3 Space-Time Geostatistical Analysis

4.3.1 Introduction

Measurements from multiple monitoring sites distributed in space and time will need to be combined using a weighted average in order to estimate the daily exposure. The daily estimation error for day *i* and region *j* is the difference between the estimated average for day *i* and region *j* $\hat{A}_{ij}$ and the true average value for the day and region A_{ij} :

$$e_{ij} = \hat{A}_{ij} - A_{ij}$$

The estimation error is characterized probabilistically by determining the mean, standard deviaiton, and shape of the distribution. Finding the optimal weights for $\hat{A}_{ij}$ (weights that minimize the estimation error

e_{ij}) requires characterizing and modeling the space-time correlation structure for the exposure measurements. This approach is illustrated using $PM_{2.5}$ mass concentrations. **Table 54** lists available monitoring sites for $PM_{2.5}$ mass from 1/1/2001 to 12/31/2003. Latitude and longitude have been converted into easting and northing coordinates to facilitate computation of distances between monitors. **Figure 65** shows the spatial arrangement of the monitoring sites. **Figure 64** shows the concentrations over time for each monitoring site. If the measurement values were entirely random, then a simple unweighed average (i.e., an average where the weights all equal 1) would be optimal. Typically, the correlation between measurements increases as the distance decreases but at different rates depending on the parameter being measured and also often shows seasonal patterns over time.

4.3.2 Characterizing Space-Time Dependence Using the Sample Variogram

Figures 66-72 show the spatial variance (the inverse of correlation) as characterized by the space-time variogram for $PM_{2.5}$ mass concentration. **Figure 65** shows a contour plot of the space-time variogram. For this function:

$$f(x,y)=f(-x,-y).$$

Darker shading indicates lower variance (higher correlation). The variogram along the horizontal line through the origin shows the variability through time (all points are at the same location but differ by date). The variogram along the vertical line through the origin shows the variability through space (all points are at the same time but differ in location). **Figure 66** shows cross sections of the space-time variogram and facilitates interpretation. The pure-time variogram shows a clear seasonal pattern - starting at about 75 $\mu g^2/m^6$ near the origin and then the variance reaches a maximum of about 100 $\mu g^2/m^6$ at 180 days of separation and drops back to 75 $\mu g^2/m^6$ at about 365 days. The pure-space variogram shows that the variance increases quickly for differences then tends to level off. The pure-space variogram at a separation distance near zero is about 60 $\mu g^2/m^6$ and by about 50 days reaches about 80 $\mu g^2/m^6$. At a space-time angle of 112.5° (in terms of the units used - a separation in terms of both space and time), the seasonality is actually more pronounced. **Figures 67 and 68** show three-dimensional renderings of the space time variogram. To reduce the effect of extreme values (due to the skewed $PM_{2.5}$ mass concentration distribution) in the variogram, the variogram was also computed for log $PM_{2.5}$ mass concentration. **Figures 69-72** shows similar relationships as **Figures66-68**, respectively, but the relationships are noticeably smoother.

The space-time geostatistical analysis for $PM_{2.5}$ mass concentration clearly shows definite patterns in the correlation structure which would need to be taken account of to produce optimal weighted daily averages. A similar analysis would need to be performed for each exposure variable.

Table 53: Coordinates for PM$_{2.5}$ monitoring sites from 1/1/2001 to 12/31/2003.

Site	Latitude	Longitude	Easting	Northing	Map ID
EPA390810016	40.36278	-80.61556	532641.7	4468094	a
EPA390810017	40.36610	-80.61500	532687.2	4468464	b
EPA390811001	40.32194	-80.60639	533440.2	4463565	c
EPA390990005	41.11111	-80.64528	529782.5	4551152	d
EPA390990014	41.09587	-80.65843	528685.3	4549456	e
EPA391550007	41.21417	-80.78750	517813.6	4562554	f
EPA420030008	40.46556	-79.96111	588074.8	4479950	g
EPA420030021	40.41361	-79.94139	589816.0	4474204	h
EPA420030064	40.32361	-79.86833	596142.2	4464290	i
EPA420030067	40.38194	-80.18556	569132.2	4470469	j
EPA420030093	40.60722	-80.02083	582837.0	4495617	k
EPA420030095	40.48694	-80.18806	568812.8	4482122	l
EPA420030116	40.47361	-80.07722	578221.7	4480735	m
EPA420030131	40.28944	-80.00500	584573.9	4460358	n
EPA420030133	40.26013	-79.88650	594687.5	4457224	o
EPA420031008	40.61861	-79.72722	607658.6	4497199	p
EPA420031301	40.40250	-79.86028	596713.8	4473056	q
EPA420033007	40.29444	-79.88667	594625.3	4461033	r
EPA420039002	40.54694	-79.78389	602975.1	4489176	s
EPA420070014	40.74778	-80.31667	557688.1	4510983	t
EPA420210011	40.30972	-78.91500	677174.7	4464220	u
EPA420850100	41.21500	-80.48500	543171.3	4562752	v
EPA421250005	40.14667	-79.90222	593506.2	4444614	w
EPA421250200	40.17056	-80.26139	562890.8	4446949	x
EPA421255001	40.44528	-80.42083	549115.0	4477342	y
EPA421290008	40.30469	-79.50567	626989.4	4462648	z
EPA540090005	40.33806	-80.59722	534210.9	4465357	A
EPA540290011	40.39450	-80.61203	532925.3	4471617	B
EPA540291004	40.42154	-80.58090	535553.5	4474630	C
EPA540490006	39.48083	-80.13528	574368.9	4370494	D
EPA540511002	39.91597	-80.73406	522728.6	4418465	E
EPA540610003	39.64944	-79.92111	592563.9	4389406	F
EPA540690008	40.06383	-80.72050	523835.5	4434879	G

Site	Latitude	Longitude	Easting	Northing	Map ID
PAQS-Schenley	40.43950	-79.94050	589856.9	4477078	H
DOE-Bruceton	40.31806	-79.98000	586662.3	4463559	I
SCAMP-Steub	40.36694	-80.64667	529998.4	4468546	J
SCAMP-North	40.53139	-80.58028	535548.1	4486824	K
SCAMP-South	40.07500	-80.69694	525840.5	4436126	L
SCAMP-East	40.31222	-79.38278	637417.8	4463667	M
SCAMP-West	40.32333	-80.89889	508590.0	4463650	N

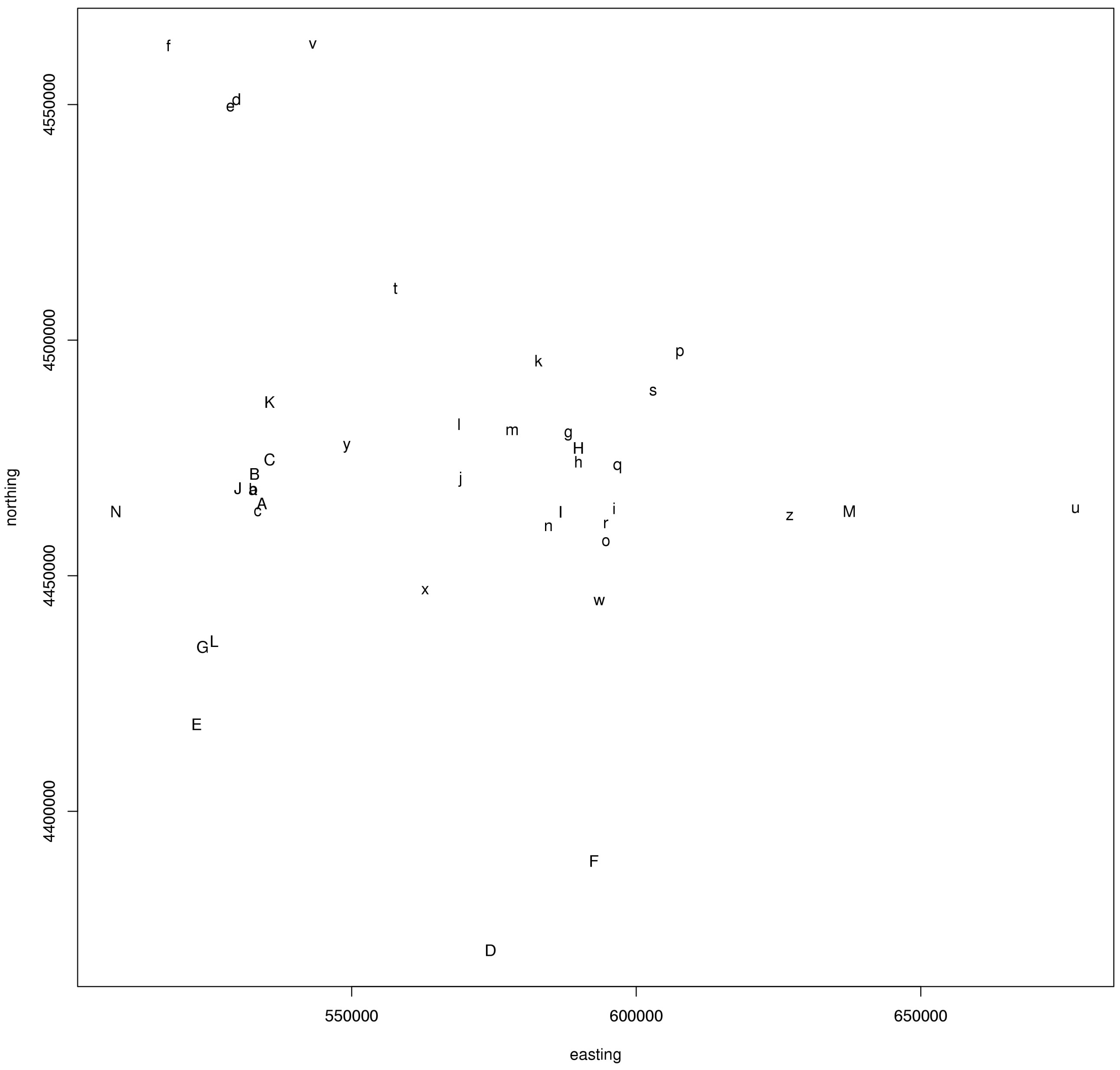


Figure 63: Spatial arrangement of $PM_{2.5}$ monitoring sites.

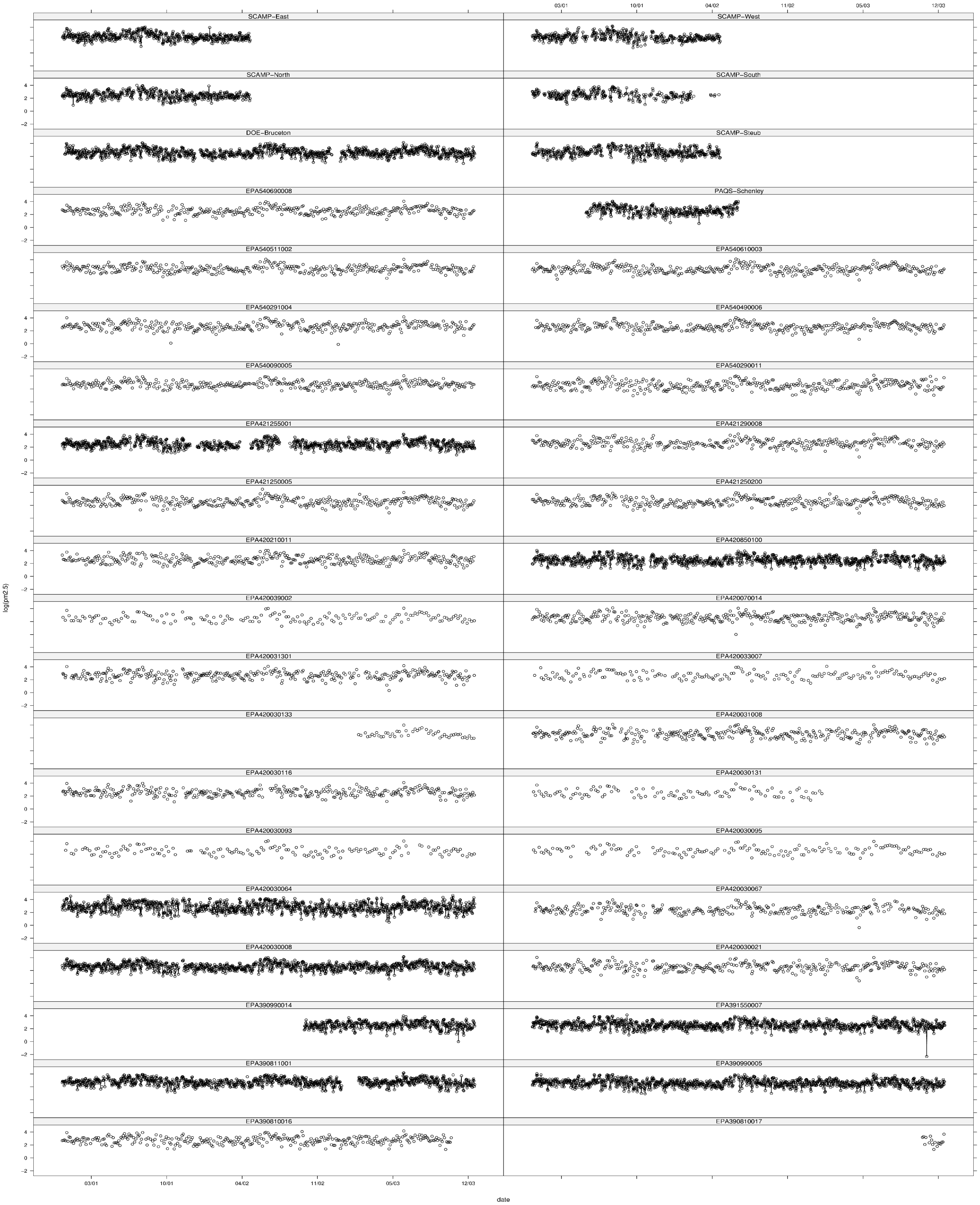


Figure 64: Log $PM_{2.5}$ concentrations from 1/1/01 to 12/31/03 for each monitoring site.

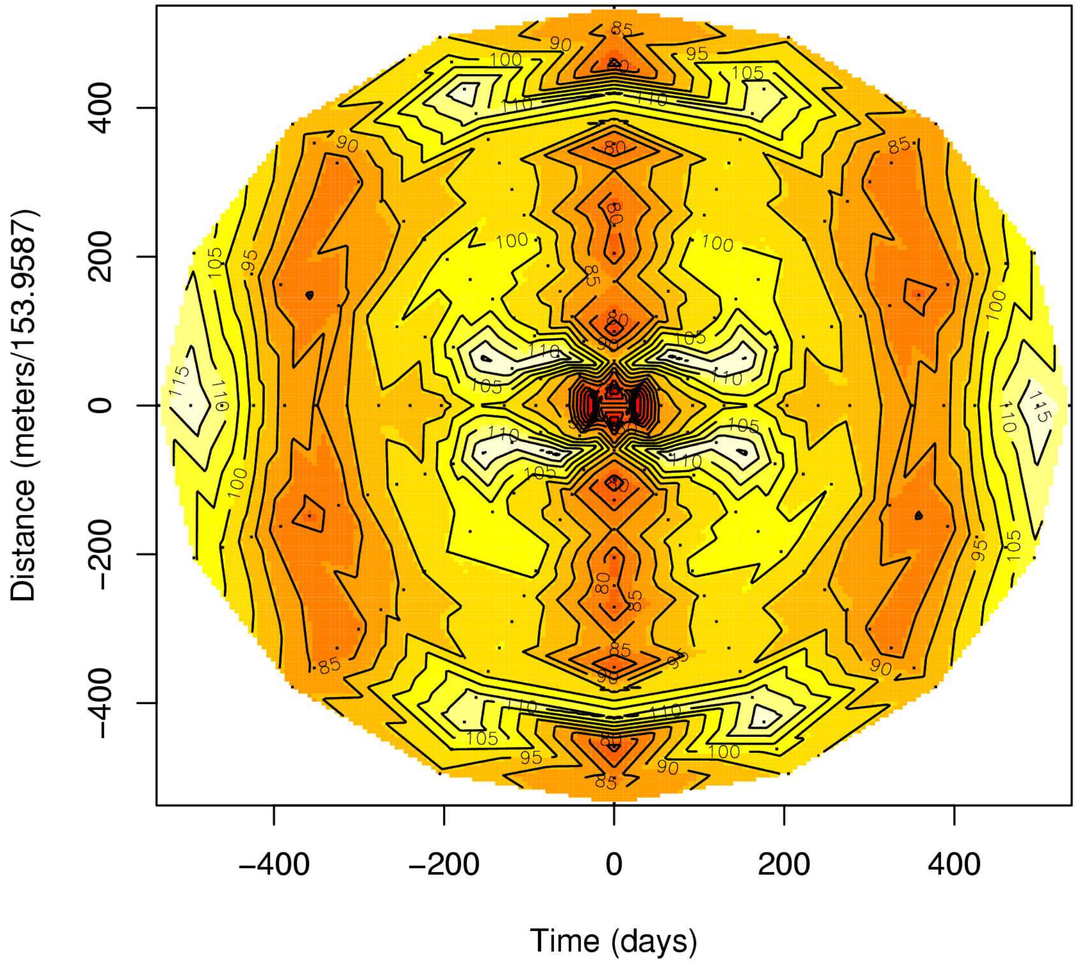


Figure 65: Contour and image map of the space-time semi-variogram for $PM_{2.5}$. Black points denote the locations at which the semi-variance was computed. Lighter shades denote higher variances

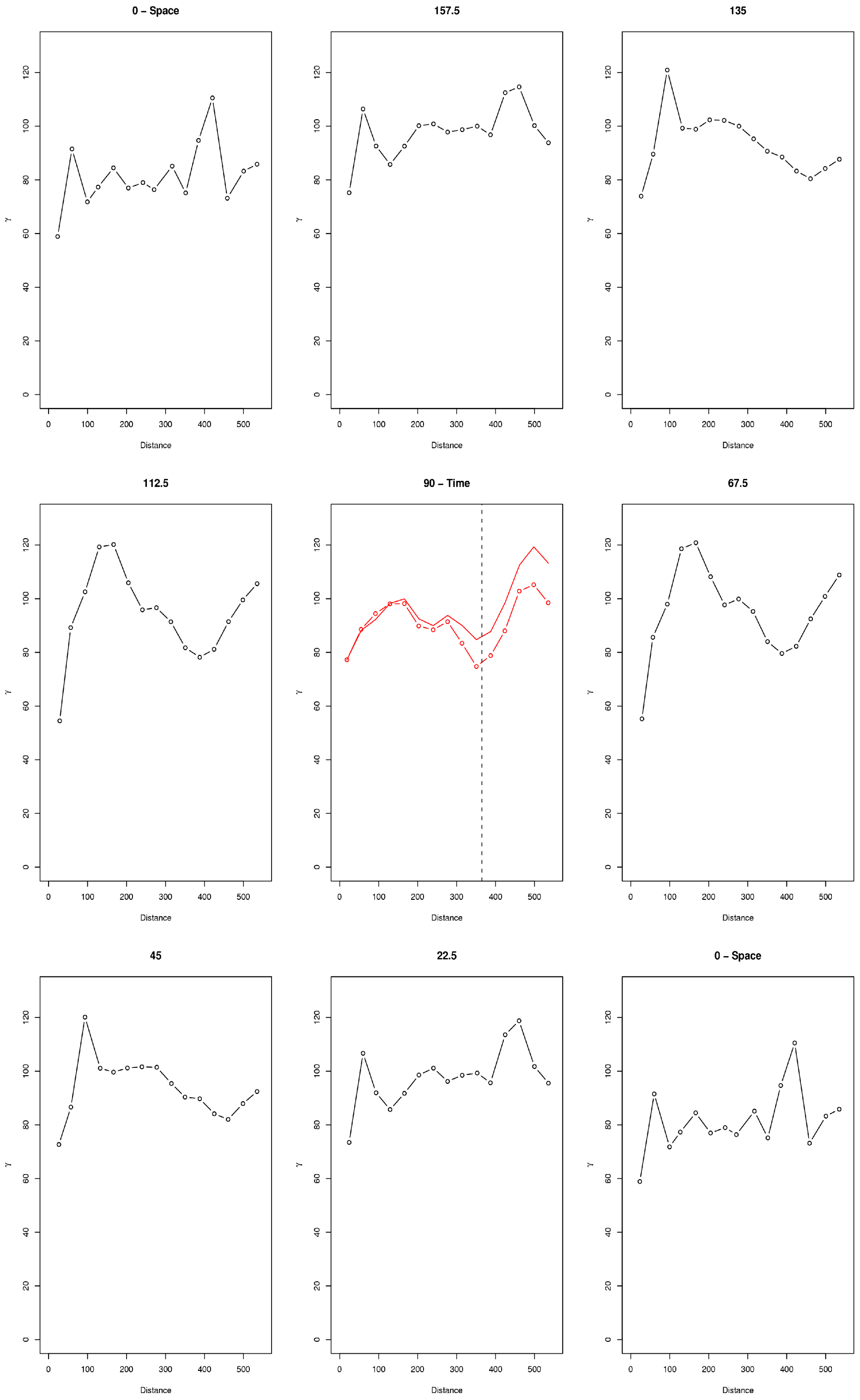


Figure 66: Cross-section plots of space-time semi-variogram for $PM_{2.5}$ mass concentration. The units for variance are $\mu g^2/m^6$.

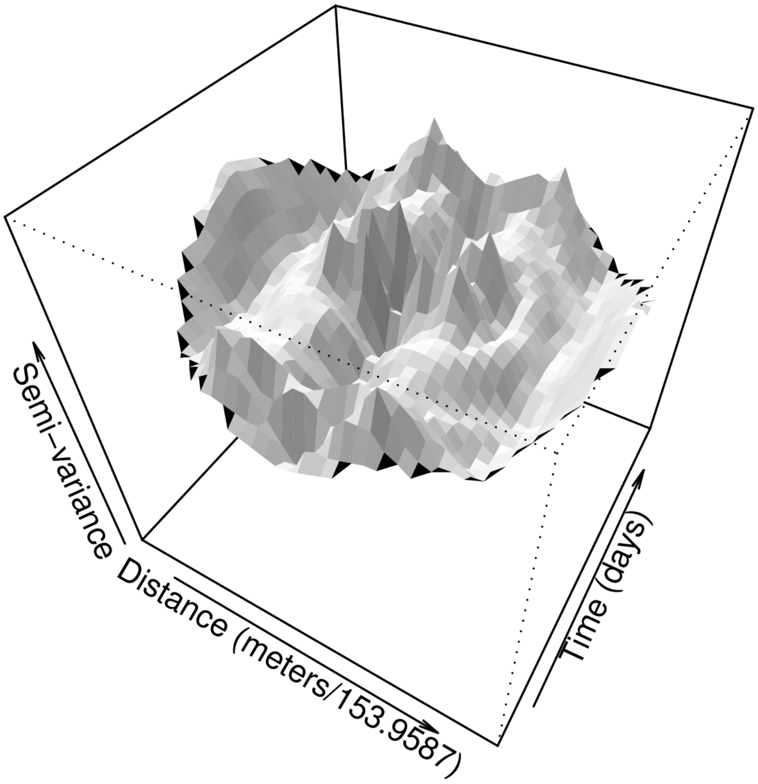


Figure 67: Perspective plot of space-time semi-variogram for $PM_{2.5}$.

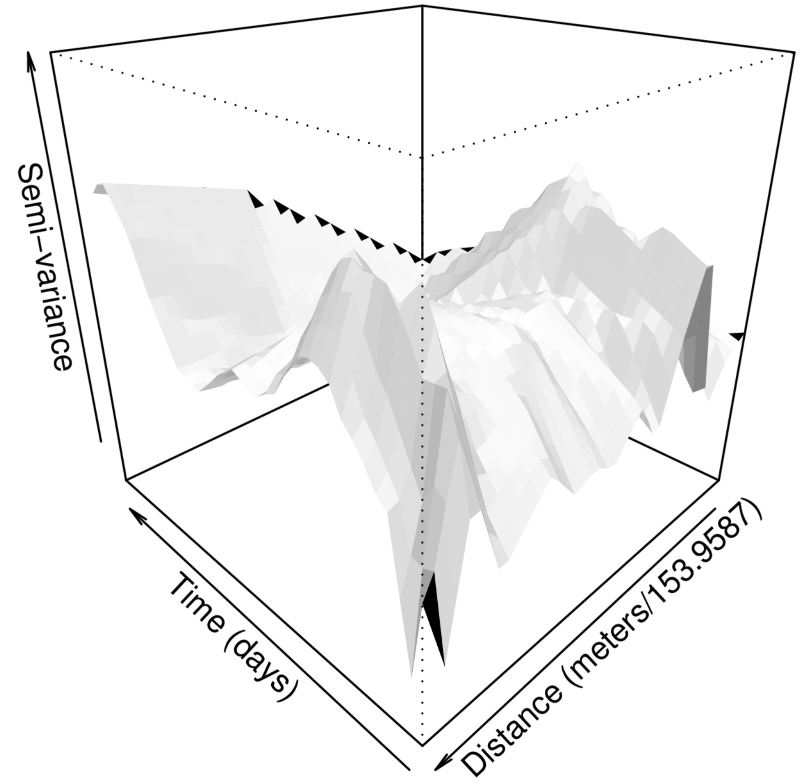


Figure 68: Perspective plot of PM$_{2.5}$ space-time variogram (one quadrant only) emphasizing the pure time and pure distance axes.

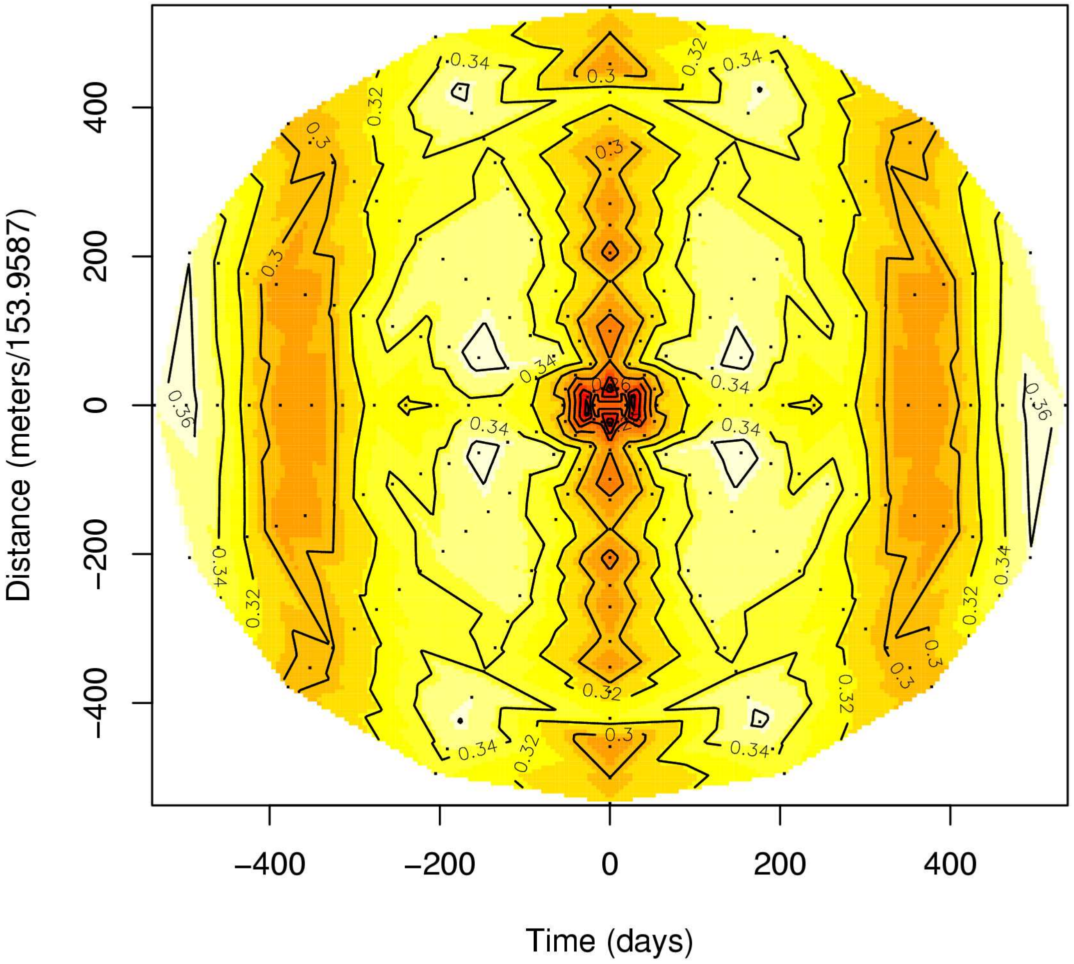


Figure 69: Contour and image map of the space-time semi-variogram for log $PM_{2.5}$. Black points denote the locations at which the semi-variance was computed. Lighter shades denote higher variances

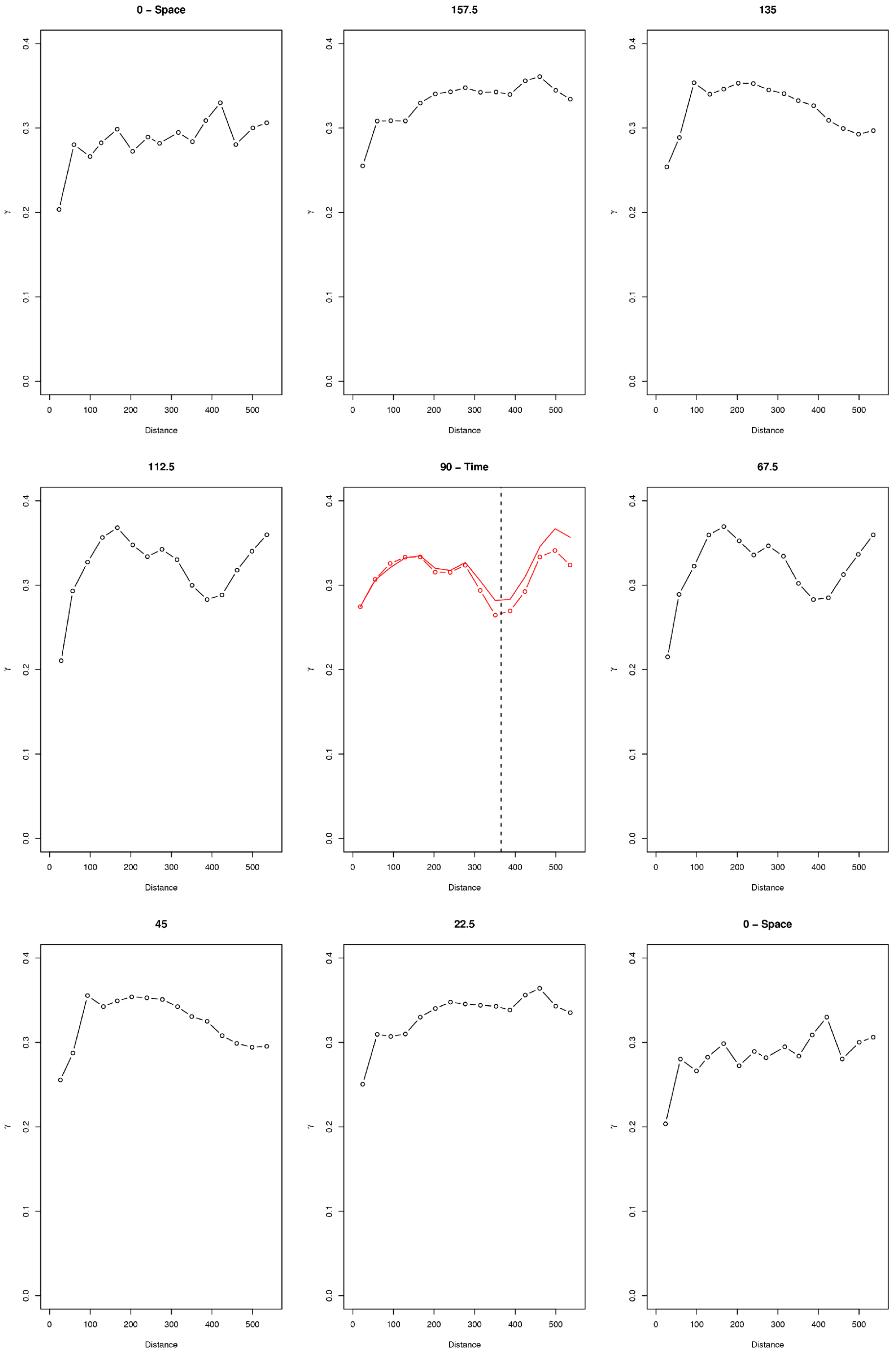


Figure 70: Cross-section plots of space-time semi-variogram for log $PM_{2.5}$. The units for variance are log $\mu g^2/m^6$.

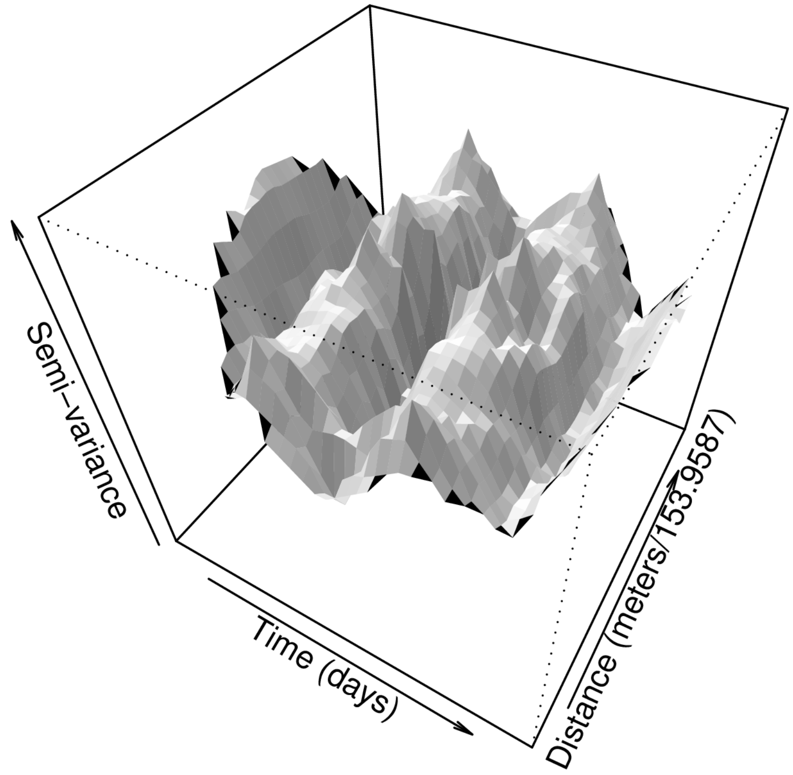


Figure 71: Perspective plot of space-time semi-variogram for log PM$_{2.5}$.

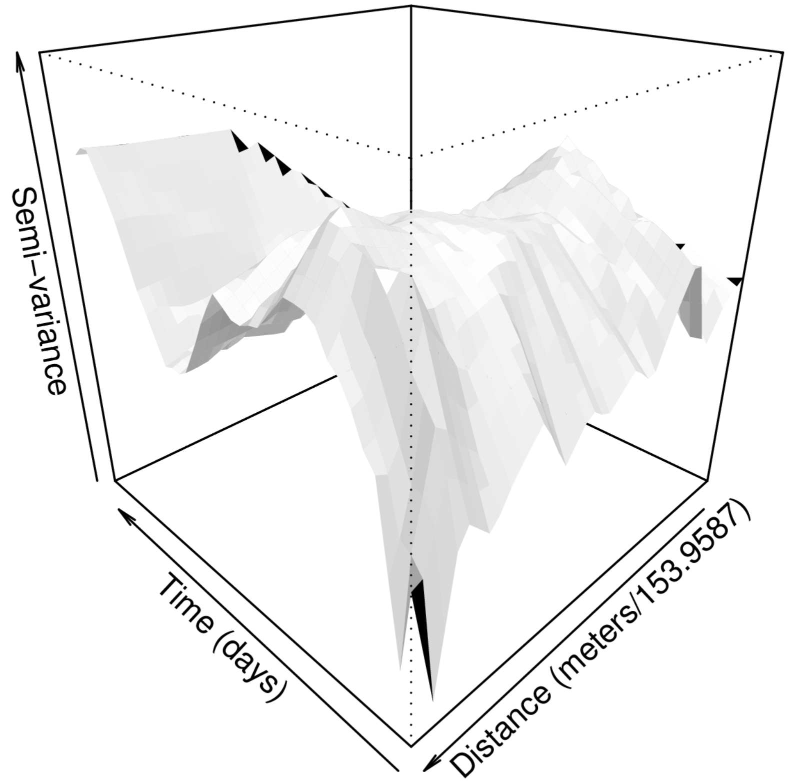


Figure 72: Perspective plot of the log $PM_{2.5}$ space-time variogram (one quadrant only) emphasizing the pure time and pure distance axes.

4.3.3 Variogram Modeling

In order to use the space-time dependence to produce optimally weighted exposure estimates, it is necessary to fit an appropriate variogram model to the sample variograms. The model is needed to ensure that any variances computed are non-negative.

4.4 Source Apportionment

4.4.1 Introduction

The goal of source apportionment is to identify the signature of the sources of the particulate material reaching the monitoring network. The method used for source apportionment depends on the information available. Typically, there is more information for the receptors than for the sources and statistical methods are required. Exploratory factor analysis has been used but is too limited to provide interpretable and trustworthy solutions. Exploratory factor analysis is a subset of a more general methodology – latent variable modeling or LVM (a. k. a., structural equation modeling or SEM). Kline (1998) provides an introduction and overview to this metthodology. This more flexible approach can better accommodate the limited source information and provide identifiable sources.

Typically, this methodology in it's simplest form assumes statistically independent samples. The assessment of statistical significance and the construction of confidence intervals are very sensitive to departures from the assumption of statistical independence. For source apportionment, the samples at each point in time at a monitoring site are correlated, i. e., statistically dependent. Simple bootstrap sampling has been successfully used to assess statistical significance and to construct realistic confidence intervals for problems where statistical independence is reasonable. When misapplied to dependent data, the results are misleading. For dependent data, special bootstrapping techniques such as block bootstrap sampling are required (Davidson and Hinkley, 1997).

4.4.2 Handling Time Dependence Using Block Bootstrapping

The goal of the method of bootstrap sampling is to recreate the variability due to sampling and determine the sampling distribution of the statistic of interest. For random sampling the samples are statistically independent. In this case, bootstrapping creates samples from the original raw data that mimic the sampling procedure to produce the sampling distribution of the statistic of interest. In practice, the method of Monte Carlo simulation is used to estimate the sampling distribution. There are no distributional assumptions. The only assumption is that the data are randomly sampled, i.e., statistically independent. This is a strong assumption and if is not reasonable, the estimate of the sampling distribution will be far from the true sampling distribution.

Measurement data used to determine source apportionment (for example, SO_2 concentrations) tend to exhibit serial correlation - measurements a day apart and a year apart tend to be more similar than measurements a month apart. This autocorrelation is an expression of statistical dependence and runs counter to the often made assumption of statistical independence or randomness which forms the backbone of many statistical methods. The statistical dependency can take many forms including seasonal dependencies. Even a "small" amount of autocorrelation can have a big impact on the estimated standard errors for estimated parameters and thus it is necessary to account for this statistical dependency. In particular, positive dependency causes the standard errors to be underestimated and thus overstates statistical significance.

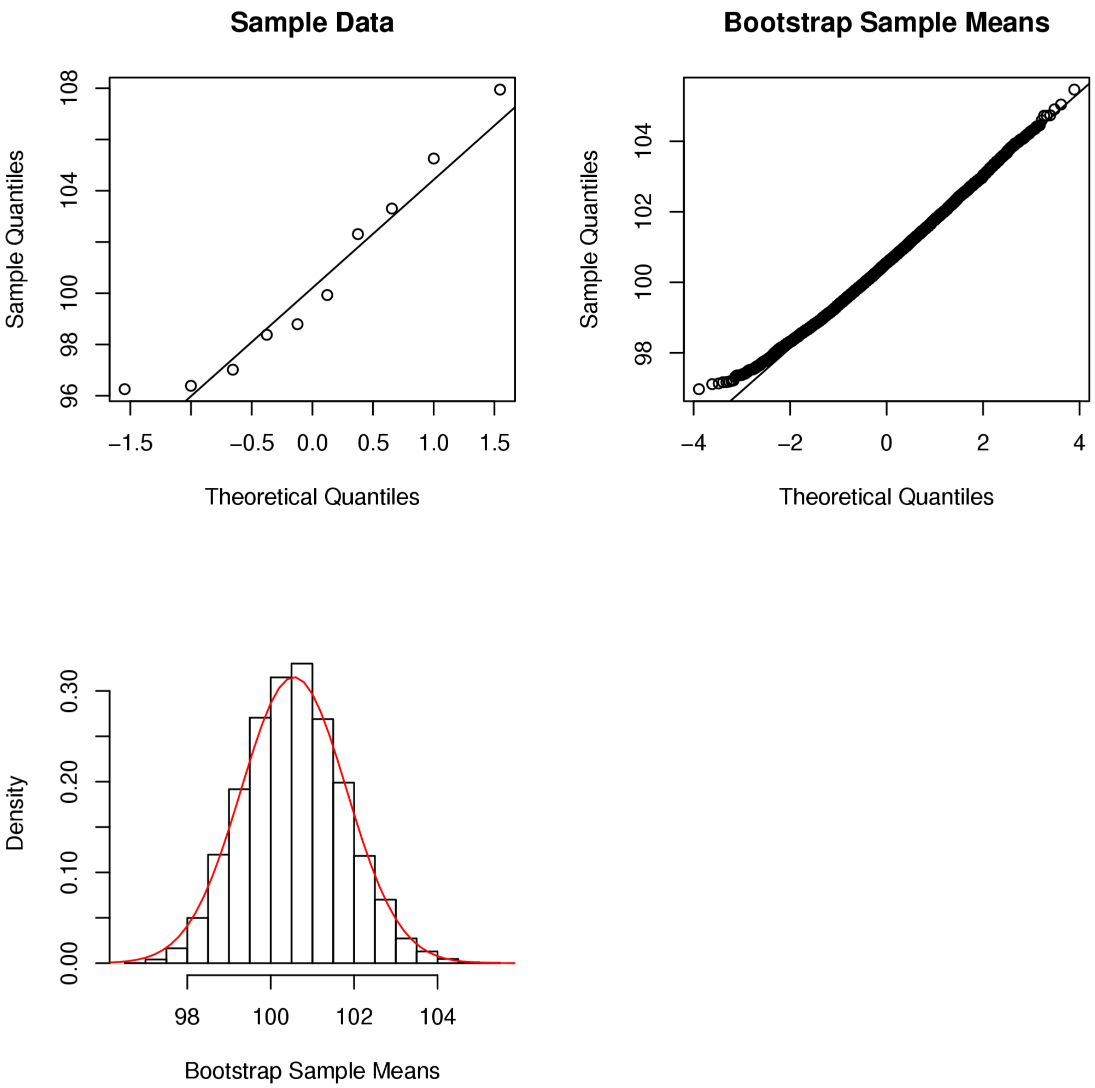
One approach to handling autocorrelated measurements is to construct a parametric model that describes the dependency and estimate the model's parameters. For example, the GLARMA modeling proposed for health counts is an example of a parametric model that accounts for autoregressive and moving average types of dependencies in order to provide optimum parameter estimates and realistic estimates of standard errors of the explanatory factors.

Listing 13: Example of estimating the standard error (SE) of sample mean from a random sample.

```
> X <- rnorm(10,100,4) # Generate a random sample n = 10 true mean = 100 true sd =
4

> X
 [1]  98.37693  98.78957 102.30517 105.25546 103.30310  97.01523  99.92815
107.94604  96.38650  96.25442

> mean(X)   # Sample mean
[1] 100.5561

> sd(X)     # Sample sd
[1] 3.999888

> sd(X)/sqrt(10)  # Estimated standard error (SE) for sample mean
[1] 1.264876

# 95% Confidence Interval - Assumptions: Normal Distribution & Statistical
Independence
> t.test(X)$conf.int
[1]  97.6947 103.4174
attr(,"conf.level")
[1] 0.95
```

Listing 14: Bootstrap estimate of the sampling distribution for the sample mean from a random sample.

```
# Basic concept - sample the raw data WITH replacement of the same size (n=10)

# sample(X,10,replace=TRUE)  # This samples the observed data with replacement n =
10

# Repeat this 10,000 times:

ave.X <- rep(NA,10000)
for(i in 1:10000) {
  ave.X[i] <- mean(sample(X,10,replace=TRUE))
  }

> mean(ave.X)   # Mean of all 10,000 bootstrapped samples of size n=10
[1] 100.5681

> sd(ave.X)     # SD of all 10,000 bootstrapped samples of size n=10
[1] 1.194682  <- compare this to conventional estimate: 1.264876


# Bootstrapped 95% confidence interval
> quantile(ave.X,c(0.025,0.5,0.975))
     2.5%       50%     97.5%
 98.34433 100.54959 102.94844  <- compare this to conventional confidence interval
```



Sample Data

Bootstrap Sample Means

Figure 73: Bootstrap results for random sample example.

Time series data typically and almost always exhibit statistical dependence. The usual approach is to try to model the dependence (in addition to making assumptions about the distribution shape). The simple bootstrap assumes statistical independence so that it is not appropriate and will give incorrect results.

Block bootstrapping is a modification of the simple bootstrap that takes the dependence pattern into account. There are a number of versions of block bootstrapping (non-overlapping blocks, overlapping blocks, nested blocks, and so forth) but we will only look at the non-overlapping block bootstrapping. (For seasonal data, nested block bootstrapping should be used.)

The block length is important. We want it long enough to capture the dependence (too short and we miss some of the dependence) but not too long (too few available blocks to sample from).

Listing 15: Analysis of raw $PM_{2.5}$ time series data.

```
# Mean of PM 2.5 time series

> mean(Y,na.rm=TRUE)
[1] 17.42429

# Median of PM 2.5 time series

> median(Y,na.rm=TRUE)
[1] 15.1

# SD of PM 2.5 time series

> sd(Y,na.rm=TRUE)
[1] 9.404735

# SE of mean under assumption of statistical independence

> sd(Y,na.rm=TRUE)/sqrt(1640)
[1] 0.2322334

# 1,000 block bootstrapped samples

# Mean of 1,000 blocked bootstrapped sample means

> mean(ave.Y)
[1] 17.41190


# SD of 1,000 blocked bootstrapped sample means = estimated SE for mean
> sd(ave.Y)
[1] 0.7319725  <= This is over 3 times the estimate assuming independence

# 95% block bootstrapped confidence interval for true mean

> quantile(ave.Y,c(0.025,0.5,0.975))
```

```
    2.5%      50%    97.5%
16.05655 17.39201 18.85222
```

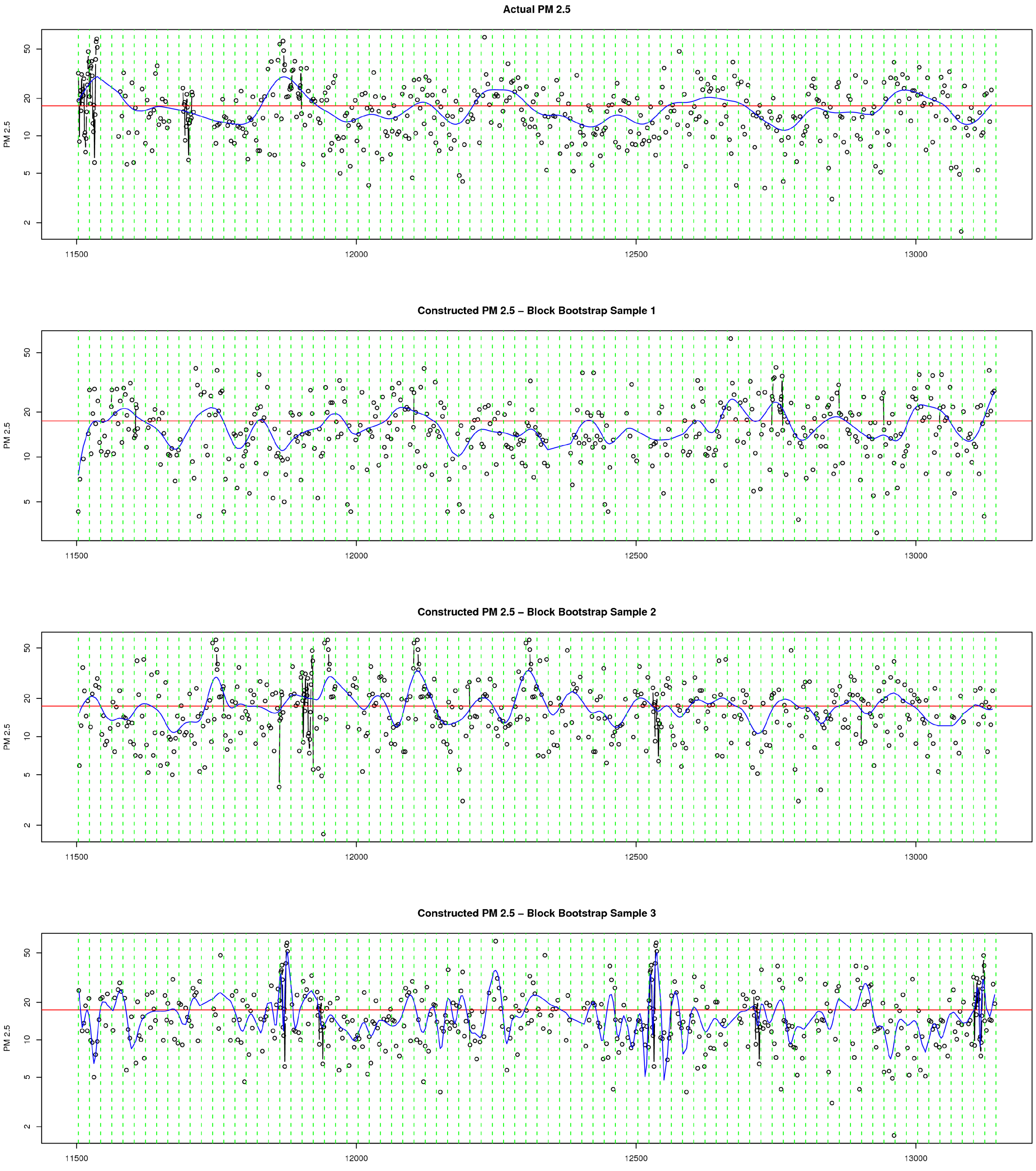
The R function to create the blocks is shown in **Listing 17**. We will need 82 blocks of length 20 (set as the defaults for the arguments). **Figure 75** shows examples of the blocks. Note that simple block boostrapping does not replicate the seasonality. For seasonal time series, a nested block bootstrap is required. An example of estimating the sampling distribution for the time series sample mean is shown in **Listing 18** and the results are shown in **Figure 77**. **Listing 16** shows the sample statistics for a $PM_{2.5}$ time series (Lawrenceville from 6/30/01 to 12/31/05). The mean of the time series is 17.42429 $\mu g/m^3$ standard deviation is 9.404735 $\mu g/m^3$ and under the assumption of statistical independence the standard error for the sample mean would be 0.2322334. The time series exhibits a great deal of statistical dependence. An autoregressive integrated moving average (ARIMA) model could be fitted to the series and the standard error computed for the sample mean assuming the fitted model. The block bootstrap does not require a model (although a parametric block bootstrap could be employed) and the results are shown in **Listing 16** and **Figure 77**. The bootstrap standard error is 0.7319725 and is over three times the estimate under the assumption of statistical independence. **Listing 19** and **Figure 77** show the results of using a log transform. The block bootstrap appears to work well even without a log transform although the log transform results are very slightly more Normally distributed. Note that the confidence interval does not depend on Normality.

Listing 16: R function to create blocks for block bootstrapping and example function call.

```
blocks <- function(Y,n=82,l=20)
{
  j <- 0
  ymat <- matrix(NA,20,n)
  for(i in 1:n) {
    ymat[,i] <- Y[(j+1):(j+l)]
    j <- j + l
    }
  ymat
}

> blocks(Y,n=82,l=20)  # series length = 20*82 = 1640
```

Listing 17: Example R code to estimate the sampling distribution for the sample mean using 1,000 bootstrapped samples.

```
# Take 1,000 samples with replacement of 82 blocks of length 20 and compute
#  the sample means

ave.Y <- rep(NA,1000)
for(i in 1:1000) {
  cat(i,"\r")
  ave.Y[i] <- mean(as.numeric(ymat[,sample(82,82,replace=TRUE)]),na.rm=TRUE)
 }
```



Actual PM 2.5

Constructed PM 2.5 – Block Bootstrap Sample 1

Constructed PM 2.5 – Block Bootstrap Sample 2

Constructed PM 2.5 – Block Bootstrap Sample 3

Figure 74: Actual $PM_{2.5}$ concentrations and three block bootstrap samples.

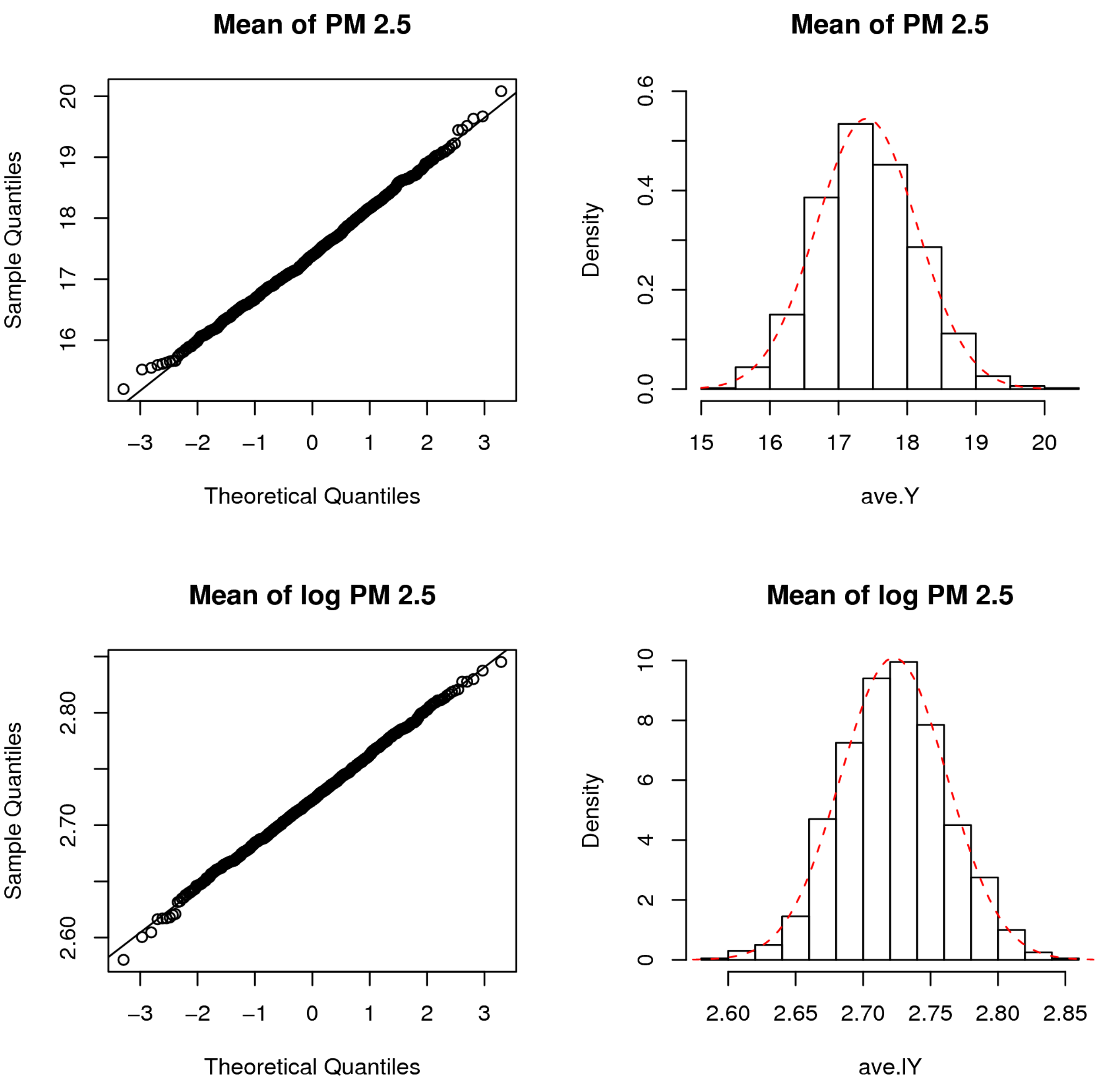


Figure 75: Results for estimating the time series mean using block bootstrapping.

Listing 18: Analysis of log transformed $PM_{2.5}$ time series data.

```
# Mean of Log PM 2.5 time series

> mean(lY,na.rm=TRUE)
[1] 2.722636

# Back transformed

> exp(mean(lY,na.rm=TRUE))
[1] 15.22039

# Median of Log PM 2.5 time series

> median(lY,na.rm=TRUE)
[1] 2.714695

# Backtransformed

> exp(median(lY,na.rm=TRUE))
[1] 15.1

# SD of Log PM 2.5 time series

> sd(lY,na.rm=TRUE)
[1] 0.5296111

# SE of mean under assumption of statistical independence

> sd(lY,na.rm=TRUE)/sqrt(1640)
[1] 0.01307781

# 1,000 block bootstrapped samples

# Mean of 1,000 blocked bootstrapped sample means
> mean(ave.lY)
[1] 2.722894

# SD of 1,000 blocked bootstrapped sample means = estimated SE for mean

> sd(ave.lY)
[1] 0.0395186 <- Over 3 times larger than SE under the assumption of independence

# 95% block bootstrapped confidence interval for true mean

> quantile(ave.lY,c(0.025,0.5,0.975))
    2.5%      50%    97.5%
2.647613 2.722427 2.800952

# Back transformed
> exp(quantile(ave.lY,c(0.025,0.5,0.975)))
    2.5%      50%    97.5%
14.12029 15.21721 16.46031
```

4.4.3 Source Apportionment via a Multivariate Receptor Model

There are various ways of determining how to apportion emissions at receptor sites (monitors) to sources. Which method is appropriate or possible depends on what source profile information is available. If extensive information is available, the problem can be approached using a chemical mass balance using the regression model:

$$x_t = \Lambda f_t + e_t$$

where x_t is a *p*-vector of observed concentrations (a.k.a, manifest variables) at time *t*,   is a $p \cdot k$ matrix of nonnegative source compositions (source profile matrix, a.k.a, factor loading matrix), f_t is a *k*-vector of nonnegative pollution source contributions (unobserved factors), and e_t is a vector of errors. For the mass balance case, the number of sources *k* is known and   is known. For example, if there are just three sources: 1) auto emissions, 2) coal fired power plant emissions, and 3) industrial emissions, then for SO^{2-}_4:

$$x_{1t} = \lambda_{11} f_{1t} + \lambda_{12} f_{2t} + \lambda_{13} f_{3t} + e_{1t}$$

= [% SO^{2-}_4 in auto emissions][concentration of auto emissions in atmosphere] +
[% SO^{2-}_4 in coal fired power plant emissions][concentration of coal fired power plant emissions in atmosphere] +
[% SO^{2-}_4 in industrial emissions][concentration of industrial emissions in atmosphere] + e_{1t}

For the mass balance case, the % SO^{2-}_4 in auto emissions, power plant emissions, and industrial emissions would be known. Then regression can be used to estimate the source contributions *f*.

Without comprehensive source profile information, the 's are not known so regression type estimation is not feasible. Exploratory factor analysis will not work because there is no unique solution and it cannot guarantee that each element of $\hat{\Lambda}$ is nonnegative and each should sum to no more than 100%. In addition, conventional factor analysis assumes that e_t is statistically independent in time. A more general class of models, referred to as latent variable models (or structural equation models) must be used. According to Christensen and Sain (2002): "We propose using a flexible latent variable model to guarantee physically valid model fits using only limited information about the relationship between the observed ambient species and the pollution sources. Latent variable modeling allows the researcher to incorporate physical constraints, laboratory measurements, past data, or other subject matter knowledge into the model so that the fitted model is interpretable." A path diagram for a hypothetical three factor model is shown in **Figure 76**. If there are *q* factors then it will be necessary to fix the factor loadings of *q* of the observed variables using available source profile information in order to eliminate factor indeterminacy. The method assumes the errors for a given species are statistically independent and from different species are uncorrelated and Normally distributed and explains the observed covariance matrix of the species concentrations as a function of the model parameters. The method of maximum likelihood is used to estimate the parameters.

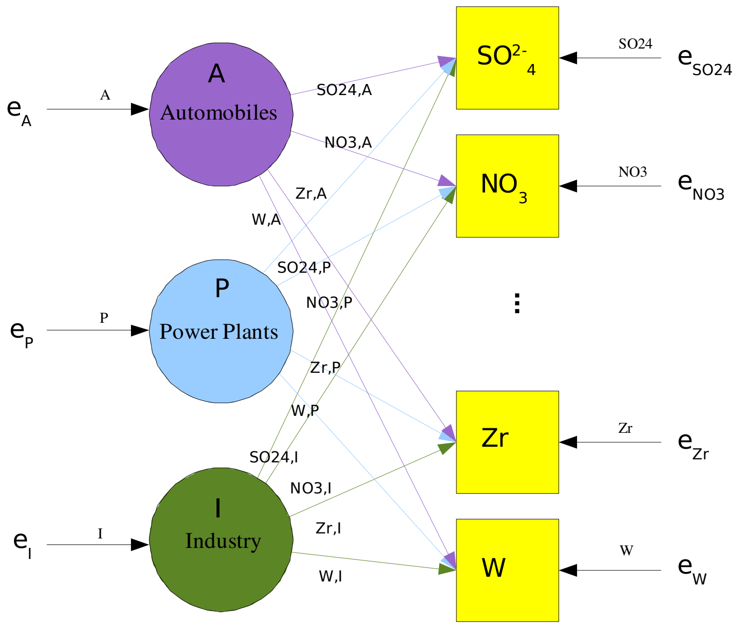


Figure 76: Multifactor latent variable model for source apportionment.

Other factor analysis related techniques include target transformation factor analysis (Alpert, 1980), positive matrix factorization (PMF) (Paatero, 1994), UNMIX (Henry, 1997; Henry, 2001), and a Bayesian methodology (Park, 2001). PMF uses nonnegative factor elements and uses weighted least squares where the standard deviations of the species are used to determine the weights.

Christensen and Sain (2002) list four problems with exploratory (unconstrained) factor analyses (EFA):

1) EFA does not prevent negative estimates for parameters that must be nonnegative,

2) EFA does not provide a unique solution (the model is not identifiable),

3) EFA cannot include partial source profile information, and

4) EFA does not allow for temporal dependence (autocorrelated errors).

PMF and UNMIX only correct the first problem. Latent variable modeling corrects for the first three problems. To account for serial dependency, Christensen and Sain (2002) propose (in addition to the latent variable model with constraints discussed above) the block bootstrap and in particular the nested block bootstrap to handle seasonal dependency. The general method of block bootstrapping is well-understood and accounts for serial dependence *without having to model the dependence*. This feature is especially important because modeling temporal dependence can be difficult especially in the case of extensive missing data.

4.4.4 Latent Variable Multivariate Receptor Model Example

In this example of using block bootstrapping to estimate the parameters of a latent variable / multivariate receptor model there are five hypothesized latent sources. The source profiles are based on the source apportionment results that ACHD obtained by applying PMF to the data from the Lawrenceville site.

The five latent sources identified by ACHD were as follows:

1. Secondary Sulfates
2. Secondary Nitrates
3. Mobile / Industrial
4. Crustal / Road Dust
5. Miscellaneous Burning / Cooking

ACHD included the following twenty-three $PM_{2.5}$ species in their model: NH_4, NO_3, SO_4, OC, EC, Al, As, Br, Ca, Cl, Cr, Cu, Fe, Pb, Mn, Hg, Ni, K, Se, Si, Ti, V, and Zn.

Table 54: Fractional contributions of five important species for five hypothesized latent sources.

	Secondary Sulfates	Secondary Nitrates	Mobile / Industrial	Crustal / Road Dust	Misc. Burning / Cooking
NO_3	0.008	0.519	0.000	0.171	0.000
SO_4	0.592	0.187	0.179	0.116	0.144
EC	0.005	0.016	0.159	0.033	0.075
Pb	0.000	0.000	0.004	0.000	0.000
Si	0.000	0.000	0.001	0.043	0.000

Table 54 shows the fractional contributions of five important species - SO_4, NO_3, EC, Pb, and Si to the mass of each source. (e.g., for SO_4 in the secondary sulfates source, the fractional contribution of 0.592 means that SO_4 accounted for 59.2% of the total mass of $PM_{2.5}$ associated with this source, based on the ACHD PMF results). These values were assigned to the corresponding parameters in the Λ matrix in

order to create an identified latent variable model. The R code shown in **Listing 20** creates a block bootstrap sample, computes and writes the variance-covariance matrix and sample size to a file and then invokes the Mx code. The Mx code inputs the variance-covariance matrix file and then computes the parameter estimates and outputs the results to a file. R then inputs the results from Mx and stores the estimates. This process was repeated 1,000 times.

The speciation data used in this example were taken from the Lawrenceville site from 6/30/2001 to 12/31/2005. The estimated variance-covariance matrix is shown in **Table 61**. The Λ parameter estimates are shown in **Listing 21**. The individual estimates are all non-negative and sum for each factor to 1 or less thus honoring the model constraints. The block bootstrapped estimated sampling distributions are plotted in **Figure 77**. The medians, means, and standard deviations of the estimated Λ sampling distributions are shown in **Table 57**. The 95% confidence interval lower and upper bounds were computed as the 2.5th and 97.5th percentiles, respectively, of the estimated sampling distributions and are shown in **Table 56**. The standard deviations represent the standard errors of the estimated parameters. For example, the actual estimate for NH4 for Factor 1 (Secondary Sulfates) is 0.355 (35.5%). The mean of the estimated sampling distribution was 0.3616 (36.16%) and the median was 0.3576 (35.76%). The estimated standard error for this estimate was 0.0138 (1.38%) and the 95% confidence interval was 0.3424 to 0.3929 (34.24% to 39.29%). The shape of the sampling distribution is non-Normal and positively skewed. Many of the estimated sampling distributions were highly non-Normal, especially when the parameter value was close to zero. The parameter estimates for Factor 4 (OC, Al, As, Fe, and K) had especially large standard errors and wide confidence intervals compared to the other factors.

Table 55: Variance-covariance matrix for Lawrenceville site based on data from 6/30/2001 to 12/31/2005.

	NH4	NO3	SO42	OC	EC	Al	As	Br	Ca	Cl	Cr	Cu	Fe	Pb	Mn	Hg	Ni	K	Se	Si	Ti	V	Zn
NH4	1.960	0.497	6.070	1.584	0.257	0.001	0.000	0.002	0.016	0.006	0.001	0.002	0.044	0.003	0.002	0	0.001	0.022	0.003	-0.014	0.002	0	0.007
NO3	0.497	2.115	-0.864	0.411	0.097	-0.004	0.000	0.002	0.001	0.028	0.000	0.001	0.010	0.001	0.000	0	0.000	0.010	0.001	-0.049	0.000	0	0.009
SO42	6.070	-0.864	23.110	4.799	0.710	0.007	0.001	0.003	0.052	-0.014	0.002	0.005	0.125	0.007	0.005	0	0.002	0.067	0.009	0.022	0.008	0	0.009
OC	1.584	0.411	4.799	4.129	0.589	0.010	0.001	0.003	0.040	0.018	0.003	0.004	0.104	0.010	0.007	0	0.001	0.056	0.008	0.051	0.005	0	0.025
EC	0.257	0.097	0.710	0.589	0.183	0.001	0.000	0.001	0.008	0.006	0.001	0.001	0.030	0.002	0.002	0	0.000	0.006	0.002	-0.004	0.001	0	0.007
Al	0.001	-0.004	0.007	0.010	0.001	0.004	0.000	0.000	0.001	0.000	0.000	0.000	0.002	0.000	0.000	0	0.000	0.002	0.000	0.027	0.000	0	0.000
As	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000	0	0.000
Br	0.002	0.002	0.003	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000	0	0.000
Ca	0.016	0.001	0.052	0.040	0.008	0.001	0.000	0.000	0.002	0.000	0.000	0.000	0.002	0.000	0.000	0	0.000	0.001	0.000	0.002	0.000	0	0.000
Cl	0.006	0.028	-0.014	0.018	0.006	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.001	0.000	0.000	0	0.000	0.000	0.000	-0.001	0.000	0	0.000
Cr	0.001	0.000	0.002	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000	0	0.000
Cu	0.002	0.001	0.005	0.004	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000	0	0.000
Fe	0.044	0.010	0.125	0.104	0.030	0.002	0.000	0.000	0.002	0.001	0.000	0.000	0.009	0.000	0.001	0	0.000	0.002	0.000	0.003	0.000	0	0.001
Pb	0.003	0.001	0.007	0.010	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000	0	0.000
Mn	0.002	0.000	0.005	0.007	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000	0	0.000
Hg	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000	0	0.000
Ni	0.001	0.000	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000	0	0.000
K	0.022	0.010	0.067	0.056	0.006	0.002	0.000	0.000	0.001	0.000	0.000	0.000	0.002	0.000	0.000	0	0.000	0.008	0.000	0.003	0.000	0	0.000
Se	0.003	0.001	0.009	0.008	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000	0	0.000
Si	-0.014	-0.049	0.022	0.051	-0.004	0.027	0.000	0.000	0.002	-0.001	0.000	0.000	0.003	0.000	0.000	0	0.000	0.003	0.000	0.619	0.000	0	0.000
Ti	0.002	0.000	0.008	0.005	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000	0	0.000
V	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000	0	0.000
Zn	0.007	0.009	0.009	0.025	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000	0	0.001

Listing 19: Estimated Λ parameters based on the speciated Lawrenceville site data computed by Mx.

```
          FACTOR-1     FACTOR-2     FACTOR-3     FACTOR-4     FACTOR-5
 NH4    3.5511E-01   1.6364E-01   7.3005E-02   1.1257E-01   0.0000E+00
 NO3    8.0000E-03   5.1900E-01   0.0000E+00   1.7100E-01   0.0000E+00
SO42    5.9200E-01   1.8700E-01   1.7900E-01   1.1600E-01   1.1400E-01
  OC    3.8216E-02   0.0000E+00   2.3096E-01   2.2319E-01   7.8353E-01
  EC    5.0000E-03   1.1600E-01   1.5900E-01   3.3000E-02   7.5000E-02
  AL    0.0000E+00   0.0000E+00   1.6019E-01   1.5394E-01   8.7086E-09
  AS    0.0000E+00   0.0000E+00   1.1625E-01   1.0971E-01   8.7506E-09
  BR    0.0000E+00   0.0000E+00   4.1964E-03   3.6121E-03   9.0809E-09
  CA    6.2582E-04   0.0000E+00   6.3889E-03   2.3358E-03   6.4958E-03
  CL    0.0000E+00   0.0000E+00   8.4529E-03   6.3564E-03   9.6178E-09
  CR    0.0000E+00   0.0000E+00   7.8932E-03   6.6863E-03   9.0087E-09
  CU    6.0181E-05   3.1872E-04   8.3532E-04   2.8553E-04   6.9677E-05
  FE    5.4357E-04   1.0199E-02   3.0879E-02   6.2569E-03   9.8234E-03
  PB    0.0000E+00   0.0000E+00   4.0000E-03   0.0000E+00   0.0000E+00
  MN    0.0000E+00   0.0000E+00   3.0619E-03   1.0231E-03   0.0000E+00
  HG    0.0000E+00   0.0000E+00   1.0920E-03   9.9407E-04   0.0000E+00
  NI    0.0000E+00   0.0000E+00   1.3159E-03   9.9405E-04   0.0000E+00
   K    4.4621E-04   0.0000E+00   3.1223E-04   4.1037E-03   1.0122E-02
  SE    0.0000E+00   0.0000E+00   2.4072E-03   1.0797E-03   0.0000E+00
  SI    0.0000E+00   0.0000E+00   1.0000E-03   4.3000E-02   0.0000E+00
  TI    0.0000E+00   0.0000E+00   1.9585E-03   1.0809E-03   0.0000E+00
   V    0.0000E+00   0.0000E+00   1.0797E-03   9.7871E-04   0.0000E+00
  ZN    0.0000E+00   3.8467E-03   6.7309E-03   1.7939E-03   9.6170E-04
```

Table 56: Block-bootstrapped 95% confidence intervals for parameters. Grayed cells indicate parameters with assigned values.

Factor 1 - Secondary Sulfates			Factor 2 - Secondary Nitrates			Factor 3 - Mobile / Industrial			Factor 4 - Crustal / Road Dust			Factor 5 - Misc. Burn./ Cook.		
	Lower	**Upper**		**Lower**	**Upper**		**Lower**	**Upper**		**Lower**	**Upper**		**Lower**	**Upper**
NH_4	34.24	39.29	NH_4	15.60	16.98	NH_4	4.87	9.49	NH_4	5.23	11.66	NH_4	0.00	0.00
NO_3	0.80	0.80	NO_3	51.90	51.90	NO_3	0.00	0.00	NO_3	17.10	17.10	NO_3	0.00	0.00
SO42	59.20	59.20	SO42	18.70	18.70	SO42	17.90	17.90	SO42	11.60	11.60	SO42	11.40	11.40
OC	0.18	5.17	OC	0.00	0.00	OC	17.93	31.35	OC	21.27	53.90	OC	77.62	81.06
EC	0.50	0.50	EC	11.60	11.60	EC	15.90	15.90	EC	3.30	3.30	EC	7.50	7.50
Al	0.00	0.00	Al	0.00	0.00	Al	11.24	19.03	Al	0.00	16.13	Al	0.00	0.00
As	0.00	0.00	As	0.00	0.00	As	8.09	13.80	As	0.00	11.56	As	0.00	0.00
Br	0.00	0.00	Br	0.00	0.00	Br	0.31	0.48	Br	0.00	0.38	Br	0.00	0.00
Ca	0.00	0.16	Ca	0.00	0.16	Ca	0.45	1.02	Ca	0.17	2.06	Ca	0.00	0.78
Cl	0.00	0.00	Cl	0.00	0.45	Cl	0.53	1.14	Cl	0.00	0.69	Cl	0.00	0.00
Cr	0.00	0.00	Cr	0.00	0.00	Cr	0.58	0.92	Cr	0.00	0.70	Cr	0.00	0.00
Cu	0.00	0.01	Cu	0.00	0.05	Cu	0.05	0.11	Cu	0.02	0.10	Cu	0.00	0.04
Fe	0.00	0.33	Fe	0.37	1.41	Fe	2.63	3.71	Fe	0.51	3.07	Fe	0.00	1.32
Pb	0.00	0.00	Pb	0.00	0.00	Pb	0.40	0.40	Pb	0.00	0.00	Pb	0.00	0.00
Mn	0.00	0.00	Mn	0.00	0.00	Mn	0.26	0.35	Mn	0.00	0.11	Mn	0.00	0.00
Hg	0.00	0.00	Hg	0.00	0.00	Hg	0.08	0.13	Hg	0.00	0.10	Hg	0.00	0.00
Ni	0.00	0.00	Ni	0.00	0.00	Ni	0.10	0.15	Ni	0.00	0.10	Ni	0.00	0.00
K	0.00	0.26	K	0.00	0.42	K	0.00	0.71	K	0.00	3.69	K	0.00	1.50
Se	0.00	0.01	Se	0.00	0.10	Se	0.17	0.34	Se	0.00	0.12	Se	0.00	0.04
Si	0.00	0.00	Si	0.00	0.00	Si	0.10	0.10	Si	4.30	4.30	Si	0.00	0.00
Ti	0.00	0.00	Ti	0.00	0.00	Ti	0.15	0.24	Ti	0.00	0.22	Ti	0.00	0.00
V	0.00	0.00	V	0.00	0.00	V	0.08	0.13	V	0.00	0.10	V	0.00	0.00
Zn	0.00	0.00	Zn	0.25	0.47	Zn	0.56	0.77	Zn	0.12	0.32	Zn	0.00	0.26

Table 57: Summary statistics for the block bootstrapped parameter sampling distributions. Grayed cells indicate parameters with assigned values.

Factor 1 - Secondary Sulfates				**Factor 2 - Secondary Nitrates**				**Factor 3 - Mobile / Industrial**				**Factor 4 - Crustal / Road Dust**				**Factor 5 - Misc. Burn./ Cook.**			
	Med.	**Mean**	**SD**		**Med.**	**Mean**	**SD**		**Med.**	**Mean**	**SD**		**Med.**	**Mean**	**SD**		**Med.**	**Mean**	**SD**
NH_4	35.76	36.16	1.38	NH_4	16.37	16.35	0.37	NH_4	7.05	7.08	1.21	NH_4	11.08	10.39	1.78	NH_4	0	0	0.07
NO_3	0.80	0.80	0.00	NO_3	51.90	51.90	0.00	NO_3	0.00	0.00	0.00	NO_3	17.10	17.10	0.00	NO_3	0.00	0.00	0.00
SO_4^2	59.20	59.20	0.00	SO_4^2	18.70	18.70	0.00	SO_4^2	17.90	17.90	0.00	SO_4^2	11.60	11.60	0.00	SO_4^2	11.40	11.40	0.00
OC	3.46	3.19	1.33	OC	0.00	0.00	0.00	OC	24.33	24.46	3.73	OC	22.39	31.11	12.57	OC	78.65	79.11	1.22
EC	0.50	0.50	0.00	EC	11.60	11.60	0.00	EC	15.90	15.90	0.00	EC	3.30	3.30	0.00	EC	7.50	7.50	0.00
Al	0.00	0.00	0.00	Al	0.00	0.00	0.00	Al	15.33	15.24	2.04	Al	15.35	10.13	7.48	Al	0.00	0.00	0.00
As	0.00	0.00	0.00	As	0.00	0.00	0.00	As	11.12	11.03	1.50	As	10.93	7.23	5.34	As	0.00	0.00	0.00
Br	0.00	0.00	0.00	Br	0.00	0.00	0.00	Br	0.40	0.40	0.05	Br	0.36	0.24	0.18	Br	0.00	0.00	0.00
Ca	0.03	0.04	0.05	Ca	0.00	0.02	0.05	Ca	0.71	0.71	0.16	Ca	0.24	0.73	0.71	Ca	0.56	0.42	0.30
Cl	0.00	0.00	0.00	Cl	0.00	0.05	0.13	Cl	0.82	0.82	0.15	Cl	0.63	0.42	0.31	Cl	0.00	0.00	0.00
Cr	0.00	0.00	0.00	Cr	0.00	0.00	0.00	Cr	0.76	0.76	0.09	Cr	0.67	0.44	0.32	Cr	0.00	0.00	0.00
Cu	0.00	0.00	0.00	Cu	0.03	0.03	0.01	Cu	0.09	0.08	0.01	Cu	0.03	0.05	0.03	Cu	0.00	0.01	0.01
Fe	0.00	0.06	0.09	Fe	0.94	0.93	0.27	Fe	3.17	3.18	0.28	Fe	0.65	1.31	0.99	Fe	0.83	0.64	0.48
Pb	0.00	0.00	0.00	Pb	0.00	0.00	0.00	Pb	0.40	0.40	0.00	Pb	0.00	0.00	0.00	Pb	0.00	0.00	0.00
Mn	0.00	0.00	0.00	Mn	0.00	0.00	0.00	Mn	0.30	0.30	0.02	Mn	0.10	0.07	0.05	Mn	0.00	0.00	0.00
Hg	0.00	0.00	0.00	Hg	0.00	0.00	0.00	Hg	0.10	0.10	0.01	Hg	0.10	0.07	0.05	Hg	0.00	0.00	0.00
Ni	0.00	0.00	0.00	Ni	0.00	0.00	0.00	Ni	0.13	0.13	0.01	Ni	0.10	0.07	0.05	Ni	0.00	0.00	0.00
K	0.00	0.05	0.08	K	0.00	0.05	0.13	K	0.09	0.19	0.22	K	0.42	1.03	1.24	K	0.91	0.80	0.59
Se	0.00	0.00	0.01	Se	0.00	0.01	0.03	Se	0.23	0.24	0.04	Se	0.11	0.07	0.05	Se	0.00	0.00	0.01
Si	0.00	0.00	0.00	Si	0.00	0.00	0.00	Si	0.10	0.10	0.00	Si	4.30	4.30	0.00	Si	0.00	0.00	0.00
Ti	0.00	0.00	0.00	Ti	0.00	0.00	0.00	Ti	0.19	0.19	0.02	Ti	0.11	0.10	0.05	Ti	0.00	0.00	0.00
V	0.00	0.00	0.00	V	0.00	0.00	0.00	V	0.10	0.10	0.01	V	0.10	0.06	0.05	V	0.00	0.00	0.00
Zn	0.00	0.00	0.00	Zn	0.37	0.37	0.06	Zn	0.67	0.67	0.05	Zn	0.18	0.18	0.05	Zn	0.11	0.11	0.07

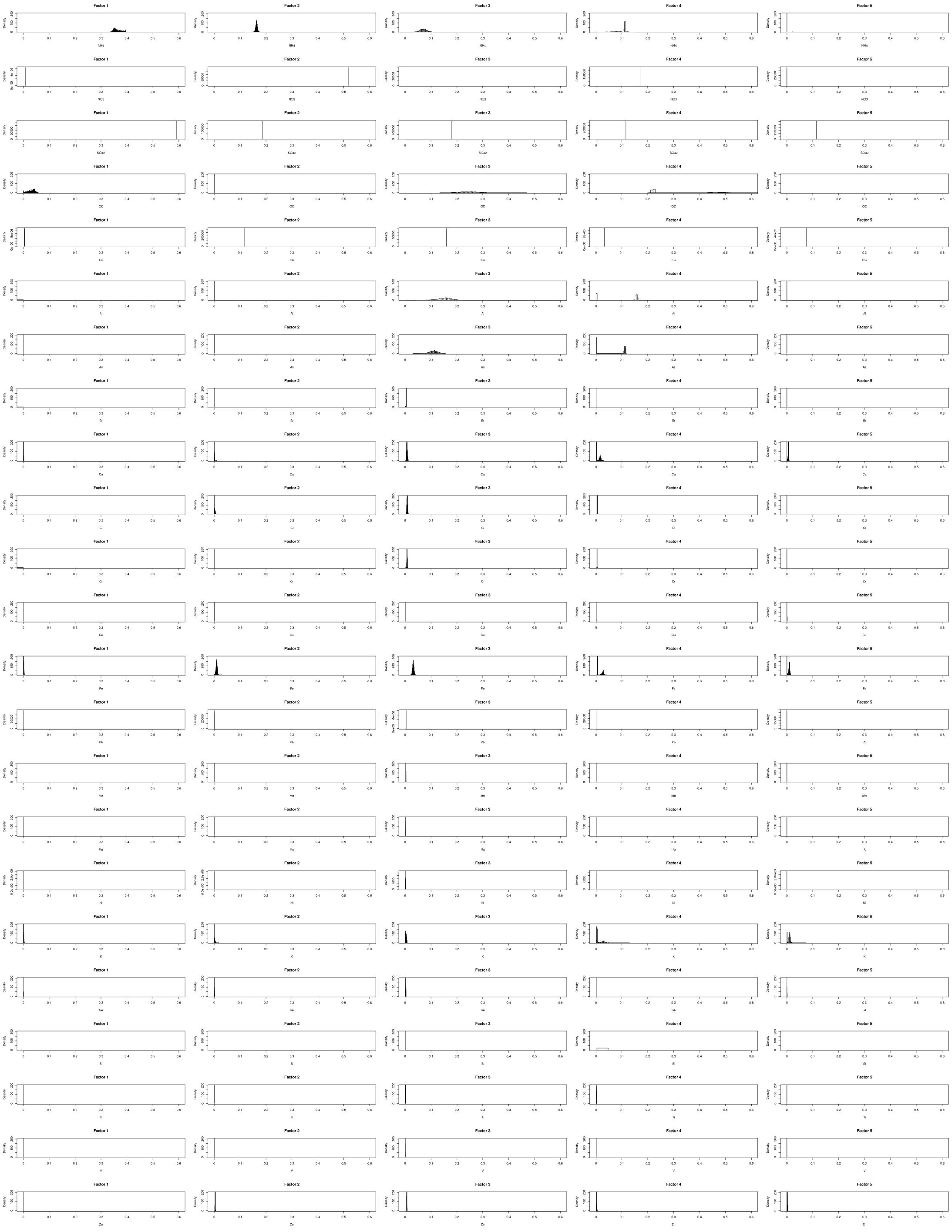


Figure 77: Block-bootstrapped sampling distributions for estimated parameters of Λ.

Listing 20: R code for block-bootstrapped latent variable / multivariate receptor model. This code repeatedly executes the structural equation program Mx code shown in Listing 22.

```
labs <- c("NH4","NO3","SO42","OC","EC","Al","As","Br","Ca",
  "Cl","Cr","Cu","Fe","Pb","Mn","Hg","Ni","K","Se","Si","Ti",
  "V","Zn")

################################
Y1 <- lawrenceville$NH4[1:1640]
Y2 <- lawrenceville$NO3[1:1640]
Y3 <- lawrenceville$SO42[1:1640]
Y4 <- lawrenceville$OC[1:1640]
Y5 <- lawrenceville$EC[1:1640]
Y6 <- lawrenceville$Al[1:1640]
Y7 <- lawrenceville$As[1:1640]
Y8 <- lawrenceville$Br[1:1640]
Y9 <- lawrenceville$Ca[1:1640]
Y10 <- lawrenceville$Cl[1:1640]
Y11 <- lawrenceville$Cr[1:1640]
Y12 <- lawrenceville$Cu[1:1640]
Y13 <- lawrenceville$Fe[1:1640]
Y14 <- lawrenceville$Pb[1:1640]
Y15 <- lawrenceville$Mn[1:1640]
Y16 <- lawrenceville$Hg[1:1640]
Y17 <- lawrenceville$Ni[1:1640]
Y18 <- lawrenceville$K[1:1640]
Y19 <- lawrenceville$Se[1:1640]
Y20 <- lawrenceville$Si[1:1640]
Y21 <- lawrenceville$Ti[1:1640]
Y22 <- lawrenceville$V[1:1640]
Y23 <- lawrenceville$Zn[1:1640]

dim(na.omit(cbind(Y1,Y2,Y3,Y4,Y5,Y6,Y7,Y8,Y9,Y10,Y11,Y12,
  Y13,Y14,Y15,Y16,Y17,Y18,Y19,Y20,Y21,Y22,Y23)))

vc.actual <- var(na.omit(cbind(Y1,Y2,Y3,Y4,Y5,Y6,Y7,Y8,Y9,Y10,Y11,
  Y12,Y13,Y14,Y15,Y16,Y17,Y18,Y19,Y20,Y21,Y22,Y23)))
dimnames(vc.actual) <- list(lab,lab)

write.table(vc.actual,
  "/projects/PITT-PM/Source_Apportionment/Data/varcov/vca23.txt",
  row.names=FALSE,col.names=FALSE,sep=" ")

round(vc.actual,3)

ymat1 <- blocks(Y1)
ymat2 <- blocks(Y2)
ymat3 <- blocks(Y3)
ymat4 <- blocks(Y4)
ymat5 <- blocks(Y5)
ymat6 <- blocks(Y6)
ymat7 <- blocks(Y7)
ymat8 <- blocks(Y8)
ymat9 <- blocks(Y9)
ymat10 <- blocks(Y10)
ymat11 <- blocks(Y11)
ymat12 <- blocks(Y12)
ymat13 <- blocks(Y13)
ymat14 <- blocks(Y14)
ymat15 <- blocks(Y15)
ymat16 <- blocks(Y16)
ymat17 <- blocks(Y17)
ymat18 <- blocks(Y18)
ymat19 <- blocks(Y19)
ymat20 <- blocks(Y20)
ymat21 <- blocks(Y21)
ymat22 <- blocks(Y22)
ymat23 <- blocks(Y23)
```

```
N <- 1000

vc <- array(NA,c(23,23,N))
nobs <- rep(NA,N)
for(i in 1:N) {
  cat(i,"\r  ")
  samp <- sample(82,82,replace=TRUE)
  x <- cbind(as.numeric(ymat1[,samp]),
             as.numeric(ymat2[,samp]),
             as.numeric(ymat3[,samp]),
             as.numeric(ymat4[,samp]),
             as.numeric(ymat5[,samp]),
             as.numeric(ymat6[,samp]),
             as.numeric(ymat7[,samp]),
             as.numeric(ymat8[,samp]),
             as.numeric(ymat9[,samp]),
             as.numeric(ymat10[,samp]),
             as.numeric(ymat11[,samp]),
             as.numeric(ymat12[,samp]),
             as.numeric(ymat13[,samp]),
             as.numeric(ymat14[,samp]),
             as.numeric(ymat15[,samp]),
             as.numeric(ymat16[,samp]),
             as.numeric(ymat17[,samp]),
             as.numeric(ymat18[,samp]),
             as.numeric(ymat19[,samp]),
             as.numeric(ymat20[,samp]),
             as.numeric(ymat21[,samp]),
             as.numeric(ymat22[,samp]),
             as.numeric(ymat23[,samp]))
        vc[,,i] <- var(x,use="complete")
        nobs[i] <- dim(na.omit(x))[1]
        print(nobs[i])
 }

lab <- c("V1","V2","V3","V4","V5","V6","V7","V8","V9",
         "V10","V11","V12","V13","V14","V15","V16","V17","V18","V19",
         "V20","V21","V22","V23")

B <- array(NA,c(23,5,N))

for(i in c(1:N))
{
  print(i)
  write.table(data.frame(vc[,,i]),
    "/projects/PITT-PM/Source_Apportionment/Data/varcov/vca23.txt",
    row.names=FALSE,col.names=FALSE,sep=" ",quote=FALSE)
  write.table(paste("Data NObservations=",nobs[i]," NInput_variables=23",sep=""),
    "/projects/PITT-PM/Source_Apportionment/Data/varcov/nobs_vars.txt",
    col.names=FALSE,row.names=FALSE,quote=FALSE)
  system(paste("mxt < /projects/PITT-PM/Source_Apportionment/R/sa_23s_5f.
    mx > /projects/PITT-PM/Source_Apportionment/Data/mx/sa_23s_5f_",i,"_output.mx",sep=""))
  A <- scan("/projects/PITT-PM/Source_Apportionment/Data/A.txt",what=" ",sep="\n")
  A <- gsub("D","E",A)
  write.table(A[2:21],"/projects/PITT-PM/Source_Apportionment/Data/B.txt",
    row.names=FALSE,col.names=FALSE,sep=" ",quote=FALSE)
  E <- read.fwf("/projects/PITT-PM/Source_Apportionment/Data/B.txt",
    widths=rep(13,6),header=FALSE)
  B[,,i] <- t(matrix(na.omit(as.vector(t(as.matrix(E)))),5,23))

}
```

Listing 21: Mx code for latent variable / multivariate receptor model for five latent factors.

```
#NGroups 6
Source Apportionment - 23 manifest variables - 5 latent factors ( To run: > mxt < sa_23s_5f.mx >
sa_23s_5f_output.txt )
!Data NObservations=488 NInput_variables=23
```

```
#Include /projects/PITT-PM/Source_Apportionment/Data/varcov/nobs_vars.txt
CMatrix Full
#Include /projects/PITT-PM/Source_Apportionment/Data/varcov/vca23.txt
Labels NH4 NO3 SO42 OC EC Al As Br Ca Cl Cr Cu Fe Pb Mn Hg Ni K Se Si Ti V Zn
Begin Matrices;
 A Full 23 5
 D Diag 23 23
 X Iden 5 5
End Matrices;
Specification A
101  201  301  401  501
0    0    0   0    0
0    0    0   0    0
104  204  304 404  504
0    0    0   0    0
106  206  306 406  506
107  207  307 407  507
108  208  308 408  508
109  209  309 409  509
1010 2010 3010 4010 5010
1011 2011 3011 4011 5011
1012 2012 3012 4012 5012
1013 2013 3013 4013 5013
0    0    0   0    0
1015 2015 3015 4015 5015
1016 2016 3016 4016 5016
1017 2017 3017 4017 5017
1018 2018 3018 4018 5018
1019 2019 3019 4019 5019
0    0    0   0    0
1021 2021 3021 4021 5021
1022 2022 3022 4022 5022
1023 2023 3023 4023 5023
Value  0.098    D 1 1
Value  0.21     D 2 2
Value  0.58     D 3 3
Value  0.06     D 4 4
Value  0.001    D 5 5
Value  0.38     D 6 6
Value  0.28     D 7 7
Value  0.0092   D 8 8
Value  0.00055  D 9 9
Value  0.016    D 10 10
Value  0.017    D 11 11
Value  0.00047  D 12 12
Value  0.004    D 13 13
Value  0.0025   D 14 14
Value  0.0025   D 15 15
Value  0.0025   D 16 16
Value  0.0025   D 17 17
Value  0.0025   D 18 18
Value  0.0025   D 19 19
Value  0.0025   D 20 20
Value  0.0025   D 21 21
Value  0.0025   D 22 22
Value  0.0025   D 23 23
Value  0.008    A 2  1
Value  0.519    A 2  2
Value  0.000    A 2  3
Value  0.171    A 2  4
Value  0.000    A 2  5
Value  0.592    A 3  1
Value  0.187    A 3  2
Value  0.179    A 3  3
Value  0.116    A 3  4
Value  0.114    A 3  5
Value  0.005    A 5  1
Value  0.116    A 5  2
Value  0.159    A 5  3
Value  0.033    A 5  4
Value  0.075    A 5  5
```

```
 Value  0.000    A 14 1
 Value  0.000    A 14 2
 Value  0.004    A 14 3
 Value  0.000    A 14 4
 Value  0.000    A 14 5
 Value  0.000    A 20 1
 Value  0.000    A 20 2
 Value  0.001    A 20 3
 Value  0.043    A 20 4
 Value  0.000    A 20 5
 Boundary 0 1 all
 Labels Row A
! 1   2    3  4  5  6  7  8  9 10 11 12 13 14 15 16 1718 19 20 2122 23
NH4 NO3 SO42 OC EC Al As Br Ca Cl Cr Cu Fe Pb Mn Hg Ni K Se Si Ti V Zn
 Labels Row D
! 1   2    3  4  5  6  7  8  9 10 11 12 13 14 15 16 1718 19 20 2122 23
NH4 NO3 SO42 OC EC Al As Br Ca Cl Cr Cu Fe Pb Mn Hg Ni K Se Si Ti V Zn
 Labels Col D
! 1   2    3  4  5  6  7  8  9 10 11 12 13 14 15 16 1718 19 20 2122 23
NH4 NO3 SO42 OC EC Al As Br Ca Cl Cr Cu Fe Pb Mn Hg Ni K Se Si Ti V Zn
 Labels Col A
! 1        2        3        4        5
  Factor-1 Factor-2 Factor-3 Factor-4 Factor-5
 Start 0 all
 Covariance_model A*X*A' + D ;
 Options RSiduals MxA=/projects/PITT-PM/Source_Apportionment/Data/A.txt
End

Constrain Factor 1 coefficients <= 1
  Constraint
  Begin Matrices;
   A Full 23 5 = A1
   T Unit 1  1
   C Full 1  4
   K Unit 23 1
  End Matrices;
  Matrix C 1 1 23 1
  Constraint \part(A,C)'*K < T;
End

Constrain Factor 2 coefficients <= 1
  Constraint
  Begin Matrices;
   A Full 23 5 = A1
   T Unit 1  1
   C Full 1  4
   K Unit 23 1
  End Matrices;
  Matrix C 1 2 23 2
  Constraint \part(A,C)'*K < T;
End

Constrain Factor 3 coefficients <= 1
  Constraint
  Begin Matrices;
   A Full 23 5 = A1
   T Unit 1  1
   C Full 1  4
   K Unit 23 1
  End Matrices;
  Matrix C 1 3 23 3
  Constraint \part(A,C)'*K < T;
End

Constrain Factor 4 coefficients <= 1
  Constraint
  Begin Matrices;
   A Full 23 5 = A1
   T Unit 1  1
   C Full 1  4
   K Unit 23 1
  End Matrices;
  Matrix C 1 4 23 4
  Constraint \part(A,C)'*K < T;
End

Constrain Factor 5 coefficients <= 1
  Constraint
  Begin Matrices;
   A Full 23 5 = A1
   T Unit 1  1
   C Full 1  4
   K Unit 23 1
  End Matrices;
  Matrix C 1 5 23 5
  Constraint \part(A,C)'*K < T;
End
```

4.5 Constructing a Roadway Exposure Time Series

4.5.1 Introduction

Particulates related to roadway transportation are not separately nor directly measured. Some of these of these particles likely reach ambient $PM_{2.5}$ and PM_{10} monitors. In order to help separate these particles from other sources, an attempt was made to create a daily proxy based on available vehicle mileage data. The usefulness of this approach in modeling will need to be explored as the time series models are constructed. The available data likely provide more information for spatial resolution compared to time resolution.

4.5.2 Annual Averaged Daily Vehicle Mileage Traveled (VMT)

The annually averaged daily VMT data is provided by Pennsylvania Department of Transportation (Penn DOT) in geo-database format that can be directly visualized and retrieved through a geo-database query system in GIS programs. The VMT data is composed of a segment identification number, a start route identification number, a segment length, a traffic pattern group (TPG) classification, an annual averaged daily VMT, an annual averaged daily truck VMT, and so on.

4.5.3 Average Day of Week by Month Factors Compiled for Total Vehicles

Average day of week by month factors for 2004 were provided by Penn DOT in table format. The data is a group factor which can be applied to a 24-hour raw traffic count taken during any day of the year to develop an annual average daily traffic as shown below.

$$\textit{Traffic Counts}_{\textit{Annual Average}} = (\textit{Traffic Counts}_{\textit{taken during any day of the year}})\cdot (\textit{average day of week by month factor})$$

The average day of week by month factors can be used to estimate daily vehicle mileage travel for each traffic segment in the Pittsburgh study area with an assumption that annual averaged daily VMT is proportional to annual averaged daily traffic counts. These daily factors data are also available for trucks. An example set of data for January 2004 is shown **Table 58**.

Table 58: Average day of week by month factors for January 2004.

DAY	TPG 1	TPG 2	TPG 3	TPG 4	TPG 5	TPG 6	TPG 7	TPG 8	TPG 9	TPG 10
Monday	1.110	1.280	1.060	1.149	1.157	1.233	1.170	1.248	1.160	1.383
Tuesday	1.071	1.282	0.998	1.111	1.096	1.203	1.116	1.198	1.096	1.291
Wednesday	1.038	1.276	0.994	1.119	1.099	1.208	1.098	1.209	1.096	1.311
Thursday	1.036	1.254	1.001	1.108	1.100	1.205	1.101	1.205	1.111	1.303
Friday	1.034	1.139	0.971	1.079	1.084	1.096	1.032	1.135	1.081	1.262
Saturday	1.297	1.434	1.355	1.306	1.226	1.372	1.213	1.310	1.283	1.247
Sunday	1.497	1.407	1.750	1.445	1.351	1.574	1.472	1.507	1.524	1.283
Day of Month	1.155	1.296	1.161	1.188	1.159	1.270	1.172	1.259	1.193	1.297

The traffic pattern group (TGP) is defined in **Table 59**.

Table 59: Traffic pattern group (TPG).

	Description
TPG 1	URBAN - INTERSTATE
TPG 2	RURAL - INTERSTATE
TPG 3	URBAN - OTHER PRINCIPAL ARTERIALS
TPG 4	RURAL - OTHER PRINCIPAL ARTERIALS
TPG 5	URBAN - MINOR ARTERIALS, COLLECTORS, LOCAL ROADS
TPG 6	NORTH RURAL - MINOR ARTERIALS
TPG 7	CENTRAL RURAL- MINOR ARTERIALS
TPG 8	NORTH RURAL - COLLECTORS AND LOCAL ROADS
TPG 9	CENTRAL RURAL- COLLECTORS AND LOCAL ROADS
TPG 10	SPECIAL RECREATIONAL

4.5.4 Hourly Percentages Compiled for Total Vehicles

Hourly percentages of total vehicles sorted by traffic pattern group for the year 2004 was provided by Penn DOT in table format. The data is a group factor which can be applied to less than 24-hour averaged raw traffic counts. The hourly percentage data can be used to estimate hourly vehicle mileage travel for each traffic segment of study area in the Pittsburgh area. An example table for hourly percentages by traffic pattern groups for the year 2004 is shown **Table 60**.

Table 60: Total vehicle hourly percentages.

Hour	TPG1	TPG2	TPG3	TPG4	TPG5
1	1.30%	1.50%	0.70%	0.90%	0.80%
2	1.00%	1.30%	0.50%	0.60%	0.50%
3	0.90%	1.20%	0.40%	0.50%	0.40%
4	0.90%	1.20%	0.50%	0.60%	0.40%
5	1.20%	1.40%	0.90%	1.00%	0.70%
6	2.50%	2.30%	2.60%	2.60%	2.10%
7	5.60%	3.80%	6.20%	5.70%	5.30%
8	7.80%	5.10%	8.50%	7.40%	7.80%
9	6.50%	5.10%	7.00%	6.40%	6.70%
10	5.40%	5.40%	5.40%	5.50%	5.30%
11	5.30%	5.90%	5.20%	5.40%	5.10%
12	5.40%	6.20%	5.30%	5.50%	5.40%
13	5.40%	6.00%	5.50%	5.70%	5.80%
14	5.40%	6.10%	5.60%	5.70%	5.80%
15	5.80%	6.40%	6.10%	6.30%	6.10%
16	6.50%	6.80%	6.70%	6.90%	6.90%
17	6.60%	6.90%	6.90%	7.10%	7.20%
18	6.40%	6.30%	6.70%	6.80%	7.00%
19	5.20%	5.10%	5.40%	5.40%	5.60%
20	4.10%	4.20%	4.10%	4.10%	4.40%
21	3.40%	3.70%	3.30%	3.30%	3.70%
22	3.00%	3.20%	2.80%	2.90%	3.00%
23	2.50%	2.80%	2.20%	2.30%	2.20%
24	1.90%	2.20%	1.40%	1.50%	1.50%
Total	100%	100%	100%	100%	100.00%

4.5.5 Daily Traffic Counts for 2000 to 2006

Daily traffic counts were provided from Penn DOT. The data were randomly collected from 2000 to 2006 by segments. Annual averaged traffic counts can be estimated based on the daily traffic counts and the average day of week by month factors.

4.5.6 Estimation of Daily Vehicle Mileage Traveled for 2004

Daily VMT can be estimated based on annual average daily VMT and the average day of week by month factors. Initially the average day of week by month factors were induced by Penn DOT to be applied to 24-hour raw traffic count taken during any day of the year to develop an annual average daily traffic. Daily VMT can be assumed to be in proportion to 24-hour raw traffic count data. The following equation can be used to estimate daily VMT by each segment for 365 days of the year 2004.

$$VMT_{Daily} = (VMT_{annual\ averaged\ daily}) / (average\ day\ of\ week\ by\ month\ factor)$$

The results will be similar to the spatial distribution map shown **Figure 78** for daily VMT on January 5, 2004.

This method has limitations in terms of accuracy of daily VMT estimation for individual days of a month. For example, the daily VMT on Mondays for a given month will be the same. However, this approach may provide reasonable temporal and spatial resolution of vehicle mileage travel pattern in the Pittsburgh area in order to compare with health outcomes and air quality pollutants.

As an example, two route segments close to a hospital in the Pittsburgh region were selected, (**Figure 79**)

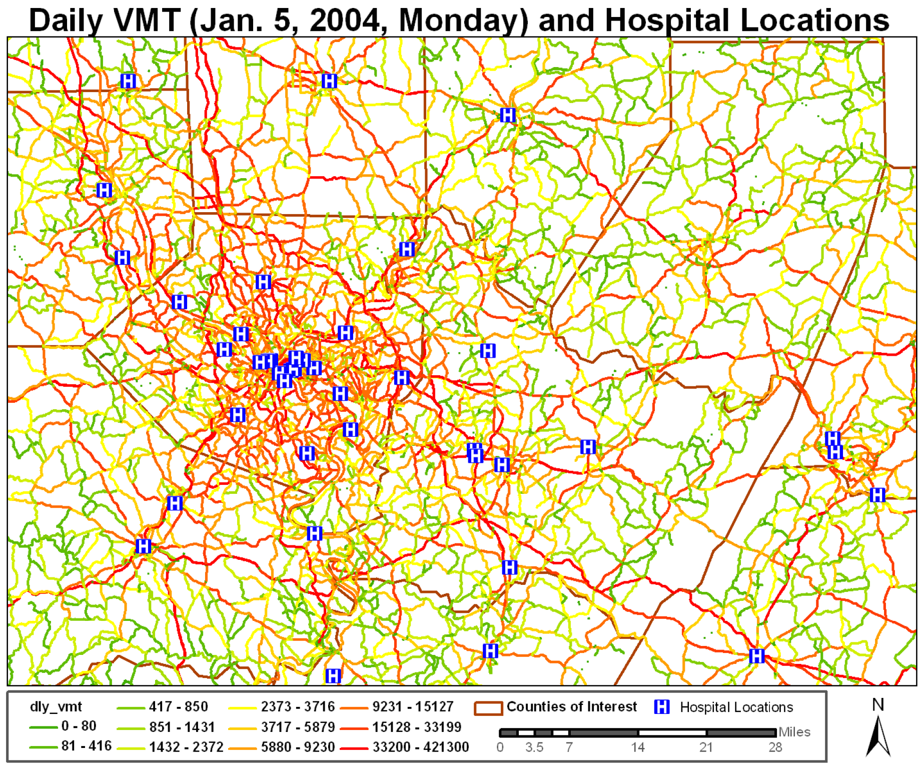


Figure 78: Daily VMT for Monday, January 5, 2004 and hospital locations.

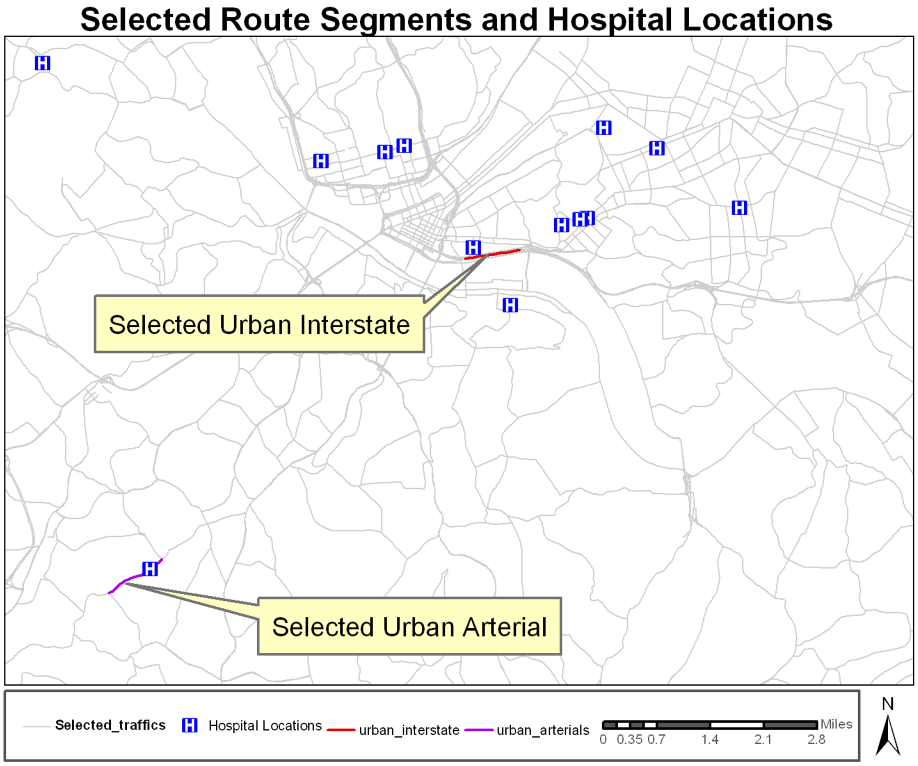


Figure 79: Selected route segments and hospital locations.

and the average day of week by month factors were applied to the annual average daily VMT of the two route segments utilizing the above equation for June to August 2004. The lengths of the selected urban interstate and urban arterial are 1144 and 1359 meters respectively. Daily VMT level was the lowest on Sundays and the highest on Fridays for the selected urban interstate and arterial routes. Daily VMT levels in the selected urban arterial was significantly lower than in the selected urban interstate during weekends.

4.5.7 Estimation of Hourly Vehicle Mileage Traveled for 2004

Hourly VMT can be estimated based on annual average daily VMT and the total vehicle hourly percentages. Initially the hourly percentages were calculated by Penn DOT to be applied to less than 24-hour raw traffic count. Hourly VMT can be assumed to be in proportion to less than 24-hour raw traffic count data. Thus the following equation can estimate hourly VMT by each segment for 24 hours of a day in the year 2004:

$$VMT_{hourly} = (VMT_{daily}) \cdot (Hourly\ percentages\ for\ total\ vehicles)$$

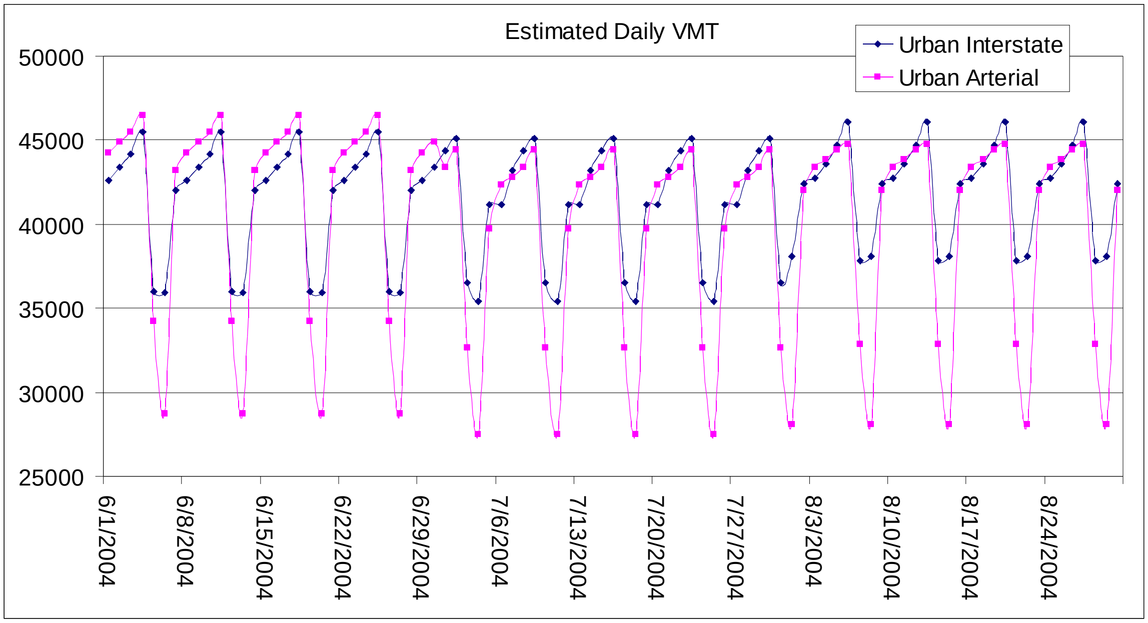


Figure 80: Estimated daily VMT.

4.6 Generalized Linear Autoregressive Moving Average Models

4.6.1 Introduction

The health outcomes will be counts so that a generalized model using a Poisson link would be needed. To appropriately account for the inevitable autocorrelation in the health outcome time series (even after accounting for the effects of explanatory factors), a generalized autoregressive moving average model (GLARMA) as developed by Davis (1999, 2003, 2005) will be used:

$$\text{Model: } Y_t | \mu_t \text{ is } Poisson(\mu_t) \text{ with } \log(\mu_t) = X_t^T \beta + Z_t$$

$$\text{For } \lambda \geq 0 \text{ define } e_t = (Y_t - \mu_t)/\mu_t^\lambda$$

Form a linear process using e_t:

$$\log \mu_t = X_t^T \beta + Z_t$$

where

$$Z_t = \sum_{i=1}^{\infty} \psi_i e_{t-i}$$

Although software for estimating the parameters of the GLARMA model is not generally available, we have GLARMA estimation code contributed by Drescher (2005, 2006) for use in the R environment for statistical computation and graphics. Analysis of the use of GLARMA models with daily $PM_{2.5}$ mass concentration and elderly hospital admissions shows that it is important to account for the autocorrelation typically exhibited in the data otherwise the coefficient estimates and their standard errors are not correctly estimated. Typically, the standard errors are underestimated causing statistical significance to be exaggerated. Researchers have focused on non-time series type models (notably generalized additive models (GAM) and generalized linear models (GLM)) and have largely ignored handling autocorrelation that persists after including explanatory covariates.

We propose forming separate time series of health counts and explanatory factors for each county and possibly at the ZIP code level and then combining these into one comprehensive model to allow an overall estimate of effects and the variation in effects by county and/or ZIP code. The county / ZIP code effect will be treated as a random effect if a large enough number of areas are available or if not, as a fixed effect.

4.6.2 Example GLARMA Model Estimation of the $PM_{2.5}$ Effect on Elderly Hospital Admissions

The ACAPS/NMMAPS data used for the power analyses was used to illustrate the advantages of GLARMA time series models over GAM and GLM non-time-series models. A GLM model was fitted for elderly hospital admissions and the results are shown in **Listing 22**. (The R code for this analysis is shown in **Listing 24**.) The relative size of the coefficients and the p-values are very similar to the regression model coefficients estimated for the power analysis. The standard deviation of the residuals is 1.67 and the range is from -11.99 to 6.10. The histogram and normal probability plots for the GLM model residuals are shown in **Figure 81**. The residual distribution is roughly Normal but somewhat negatively skewed. GLM (and GAM) models do not provide a way to model autocorrelation in the residuals. The deficiency of the fitted GLM model is illustrated in the time series plot and the ACF and PACF graphs shown in **Figure 82**. The PACF indicates the need for autoregressive components up to and including lag 8. The correlation for lag 1 in the PACF is about 0.5. In most models, it is virtually impossible to include all the relevant explanatory factors that would account for all the dependence in the response.

A GLARMA model with eight autoregressive parameters (lags 1 through 8: $\phi_1, \phi_2, \dots, \phi_8$) fitted to the same data with the same explanatory factors is shown in **Listing 23**. The estimated autoregressive coefficients are all highly statistically significant with the exception of lag 7. The standard deviation of the residuals for the GLARMA model is 1.42 and they range from -8.5 to 4.30. **Table 61** provides a comparison of the descriptive statistics for the residual errors from both models. The standard deviation, the interquartile range, and the range are substantially smaller for the GLARMA residuals compared to the GLM residuals. The histogram and normal probability plots shown in **Figure 83** are more Normal (with less negative skewness) compared to those for the GLM residuals. The time series plot and ACF and PACF graphs shown in **Figure 84** show little if any evidence of non-randomness and all of the correlations in both the ACF and PACF graphs fall within the two standard error limits. All the evidence in **Table 61** indicates that the residuals for the GLARMA model have less variability. The residual time series plot is much more homogeneous and shows almost no non-random patterns in contrast for the GLM plot. Clearly, the GLARMA time-series model is a substantial improvement over the GLM non-time-series model. The estimated coefficient for $PM_{2.5}$ for the GLARMA model (0.0002) is less than half of the estimate (0.0005) for the GLM model. Similarly, the p-value for the GLARMA model (0.6489) is more than two and one-half times that of the GLM model (0.235).

Listings 24 - 27show the R code used for GLARMA estimates. The routines for GLARMA estimation were written by and kindly provided by Daniel Drescher.

Listing 22: R output for Poisson non-time-series estimation using GLM. The coefficient estimates are used as starting values for the GLARMA estimation.

```
> #Poisson non-timeseries estimation
> GLM <- glm(Y~-1+X, family = poisson, x = T)
> summary(GLM)

Call:
glm(formula = Y ~ -1 + X, family = poisson, x = T)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-11.98624   -0.82050    0.07845    0.91292    6.09621

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
Xsun       5.9088416  0.2450335  24.114  < 2e-16 ***
Xmon       6.4073235  0.2435913  26.304  < 2e-16 ***
Xtue       6.3714463  0.2435174  26.164  < 2e-16 ***
Xwed       6.3362099  0.2434679  26.025  < 2e-16 ***
Xthu       6.3117545  0.2437341  25.896  < 2e-16 ***
Xfri       6.2745184  0.2447092  25.641  < 2e-16 ***
Xsat       5.9423186  0.2447792  24.276  < 2e-16 ***
Xtp       -0.0001528  0.0000218  -7.007 2.43e-12 ***
Xcs        0.1265367  0.0135518   9.337  < 2e-16 ***
Xsn        0.0924945  0.0070599  13.101  < 2e-16 ***
Xpm25mean -0.0005034  0.0004237  -1.188    0.235
Xso2      -0.9923606  0.9111303  -1.089    0.276
Xno2       0.6529158  0.7861352   0.831    0.406
Xno        0.0249012  0.2649844   0.094    0.925
Xozone     0.2586509  0.4735412   0.546    0.585
Xmntp      0.0028332  0.0004921   5.758 8.53e-09 ***
Xmnrh      0.0001718  0.0002523   0.681    0.496
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 525758.1  on 607  degrees of freedom
Residual deviance:   1684.2  on 590  degrees of freedom
AIC: 5689.2

Number of Fisher Scoring iterations: 4

# sd of GLM residuals
> sd(resid(GLM))
[1] 1.666426

# Interquartile Range
> IQR(resid(GLM))
[1] 1.733423
```

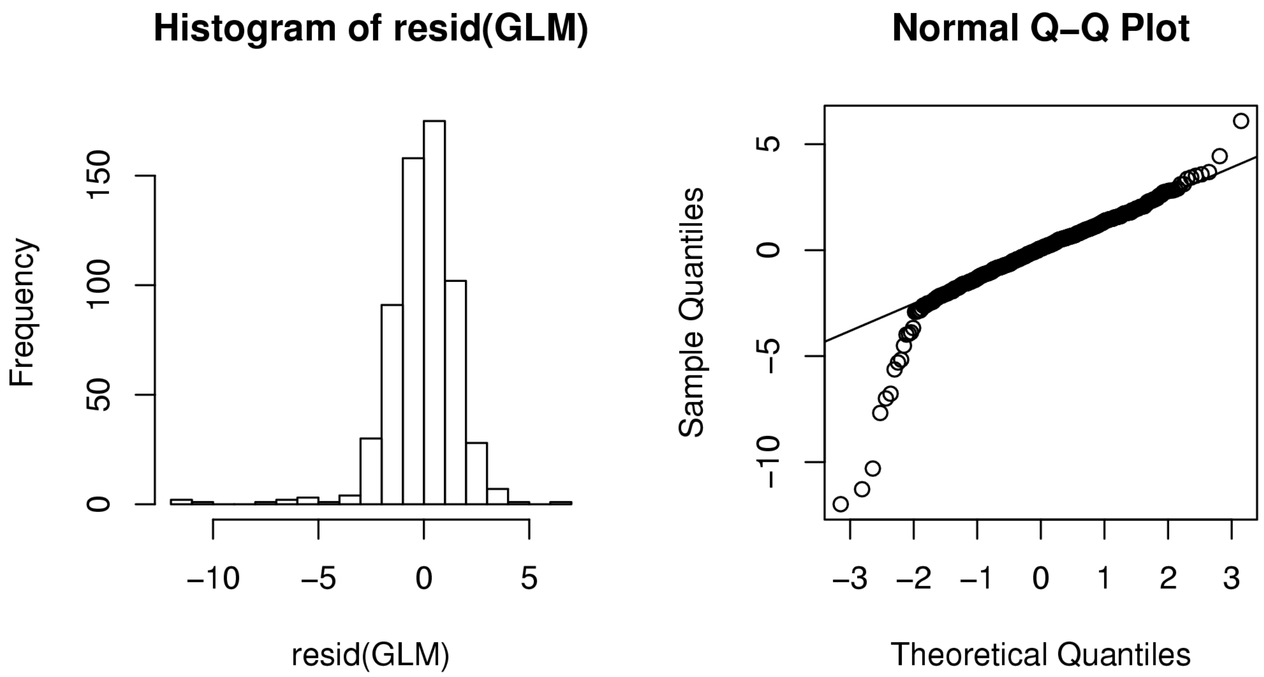


Figure 81: Histogram and normal probability plot for GLM residuals.

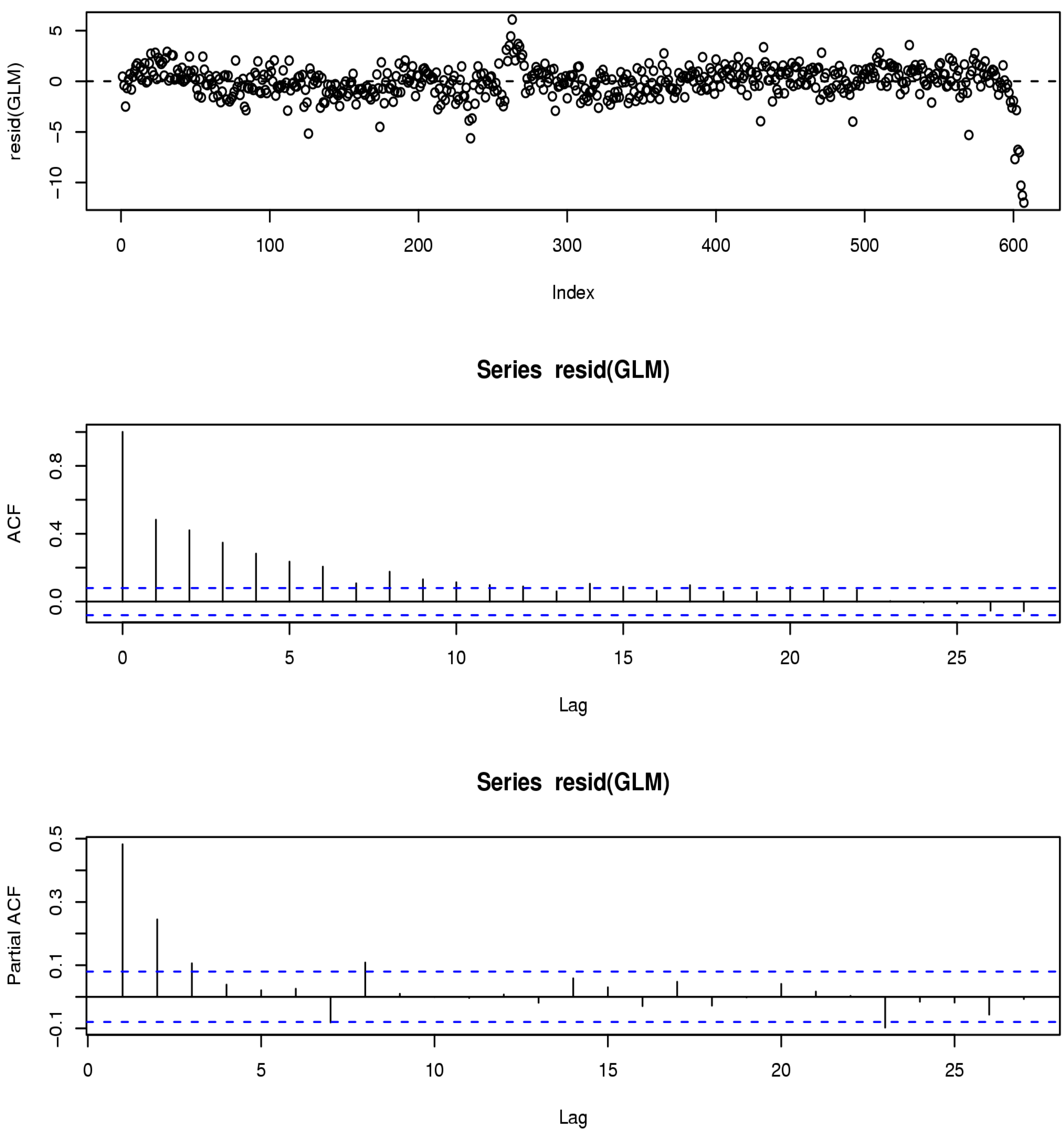


Figure 82: Time series plot and ACF and PACF plots for GLM residuals.

Listing 23: R output results for GLARMA estimation.

```
> GL <- Poisson_GLARMA_NR(Y,X,delta0,length(GLM$coefficients),phi.lags,theta.lags,
+   round, maxit, conv,lambda, type)
There were 50 or more warnings (use warnings() to see the first 50)
> GL$results
      estimate   s.e. t-value p-value derivative of ll
 [1,]   6.4706 0.6930  9.3367  0.0000    -6.473044e-12   Sunday
 [2,]   6.9756 0.6928 10.0692  0.0000     1.083011e-11   Monday
 [3,]   6.9311 0.6928 10.0051  0.0000    -9.616530e-12   Tuesday
 [4,]   6.8910 0.6927  9.9474  0.0000     4.654055e-12   Wednesday
 [5,]   6.8767 0.6928  9.9255  0.0000    -7.861267e-12   Thursday
 [6,]   6.8329 0.6931  9.8581  0.0000     1.525891e-12   Friday
 [7,]   6.5068 0.6929  9.3900  0.0000     5.886847e-12   Satday
 [8,]  -0.0002 0.0001 -3.1855  0.0014    -6.810296e-09   tp
 [9,]   0.0891 0.0219  4.0722  0.0000    -3.286260e-13   cs
[10,]   0.0754 0.0186  4.0566  0.0000    -5.528911e-13   sn
[11,]  -0.0002 0.0004 -0.4553  0.6489     4.092726e-11   PM2.5
[12,]   1.2461 0.8820  1.4127  0.1577    -2.405021e-14   SO2
[13,]   0.0117 0.7699  0.0152  0.9879    -1.476597e-14   NO2
[14,]  -0.2970 0.2827 -1.0507  0.2934    -3.186340e-14   NO
[15,]  -0.4008 0.4889 -0.8198  0.4123     2.285672e-14   ozone
[16,]   0.0023 0.0006  3.7422  0.0002    -3.899459e-11   mntp
[17,]   0.0000 0.0003  0.0000  1.0000    -5.115908e-12   mnrh
[18,]   0.0336 0.0030 11.2462  0.0000     4.348522e-12   1
[19,]   0.0322 0.0032 10.2083  0.0000    -2.643219e-12   2
[20,]   0.0289 0.0033  8.7038  0.0000     8.071765e-12   3
[21,]   0.0246 0.0033  7.4138  0.0000     4.789058e-12   4
[22,]   0.0202 0.0032  6.3266  0.0000    -3.055334e-12   5
[23,]   0.0128 0.0031  4.0968  0.0000    -2.404477e-11   6
[24,]  -0.0006 0.0030 -0.2026  0.8394     5.524470e-12   7
[25,]   0.0125 0.0031  4.0248  0.0001     1.359268e-11   8

# sd of GLARMA residuals
> sd(GL$e)
[1] 1.419827

# Minimum and Maximum
> range(GL$e)
[1] -8.49780  4.29599

# Interquartile Range
> IQR(GL$e)
[1] 1.603282

# Range
> diff(range(GL$e))
[1] 12.79379

> summary(GL$e)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-8.49800 -0.77450  0.02851 -0.03108  0.82880  4.29600
```

Table 61: Comparison of residual error for the GLM non-time-series model and the GLARMA time series model.

Summary Statistic	***Model***	
	GLM	***GLARMA***
Mean	-0.05	-0.03
Median	0.08	0.03
Standard deviation	1.67	1.42
Interquartile range	1.73	1.60
Range	18.08	12.79
Minimum	-11.99	-8.50
Maximum	6.10	4.30

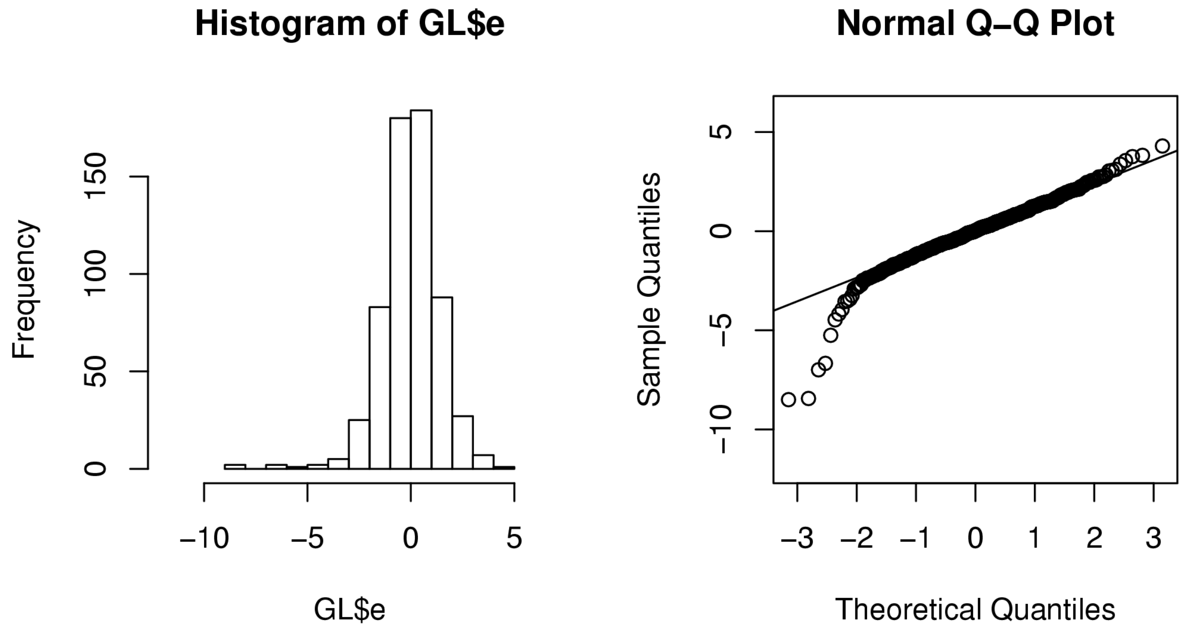


Figure 83: Histogram and normal probability plot for GLARMA residuals.

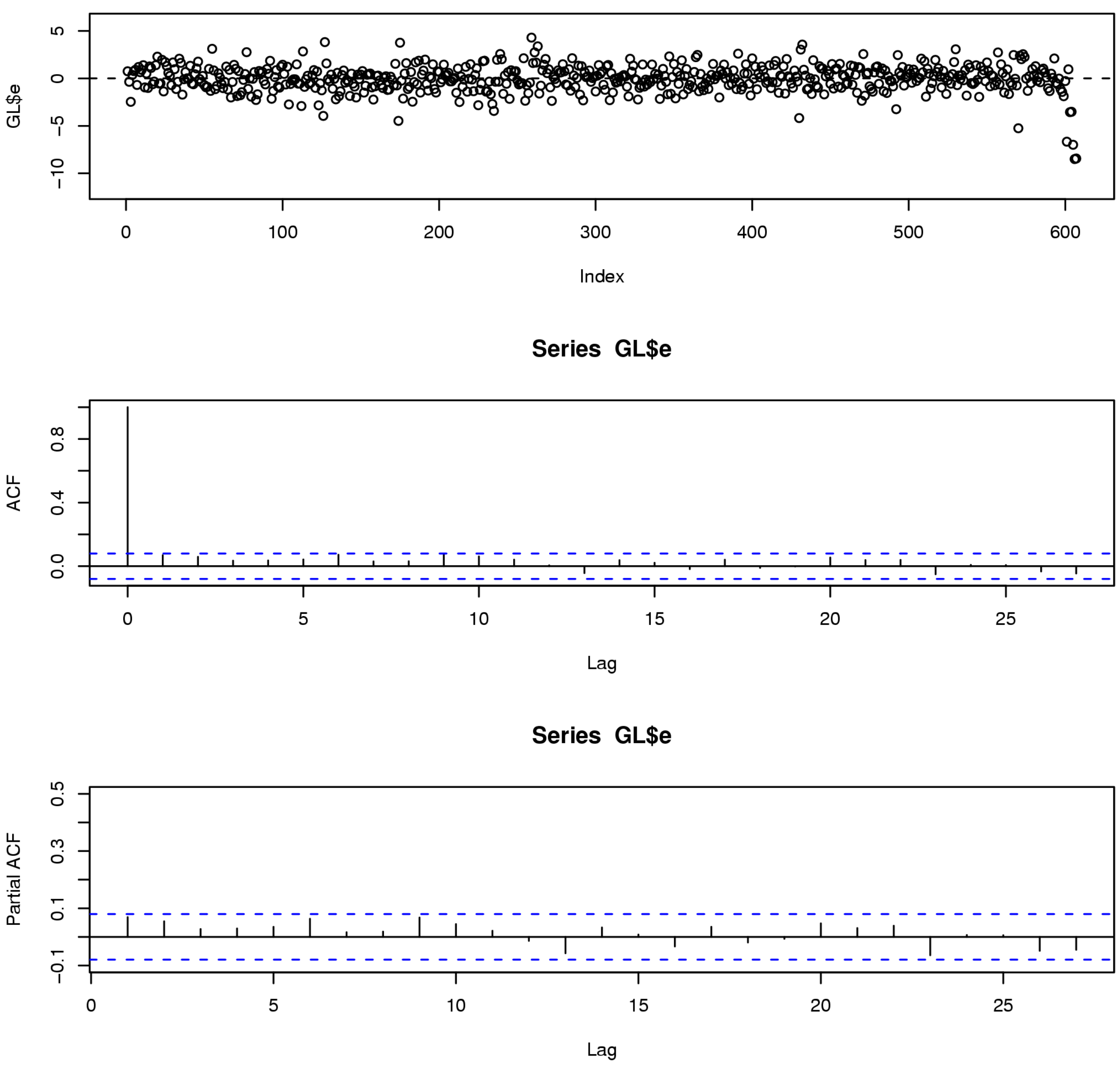


Figure 84: Time series plot and ACF and PACF plots for GLARMA residuals.

Listing 24: R code for GLM and GLARMA estimation of the $PM_{2.5}$ effect on elderly hospital admissions.

```
# Modified slightly by Rick Bilonick
####################################################
# This program estimates an POisson-GLARMA(p,q)    #
# model as proposed by Davis et al. (1999, 2003).  #
# The program offers the option to use alternative #
# model-driving residuals in the GLARMA model.     #
# This program has been written by Daniel Drescher #
# July 2005. (Drescher.D@gmx.net)                  #
####################################################

# Full data set with missing values
dow <- factor(as.character(weekdays(complete.copy$date)),
  levels=c("Sat","Sun","Mon","Tue","Wed","Thu","Fri"))
tp <- as.numeric(complete.copy$date)
sun <- ifelse(dow=="Sun",1,0)
mon <- ifelse(dow=="Mon",1,0)
tue <- ifelse(dow=="Tue",1,0)
wed <- ifelse(dow=="Wed",1,0)
thu <- ifelse(dow=="Thu",1,0)
fri <- ifelse(dow=="Fri",1,0)
sat <- ifelse(dow=="Sat",1,0)

cs <- cos(2 * pi * tp/365)
sn <- sin(2 * pi * tp/365)

x3 <- with(complete.copy,data.frame(sun,mon,
  tue,wed,thu,fri,sat,tp=tp,cs=cos(2 * pi * tp/365),
  sn=sin(2 * pi * tp/365),
  pm25mean,so2,no2,no,ozone,mntp,mnrh))

#read functions - changed the paths - rab
source("/projects/glarma/R/GLARMA_Example/Methods/Poisson_GLARMA_NR.txt")
source("/projects/glarma/R/GLARMA_Example/Methods/Poisson_DDW.txt")

dow <-
factor(as.character(weekdays(complete.copy$date[!is.na(complete.copy$pm25mean)])),
  levels=c("Sat","Sun","Mon","Tue","Wed","Thu","Fri"))
tp <- as.numeric(complete.copy$date[!is.na(complete.copy$pm25mean)])
sun <- ifelse(dow=="Sun",1,0)
mon <- ifelse(dow=="Mon",1,0)
tue <- ifelse(dow=="Tue",1,0)
wed <- ifelse(dow=="Wed",1,0)
thu <- ifelse(dow=="Thu",1,0)
fri <- ifelse(dow=="Fri",1,0)
sat <- ifelse(dow=="Sat",1,0)

cs <- cos(2 * pi * tp/365)
sn <- sin(2 * pi * tp/365)

x <- with(complete.copy[!is.na(complete.copy$pm25mean),],data.frame(sun,mon,
  tue,wed,thu,fri,sat,tp=tp,cs=cos(2 * pi * tp/365),
  sn=sin(2 * pi * tp/365),
  pm25mean,so2,no2,no,ozone,mntp,mnrh))
x2 <- with(complete.copy[!is.na(complete.copy$pm25mean),],data.frame(dow,
  tp,cs=cos(2 * pi * tp/365),
  sn=sin(2 * pi * tp/365),
  pm25mean,so2,no2,no,ozone,mntp,mnrh))

#read data - changed the path - rab
#Data <- read.csv("/projects/glarma/R/GLARMA_Example/Data/Asthma_CTown.csv",
sep=";")
Y<-complete.copy[!is.na(complete.copy$pm25mean),2]
#Y<-complete.copy[,2]
X<-as.matrix(x[,1:17])
#X2<-as.matrix(x2[,1:11])
#X<-as.matrix(x3[,1:17])


# Set model parameters

#theta.lags <- c(7)
#phi.lags <- rep(0,0)
theta.lags <- c(1,2,3)
phi.lags <- rep(0,0)

round <- 4          # sets the numbers of digits after the dot for the results
conv <- 1e-10             # convergence criterion based on the gradient
maxit <- 10         # maximum number of iterations
lambda <- 1.0             # lambda for the scaled deviation
type <-  "FT"             # type of model-driving residual:
                    #    "SD" - scaled deviation,
                    #    "VS" - variance stabilized residuals
                    #    "A"  - Anscombe residuals,
                    #    "FT" - Freeman tukey residuals,
                    #    "NP" - normal pseudo residuals,

# Set Initial Values
#theta.init <- c(0.0)
#phi.init <-  rep(0,0)
theta.init <- rep(0,3)
phi.init <-  rep(0,0)
N <-  length(Y)

#Poisson non-timeseries estimation
GLM <- glm(Y~-1+X, family = poisson, x = T)
#GLM2 <- glm(Y~dow+tp+cs+sn+pm25mean+so2+no2+no+ozone+mntp+mnrh, family = poisson,
x = T,data=x)

# initial parameter vector for time series model
delta0 <- c(GLM$coefficients, phi.init, theta.init)


# time series estimation by using Newton-Raphson methode
GL <- Poisson_GLARMA_NR(Y,X,delta0,length(GLM$coefficients),phi.lags,theta.lags,
  round, maxit, conv,lambda, type)
GL$results

# value of the loglikelihood function at the estimates
GL$ll
```

```
#numbers of iterations for reaching convergence
GL$iterations


# results estimates (1st column), standard error (2nd column), t-value (3rd
column),
#     two-sided probability (4 column), value of the gradient at the estimates (5th
column)
GL$results

acf(GL$e)
```

Listing 25: R function "Poisson_GLARMA_NR." Written by Daniel Drescher.

```
Poisson_GLARMA_NR <- function(Y, X, delta, r, phi.lags, theta.lags, round, maxit, conv,  lambda, type)
{
counter<-0
maxgrad<- 1

        n <- length(Y)
        p <- length(phi.lags)
        q <- length(theta.lags)


while((counter< maxit)&(maxgrad>conv)){

        beta <- delta[1:r]
        phi <- delta[(r + 1):(r + p)]
        theta <- delta[(r + p + 1):(r + p + q)]
        mpq <- 0
        if((p + q) > 0) {
                mpq <- max(phi.lags[p], theta.lags[q])
        }
        nmpq <- n + mpq
        s <- r + p + q
        e <- array(0, nmpq)
        Z <- array(0, nmpq)
        W <- array(0, nmpq)
        mu <- array(0, nmpq)
        e.d <- array(0, c(s, nmpq))
        Z.d <- array(0, c(s, nmpq))
        W.d <- array(0, c(s, nmpq))
        e.dd <- array(0, c(s, s, nmpq))
        Z.dd <- array(0, c(s, s, nmpq))
        W.dd <- array(0, c(s, s, nmpq))
        eta <- X %*% beta

        ll <- 0
        ll.d <- matrix(0, ncol = 1, nrow = s)
        ll.dd <- matrix(0, ncol = s, nrow = s)
        for(time in 1:n) {
                tmpq <- time + mpq

                if(p > 0) {

                        Z.d[(r + 1):(r + p), tmpq] <- Z[tmpq -phi.lags] + e[tmpq - phi.lags]
                        Z.dd[(r + 1):(r + p),  , tmpq] <- t((Z.d + e.d)[, (tmpq -phi.lags)])
                        Z.dd[, (r + 1):(r + p), tmpq] <- Z.dd[, (r + 1):(r + p), tmpq] +
                                                          (Z.d + e.d)[, (tmpq -phi.lags)]

                        for(i in 1:p) {

                                Z[tmpq] <- Z[tmpq] + phi[i] * (Z + e)[tmpq - phi.lags[i]]
              Z.d[, tmpq] <- Z.d[, tmpq] + phi[i] * (Z.d[, tmpq - phi.lags[i]] +
                              e.d[, tmpq - phi.lags[i]])
                                Z.dd[,  , tmpq] <- Z.dd[,  ,tmpq] + phi[i] * (Z.dd[,  ,tmpq - phi.lags[i]]
+
                                                  e.dd[,  , tmpq - phi.lags[i]])
                        }}

                if(q > 0) {

                        Z.d[(r + p + 1):(r + p + q), tmpq] <- e[tmpq - theta.lags]
                        Z.dd[(r + p + 1):(r + p + q),  ,tmpq] <- t(e.d[, tmpq - theta.lags])
                        Z.dd[, (r + p + 1):(r + p + q), tmpq] <- Z.dd[, (r + p + 1):(r + p + q), tmpq] +
                                                                e.d[, tmpq - theta.lags]

                        for(i in 1:q) {

                                Z[tmpq] <- Z[tmpq] + theta[i] * e[tmpq - theta.lags[i]]
                                Z.d[, tmpq] <- Z.d[, tmpq] + theta[i] * e.d[, tmpq - theta.lags[i]]
                                Z.dd[,  , tmpq] <- Z.dd[,  ,tmpq] + theta[i] * e.dd[,  ,tmpq -
theta.lags[i]]
                        } }


                W[tmpq] <- eta[time] + Z[tmpq]
                W.d[, tmpq] <- matrix(c(X[time,  ], rep(0, p +q)), ncol = 1) + Z.d[, tmpq]
                W.dd[,  , tmpq] <- Z.dd[,  , tmpq]
                mu[tmpq] <- exp(W[tmpq])

        if(type=="SD"){
                e[tmpq] <- (Y[time] - mu[tmpq])/mu[tmpq]^(lambda)
                e.W<- -(mu[tmpq]^(1-lambda) + lambda * e[tmpq])
                e.WW<- ((2*lambda-1)*mu[tmpq]^(1-lambda)+ (lambda^2)*e[tmpq])
                }

        if(type=="VS"){
                e[tmpq] <- 2*(Y[time]^(0.5) - mu[tmpq]^(0.5))
                e.W<- -(mu[tmpq]^(0.5))
                e.WW<- - 0.5*(mu[tmpq]^(0.5))
                }

        if(type=="A"){
                e[tmpq]<- 1.5*(Y[time]^(2/3) - mu[tmpq]^(2/3))/(mu[tmpq]^(1/6))
                e.W<-  -(mu[tmpq]^(0.5)+ (1/6)*e[tmpq])
                e.WW<-  -(1/3)*mu[tmpq]^(0.5)+ (1/36)*e[tmpq]
                }

        if(type=="FT"){
                e[tmpq] <-   Y[time]^(1/2) +( Y[time]+1)^(1/2) - (4*mu[tmpq]+1)^(1/2)
                e.W<- -(2*mu[tmpq])/((4*mu[tmpq]+1)^(1/2))
                e.WW<- -(2*mu[tmpq]*(2*mu[tmpq]+1))/((4*mu[tmpq]+1)^(3/2))
                }

        if(type=="NP"){
                e[tmpq] <- qnorm(ppois(Y[time],mu[tmpq], lower.tail=TRUE, log.p=FALSE) , mean=0, sd=1,
lower.tail=TRUE, log.p=FALSE)
                DFW<-Poisson_DDW(Y[time],mu[tmpq])
                e.W<- DFW$DF*(1/(dnorm(e[tmpq], mean=0, sd=1, log=FALSE)))
                e.WW<- e[tmpq]*(e.W^2)+DFW$DDF*(1/(dnorm(e[tmpq], mean=0, sd=1, log=FALSE)))
                }


                e.d[, tmpq] <- e.W*W.d[, tmpq]
                e.dd[,  , tmpq] <- e.W*W.dd[,  , tmpq]+e.WW*W.d[, tmpq] %o% W.d[, tmpq]

        #update likelihood and derivatives.
                ll <- ll + Y[time] * W[tmpq] - mu[tmpq] - log(factorial(Y[time]))
                ll.d <- ll.d + (Y[time] - mu[tmpq]) * W.d[,tmpq]
                ll.dd <- ll.dd + (Y[time] - mu[tmpq]) * W.dd[,  , tmpq] -
                          mu[tmpq] * W.d[, tmpq] %o% W.d[, tmpq]
```

Listing 26: R function "Poisson_DDW."

```
Poisson_DDW <- function(Y, Mu)
{
DF<-0
DDF<-0
        for(k in 0:Y) {
                DF<- DF + (k-Mu)*(Mu^(k)*exp(-Mu))/(factorial(k))
                DDF<- DDF + (-Mu + (k-Mu)^2)*(Mu^(k)*exp(-Mu))/(factorial(k))
                 }

list(DF=DF, DDF=DDF)
```

Listing 27: R code for GLM diagnostic residual plots.

```
postscript("/projects/PITT-PM/GLARMA_PM2.5/PS/glm_acf_pacf.eps",height=7,
  width=7,onefile=FALSE,
  horizontal=FALSE,paper="special")
par(mfrow=c(3,1))
plot(resid(GLM))
abline(h=0,lty=2)
acf(resid(GLM))
pacf(resid(GLM))
dev.off()
system("evince /projects/PITT-PM/GLARMA_PM2.5/PS/glm_acf_pacf.eps &")
system("oodraw /projects/PITT-PM/GLARMA_PM2.5/PS/glm_acf_pacf.eps &")

postscript("/projects/PITT-
PM/GLARMA_PM2.5/PS/glm_hist.eps",height=4,width=7,onefile=FALSE,
  horizontal=FALSE,paper="special")
par(mfrow=c(1,2))
hist(resid(GLM),breaks=seq(-12,7,1))
qqnorm(resid(GLM))
qqline(resid(GLM))
dev.off()
system("evince /projects/PITT-PM/GLARMA_PM2.5/PS/glm_hist.eps &")
system("oodraw /projects/PITT-PM/GLARMA_PM2.5/PS/glm_hist.eps &")
```

4.7 Case-Crossover Analysis

The case-crossover design, initially proposed by Maclure (1991), provides an attractive approach to estimating the effects of environmental triggers on acute health outcomes. The design is an alternative to time-series analysis for assessing the acute health effects of air pollution. The application of case-crossover analysis of daily mortality and particulate matter was first carried out by Neas et al. (1999) in Philadelphia, Pennsylvania. Subsequently, the case-crossover design has been widely applied to assess the association between air pollution and adverse health effects, including mortality and cardiopulmonary hospitalizations (D'Ippoliti et al., 2003; Kwon et al., 2001; Sunyer et al., 2000; Tsai et al., 2006; Zanobetti and Schwartz, 2005).

In the case-crossover design, only cases are involved and the exposure of each case during an at-risk "hazard period" just before the event is compared with the exposure levels during one or more reference periods when the event did not occur. Cases serve as their own controls. Therefore, time-invariant confounding factors, such as individual characteristics (e.g., age, gender, race, socioeconomic status, etc.) are controlled by design rather than by statistical adjustment.

The case period is usually defined as the current day of the event or alternatively 0 to 1 or 2 days before the event. The control periods are chosen from time periods that precede the event. However, if the exposures exhibit a time trend, risk estimates from unidirectional sampling could be confounded by the time trend in such an exposure (Greenland, 1996). Bateson and Schwartz (1999) proposed a symmetric bidirectional case-crossover design, in which "control" periods are selected as the same day of week as the case period both before and after the event. This strategy of referent sampling is used in most case-crossover studies of air pollution and would be used for the proposed retrospective study.

In many air pollution studies, the case-crossover analysis will produce results similar to those from a time-series analysis (Basu et al., 2005; Lu and Zeger, 2006). In comparison with time series analysis, the case-crossover approach has a few significant strengths. First, it avoids complex mathematical modeling and adjusting for seasonality because this approach controls some confounding factors such as long-term trend, seasonality and day of week by design rather than by modeling. Moreover, personal characteristics and other time-invariant variables are also controlled by the design. However, the drawback of case crossover design is that the efficiency of case-crossover design estimators has been shown to be lower than that of time series analysis (Bateson and Schwartz, 1999; Pope, 1999). Therefore, both approaches have their strengths and weaknesses.

In the proposed retrospective study, both types of analyses will be used to examine the associations between $PM_{2.5}$ and health outcomes and the comparability and consistency of results will be assessed.

4.8 Spatial Bootstrap Sampling for Confidence Intervals and P-Values

The daily health outcome count time series for each ZIP code area are likely to be positively correlated and this correlation is likely to increase as the areas in question are closer together. In order to properly assess the statistical significance of an estimated parameter or construct the confidence interval for the parameter in the GLARMA model which incorporates all ZIP code areas, it will be necessary to take this spatial correlation into account. The basic idea of spatial bootstrap sampling (Lahiri, 2003) is similar to block bootstrap sampling discussed in section 4.4.2 that would be used for handling dependence in one dimension for the latent variable multivariate receptor model. The basic idea is to randomly select clusters of ZIP code areas where the included ZIP code areas are adjacent or close together in order to preserve the spatial dependence.

5 Work Plan

Tasks To Be Performed

The PITT-PM study will characterize the relationship between human health and ambient airborne fine particles ($PM_{2.5}$) from coal-fired power plants and other emission sources in the Pittsburgh, Pennsylvania region. The proposed study area has approximately 2.4 million people, almost one fifth of the total Pennsylvania population in 2002. More than one half of the population of the Pittsburgh Metropolitan Statistical Area (MSA) resides in Allegheny County (1.24 million; 83.7% Caucasian, 13.0% African American; Other 2.3%). The time period of the study will span 1999 to 2006 for $PM_{2.5}$ mass and 1999 to 2003 for speciated $PM_{2.5}$ components which are constrained by the availability of sufficient component data. Air monitoring and meteorological data from the larger 35-county region will be used to help inform exposure estimates for the MSA. Analysis performed in this feasibility study has demonstrated that, with the inclusion of data from archived air monitoring filters, sufficient exposure and health outcomes data exists for the characterization of health effects over the period from 1999 to 2006.

There are four main tasks to be completed:

1. Assembly of an air monitoring/exposure daily database using existing datasets and yet-to-be-analyzed archived filters for speciated $PM_{2.5}$ components (see the blue boxes in **Figure 1, Section 1**),

2. Assembly of a health outcomes daily database (see the yellow boxes in **Figure 1**, **Section 1**,

3. Statistical analysis and modeling to characterize the relationship between various health outcomes and $PM_{2.5}$ mass concentration, components, and emissions from coal-fired power plants while adjusting for confounding factors (see the orange boxes in **Figure 1, Section 1)**, and

4. Writing of interim progress reports and a final scientific report providing a comprehensive description of the methods, results and conclusions.

These tasks are described in more detail below. During the completion of Task 1 and Task 2, it will be necessary to acquire numerous data sets. These data sets will be stored in an interim database for analysis and manipulation using the R statistical programming language before being assembled into a comprehensive daily database suitable for statistical time series modeling and other analyses. PostgreSQL will be used to house the database on a secure server. Because the air monitoring/exposure data were collected using various monitors distributed irregularly in space and time, considerable effort will be required to estimate daily exposure for each ZIP code area. Given the availability of sufficient speciated $PM_{2.5}$ data only from August 1999 to August 2004, statistical models focusing on $PM_{2.5}$ components (while accounting for confounding factors) can only be constructed for this time period. Statistical models

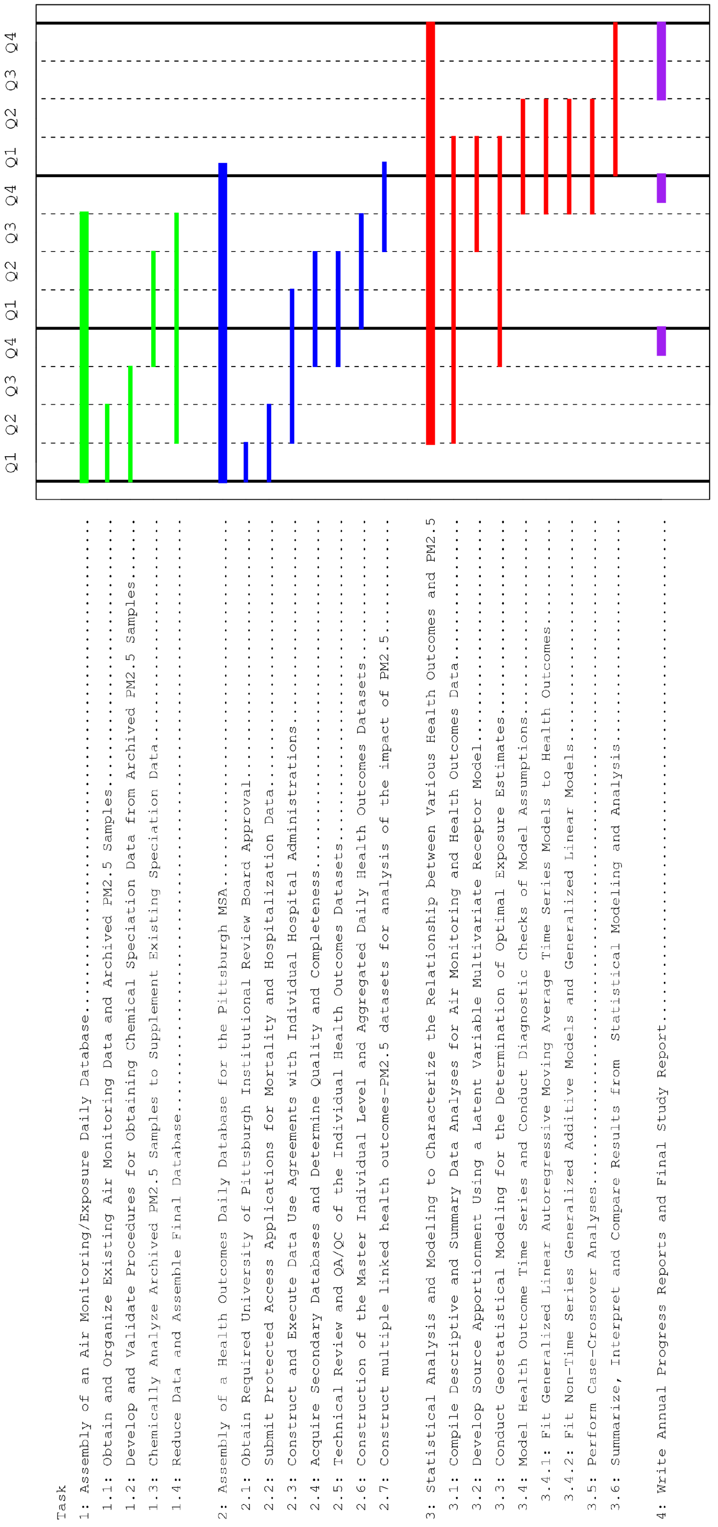


Figure 85: Gantt chart displaying the expected start dates and durations for each study task.

focusing on $PM_{2.5}$ mass concentrations, however, can be constructed for the entire period from 1999 to 2006.

It is expected that the study will take three years to complete. **Figure 85** provides a Gantt chart illustrating the start times and durations of the various tasks comprising the proposed study. Most of the first two years will consist of acquiring the air monitoring/exposure data and health outcomes data, organizing and performing quality assurance, quality control and statistical calibration procedures, and conducting extensive geostatistical modeling and latent variable modeling to help construct daily time series of health outcomes and air monitoring/exposure parameters for ZIP code areas within the study region. Once the daily database is constructed, the last 15 months will be devoted to time series modeling/analysis and other applicable techniques. These methods will be used to characterize the relationships between $PM_{2.5}$ mass, components, latent parameters determined by source apportionment and health outcomes while accounting for confounding effects of gaseous co-pollutants and meteorological factors among others.

Task 1 – Assembly of an Air Monitoring/Exposure Daily Database

The objective of this task is to assemble a database of daily ambient $PM_{2.5}$ mass, $PM_{2.5}$ speciation, co-pollutant, and meteorological data for use in representing the exposures of the population of the Pittsburgh Metropolitan Statistical Area (MSA) to these parameters in a retrospective time series epidemiology study according to the design that was developed. It was concluded that there are a sufficient quantity and quality of existing $PM_{2.5}$ samples and archived $PM_{2.5}$ samples to permit a retrospective epidemiology study of $PM_{2.5}$ from coal-fired power plants and other emission sources in the Pittsburgh region. A four-year study focusing on the Pittsburgh MSA between August 3, 1999, and August 2, 2003 was recommended. To develop exposure estimates for chemical components of $PM_{2.5}$, the study will utilize existing $PM_{2.5}$ speciation data collected by seven monitoring sites (i.e., the Bruceton, Hazelwood, Lawrenceville, and Schenley Park sites in Allegheny County, the Florence site in Washington County, and the Greensburg and St. Vincent College sites in Westmoreland County) that operated in the Pittsburgh MSA during some or all of the four-year study period, as well as additional $PM_{2.5}$ speciation data that will be obtained by chemically analyzing archived $PM_{2.5}$ samples collected by the Bruceton, Florence, Lawrenceville, Liberty Borough, and St. Vincent College sites during that period. Existing $PM_{2.5}$ speciation data that were collected by eight monitoring sites (i.e., Franciscan University of Steubenville, Holbrook, Hopedale, M.K. Goddard, Quaker City, Tomlinson Run State Park, Wheeling Jesuit University, and Youngstown) located in a larger region surrounding the Pittsburgh MSA will also be included in the database for possible use in geostatistical modeling to help inform the exposure estimates developed for the MSA. Estimates of $PM_{2.5}$ total mass concentrations, co-pollutant (i.e., $PM_{10-2.5}$, SO_2, CO, NO_2, and O_3) concentrations, and meteorological conditions in the Pittsburgh MSA will be derived by geostatistically modeling data that were collected during the study period at numerous sites located in a 35-county region centered on Pittsburgh.

Table 24 (Section 2) summarizes the expected availability of $PM_{2.5}$ speciation data, including both preexisting data and data that will be obtained by analyzing archived $PM_{2.5}$ samples, for the four-year study. (The statistics presented in the table do not include $PM_{2.5}$ speciation data from outside of the Pittsburgh MSA). The study will include, at a minimum, the following $PM_{2.5}$ species: sulfate, nitrate, elemental carbon (EC), organic carbon (OC), and 14 trace and crustal elements (As, Br, Ca, Cr, Cu, Fe, K, Mn, Ni, Pb, Se, Si, Ti, Zn). Archived $PM_{2.5}$ samples were collected on either Teflon or quartz filters; all of the archived quartz filters and some of the archived Teflon filters have been stored under refrigeration since collection. The numbers shown in the table were tabulated under the assumptions that all archived Teflon-filter-based samples can be analyzed to provide valid fine particulate sulfate and trace/crustal element data, that refrigerated Teflon-filter-based samples can additionally be analyzed to provide valid fine particulate nitrate data, and that all archived quartz-filter-based samples can be analyzed to provide valid fine particulate sulfate, nitrate, and EC/OC data. As indicated in **Table 24 (Section 2)**, for each $PM_{2.5}$ species, data are expected to be available from at least one monitoring site in the Pittsburgh MSA on greater than 90% of the 1460 study days, with many days (90% for sulfate, 66% for nitrate, 42% for EC/OC, and 88% for trace/crustal elements) having data available from multiple sites. Although not shown in **Table 24 (Section 2),** >90% data availability is also expected for $PM_{2.5}$ total mass, co-pollutants, and meteorological parameters.

In addition to the four-year retrospective study focusing on chemical components of $PM_{2.5}$, performing a longer study focusing on $PM_{2.5}$ total mass and co-pollutants is also proposed. This study would take advantage of the additional statistical power afforded by the Pittsburgh region's vast record of daily data for these criteria pollutants. Hence, rather than focusing only on the August 1999 to August 2003 period, the database of air monitoring data will include all publicly available $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, and O_3 data that were collected in the 35-county greater Pittsburgh region between 1999 and 2006. There were at least 47 sites that measured $PM_{2.5}$, 56 sites that measured PM_{10}, 49 sites that measured SO_2, 20 sites that measured CO, 16 sites that measured NO_2, and 34 sites that measured O_3 in the region during some or all of the 1999-2006 period.

Work on Task 1 will be carried out according to a series of four subtasks, as detailed below.

Task 1.1 – Obtain and Organize Existing Air Monitoring Data and Archived $PM_{2.5}$ Samples

Work to be performed under this task includes:

1. Obtaining and organizing existing $PM_{2.5}$ chemical speciation data that were collected at monitoring sites in the greater Pittsburgh region between August 3, 1999, and August 2, 2003. All associated metadata (e.g., validation flags, sampling start and end times, etc.) will also be gathered. Data will be obtained from the following sources, per the inventory that was conducted as part of the DOE-

sponsored PITT-PM feasibility assessment:

- Speciation Trends Network (STN) – Florence, Greensburg, Hazelwood, Lawrenceville, and Youngstown sites
- National Energy Technology Laboratory, Office of Science and Technology (NETL/OST) – Bruceton site
- Pittsburgh Air Quality Study (PAQS) – Schenley Park site
- Upper Ohio River Valley Project (UORVP) – Lawrenceville and Holbrook sites
- Steubenville Comprehensive Air Monitoring Program (SCAMP) – Franciscan University of Steubenville, Hopedale, Tomlinson Run State Park, St. Vincent College, and Wheeling Jesuit University sites
- Clean Air Status and Trends Network (CASTNet) – M.K. Goddard and Quaker City sites
- Interagency Monitoring of Protected Visual Environments (IMPROVE) Network – M.K. Goddard and Quaker City sites

These data are all publicly available; most have already been obtained by CONSOL as part of the DOE-sponsored feasibility assessment. Under this task, the remaining data will be obtained and all data organized in an electronic database by data source, monitoring site, and parameter. Electronic data files will be kept on a secure server and routinely backed up to prevent catastrophic loss.

2. Obtaining and organizing existing $PM_{2.5}$ mass, co-pollutant (i.e., PM_{10}, SO_2, CO, NO_2, and O_3), and meteorological (i.e., temperature, relative humidity, dew point, wind speed, and wind direction) data that were collected by monitoring sites in the 35-county greater Pittsburgh region between 1999 and 2006. All associated metadata (e.g., validation flags, sampling start and end times, etc.) will also be gathered. Data will be obtained from the following sources, per the inventory that was conducted as part of the DOE-sponsored PITT-PM feasibility assessment:

- U.S. Environmental Protection Agency’s Air Quality System (AQS)
- National Energy Technology Laboratory, Office of Science and Technology (NETL/OST)
- Pittsburgh Air Quality Study (PAQS)
- Upper Ohio River Valley Project (UORVP)
- Steubenville Comprehensive Air Monitoring Program (SCAMP)

- Clean Air Status and Trends Network (CASTNet)
- Interagency Monitoring of Protected Visual Environments (IMPROVE) Network
- Automated Surface Observing System (ASOS) and Automated Weather Observing System (AWOS) Stations
- Roadway Weather Information System (RWIS)

These data are all publicly available; many have already been obtained as part of the DOE-sponsored feasibility assessment. Under this task, the remaining data will be obtained and all data will be organized in an electronic database by data source, monitoring site, and parameter. Electronic data files will be kept on a secure server and routinely backed up to prevent catastrophic loss.

3. Requesting, obtaining, and organizing all archived $PM_{2.5}$ samples (including blanks and duplicates) that were collected at the Bruceton, Lawrenceville, Florence, Liberty Borough, and St. Vincent College monitoring sites between August 3, 1999, and August 2, 2003 (as well as any samples from outside of this date range that are needed to complete Task 1.2.2 – e.g., for the Liberty Borough site). Sampler operating data and QA/QC data regarding the collection of these samples will also be gathered. Samples will be obtained from the Allegheny County Health Department (for the Lawrenceville and Liberty Borough sites), NETL/OST (for the Bruceton site), Pennsylvania DEP (for the Florence site), Desert Research Institute (for the Lawrenceville site), and CONSOL (for the St. Vincent College site). Activities to be performed include:

   - Requesting and securing permission to obtain and analyze the archived $PM_{2.5}$ samples. Discussions in this regard have already been initiated with all of the groups listed above as part of the DOE-sponsored feasibility assessment, and all were preliminarily agreeable to contributing archived samples for use in the study.
   - Physically obtaining the samples, organizing them (i.e., according to monitoring site, filter type, collection date, blank vs. non-blank, etc.), and storing them prior to analysis. Samples will be stored at the CONSOL R&D facility in South Park, PA. Refrigerated storage will be provided for samples that previously were being stored under refrigeration. Standard chain of custody forms will be used to track sample stewardship.

Task 1.2 – Develop and Validate Procedures for Obtaining Chemical Speciation Data from Archived $PM_{2.5}$ Samples

In order to maximize data quality and avoid unnecessary costs, prior to beginning full-scale analysis of archived $PM_{2.5}$ samples, a pilot study will be conducted to confirm the quality of results that can be obtained from the filters obtained in Task 1.1.3. The study will take advantage of the fact that, for all sites except St. Vincent College, there are days from which both existing $PM_{2.5}$ speciation data and an archived $PM_{2.5}$ sample are available. Hence, pairwise comparisons between the existing speciation data and the speciation data obtained by analyzing the archived samples can be used to establish the validity of the archived sample results and to allow any artifacts resulting from the use of these archived samples to be corrected. Specific objectives of the pilot study include resolving uncertainties regarding the quality of trace/crustal element results that can be obtained by analyzing archived $PM_{2.5}$ samples and the feasibility of obtaining nitrate and ammonium data from archived samples that have not been stored under refrigeration. Results of the pilot study will be used to refine the plan for archived filter analysis (including the specific filters to be analyzed and the methods used to analyze them), thereby ensuring that analyses that would not contribute any valuable data to the study are not performed. Subtasks to be performed under Task 1.2 include:

1. Developing QA/QC protocols and standard operating procedures (SOPs) for the determination of inorganic ions (by ion chromatography), elemental and organic carbon (by thermal optical transmittance), and trace and crustal elements (by X-ray fluorescence spectroscopy) from archived $PM_{2.5}$ samples. These will be adapted from existing protocols and SOPs where possible.

2. For each of the Bruceton, Florence, Lawrenceville, Liberty Borough, and St. Vincent College sites, chemically analyzing up to 100 archived $PM_{2.5}$ samples that were collected on days from which collocated $PM_{2.5}$ speciation data are already available. (In the absence of collocated, preexisting speciation data from the monitoring site under consideration, preexisting data from a nearby site will be used for comparison with the archived sample results – e.g., archived sample results from the St. Vincent College site will be compared with preexisting data from the Greensburg site, which was located about 10 km away). Teflon-filter-based samples will be analyzed for trace and crustal elements by X-ray fluorescence spectroscopy and for inorganic ions by ion chromatography, and quartz-filter-based samples will be analyzed for EC and OC by thermal optical transmittance and for inorganic ions by ion chromatography.

3. Applying latent variable modeling and Bland-Altman analyses to develop calibrations relating the archived sample results to the existing speciation data. Random and systematic errors resulting from use of the archived sample results will be rigorously characterized.

4. Finalizing plans for the chemical analysis of archived $PM_{2.5}$ samples, based on the results of Task 1.2.3. As discussed above, the purpose of this task is to provide a means for avoiding unnecessary project costs by ensuring that only analyses that will contribute useful information to the retrospective epidemiology study are performed. Modifications to analytical methods may be explored and implemented if necessary to improve data quality (e.g., by providing better sensitivity for determining trace element species); revisions to the SOPs developed under Subtask 1.2.1 will be made accordingly. The final plan will specify in detail the archived $PM_{2.5}$ samples to be analyzed, the chemical species to be determined from each sample, and the analytical techniques to be used to perform these analyses. A

final budget for archived sample analysis will also be developed.

Task 1.3 – Chemically Analyze Archived $PM_{2.5}$ Samples to Supplement Existing Speciation Data

Work to be performed under this task includes:

1. Chemically analyzing archived $PM_{2.5}$ samples according to the final plan developed under Task 1.2.4.

2. Reducing and quality assuring all data produced by the chemical analysis of archived $PM_{2.5}$ samples in Task 1.3.1. This includes converting laboratory results to ambient air concentration units using the sampler operating data collected as part of Task 1.1.3. QA/QC will be conducted according to the protocols developed under Task 1.2.1.

Task 1.4 – Reduce Data and Assemble Final Database

The ambient air monitoring data that will be used to represent exposures in the retrospective epidemiology study in many cases were collected by different groups who used different sampling and analytical techniques and employed different QA/QC protocols. Hence, a major objective of this task is to provide for consistency among these data before they are used in geostatistical and epidemiological modeling. An air monitoring database for the study will then be assembled. Specific tasks to be performed include:

1. Applying a consistent set of QA/QC standards to the data. Data will be reviewed to identify inconsistencies in QA/QC procedures, and a consistent set of QA/QC criteria will be applied where practical. All data will be qualified using NARSTO standard validation flags. Data will be vetted using descriptive statistics and graphs, and any outliers or anomalies will be investigated and documented.

2. Mathematically adjusting data (e.g., using calibrations such as those developed as part of Task 1.2.3) to account for relative biases resulting from discrepancies in sampling and analytical techniques, blank correction practices, archiving procedures, etc. Calibrations will be performed using latent variable modeling or Bland-Altman analysis, as appropriate. This task will also include, for example, conversion of PM_{10} concentration data, which are commonly reported at standard temperature and pressure, to local conditions so that these data can be used in conjunction with $PM_{2.5}$ data (reported at local conditions) to derive $PM_{10-2.5}$ estimates. All procedures will be thoroughly documented and reported with the results of the study.

3. Aggregating data, as required, to compute daily, midnight-to-midnight average values for each parameter at each monitoring site. This includes:

   - Aggregating measurements that were measured on a finer-than-daily time resolution (e.g., hourly, 3-hour, etc.) to compute 24-hour mean values (or other metrics appropriate for quantifying exposure, such as maximum 1-hour average concentration, maximum 8-hour average concentration, etc.), provided that the data satisfy appropriate completeness criteria (e.g., >75% valid data availability per day).

- For 24-hour integrated data that were not measured from midnight-to-midnight, combining these data (e.g., using time-weighted means) to derive midnight-to-midnight averages. This applies to 24-hour integrated data from the Bruceton monitoring site, which were routinely measured from noon to noon, and to 24-hour integrated data from the SCAMP monitoring sites, which were routinely measured from 9:00 am to 9:00 am. Where available, hourly $PM_{2.5}$ measurements may be used to inform the averaging process.

- All procedures will be thoroughly documented and reported with the results of the study.

4. Assembling the reduced, validated, daily data from all sites into a comprehensive database for use in the study. The database will be formatted so that it can easily be imported into the geostatistical and epidemiological models, and it will be housed on a secure server and routinely backed up to prevent catastrophic loss.

Task 2 – Assembly of a Health Outcomes Daily Database for the Pittsburgh MSA

The overall objective of this task is to assemble a database of daily health outcomes focusing on the Pittsburgh MSA for the period from 1999 to 2006. The retrospective nature of this study will necessitate the use of existing secondary data on mortality, hospitalizations and emergency department (ED) visits within the seven-county region of interest. These data will ultimately be linked to $PM_{2.5}$ (both mass and chemical components) for the same region to examine the relationship between $PM_{2.5}$ and these daily health outcomes. Epidemiological studies and the time series power analysis (**Section 4.1**) have suggested that a minimum of three years of data are needed to acquire the statistical power necessary to identify the health effects of typical $PM_{2.5}$ exposures. The proposed study period for optimal retrospective data coverage for speciated $PM_{2.5}$ is 4 years (August 3, 1999-August 2, 2003). However, health outcomes data will be obtained for the period from 1999-2006 to take advantage in a longer study of the extensive $PM_{2.5}$ mass data available for the Pittsburgh region. It is desirable to have access to various categories of specific health outcomes (hospitalizations, ED visits) as well as total, cardiovascular and respiratory mortality for this region that can then be correlated with $PM_{2.5}$ in a rigorous manner.

It was determined that standardized data for both mortality and hospitalizations are available from 1999-2003 (and through 2006) for the region of interest. It was also demonstrated that an ED visit dataset can be constructed but will likely be more limited in geographic coverage. ED data are available from 40% of the area hospitals from 1999-2004 and the hospitals with the most complete data are associated with the large healthcare systems (UPMC Health System, West-Penn Allegheny Health System, Mercy Health

System). These hospitals are more likely to partner with university and industry-based research groups.

Based on a complete comprehensive inventory and assessment of available mortality and morbidity datasets, the primary identified sources of secondary data include:

1. National Center for Health Statistics, Division of Vital Statistics
2. Pennsylvania Department of Health, Bureau of Health Statistics and Research
3. Allegheny County Health Department
4. PA Health Care Cost Containment Council Hospital Discharge Datasets (l999-2004)
5. Ohio Department of Health Hospital Discharge Datasets (l999-2004)
6. West Virginia Health Care Authority Hospital Discharge Datasets (l999-2004)
7. Emergency Department Visit Data (from hospital systems and individual hospitals)
8. UPMC Medical Archival Retrieval System (MARS)

Unlike the $PM_{2.5}$ data, individual level health outcomes data are not publicly available and must be requested by specific protected access application to the agency or institution responsible for data collection. We intend to also acquire mortality and hospitalization data for counties in West Virginia and Ohio that border the Pittsburgh MSA and are considered a part of the broader Ohio River Valley.

Task 2.1 – Obtain Required Institutional Review Board Approval for Acquisition of Secondary Limited Datasets from the Pennsylvania Department of Health, the Pennsylvania Health Care Cost Containment Council, West Virginia and Ohio Hospital Associations, Hospital Systems and Individual Hospitals

Institutional review board (IRB) approval of a project is mandatory prior to the requesting of data if a study or project involves human subjects and requires the receipt of records with personal identifiers from medical or other agencies. This task will be completed within the first three months of the project period.

Task 2.2 – Submit Protected Access Applications for Mortality and Hospitalization Data to the Pennsylvania Department of Health, PA Health Care Cost Containment Council and Allegheny County Health Department

The process of obtaining confidential data is initiated with the submission of a completed “Application for Access to Protected Data” to the institutions or agencies. Guidelines and procedures for these “follow-back" activities using Pennsylvania records are covered in detail in the Pennsylvania "User's Guide for Access to Protected Data." The application must be reviewed and approved by the Department of Health and/or the PHC4 prior to release of the information. Applications can be obtained by writing to the Director, Division of Health Statistics and Research and the PHC4 Special Requests Unit. Similar procedures are in place in West Virginian and Ohio. This task with be completed within the first 6 months of the project period.

Task 2.3 – Construct and Execute Data Use Agreements with Individual Hospital Administrations for Access to Limited Datasets for Emergency Room Data That Has Been De-identified but Provides ZIP Code of Residence, Age, Race and Gender of Individual, Date of Visit/Admission and Discharge (If Admitted to the Hospital), Both Admission and Discharge Diagnoses and Vital Status Outcome at Discharge

Unlike hospital admission data, emergency department (ED) visit data for population-based epidemiological assessments are not captured by a single centralized agency in Pennsylvania. Therefore ED data will need to be accessible in electronic format from hospital systems or individual hospitals in the Pittsburgh MSA. Data use agreements will be constructed and executed with each hospital system and/or hospital of interest. This task will be completed by the end of the first quarter of the second project year.

Task 2.4 – Acquire Secondary Databases and Determine Quality and Completeness for Daily Total, Respiratory and Cardiovascular Mortality, Hospitalizations and ED Visits

Health outcomes data will be obtained from secondary sources per the inventory conducted as a part of the feasibility assessment. These sources are outlined briefly below.

Daily Mortality Data for the Pittsburgh MSA

Mortality data in the Pittsburgh MSA and the region are relatively well characterized for the 1999-2004 time period of interest through the National Center for Health Statistics (NCHS) Division of Vital Statistics and the Pennsylvania Department of Health. Pennsylvania mortality data is also available directly through the Pennsylvania Department of Health Bureau of Health Statistics and Research. Complete datasets are currently available from 1999-2004. Protected access datasets from the Bureau of Health Statistics and Research include street address and ZIP code of residence (ZIP+4), as well as demographic variables such as age, race, education, marital status and occupation**.** During the proposed 4-year study period, approximately 114,000 all-cause deaths were observed in the Pittsburgh MSA (**Table**

26, Section 3).

Daily Hospitalizations

Hospitalization data are readily accessible from 1999 (and earlier) through 2006 at the geographic level of 5-digit ZIP code from the Pennsylvania Health Care Cost Containment Council (PHC4). Through a protected access data user agreement with PHC4, researchers can obtain customized datasets that contain dates of hospital admission and discharge in addition to age, gender, race, ZIP code of residence and other variables of interest for all subjects. Admission type and admission source codes are also available to determine if the subject was admitted through the emergency department or other external entity in an unscheduled admission. This is an important consideration for time-series studies of air pollution and health outcomes. Individual street addresses, however, are not available.

Table 62 presents a summary of admissions to hospitals located within the Pittsburgh MSA by county of residence and specific hospital from 1999-2004. As observed in the summary table, more than half (53.4%) of the total hospital admissions from 1999-2004 in the 7-county Pittsburgh MSA (2.23 million) involve residents of Allegheny County (1.19 million). Of the 1.19 million Allegheny County residents admitted to hospitals during this time period, 98.7% were admitted to hospitals located within the boundaries of Allegheny County. A total of 78% of all hospital admissions represent residents of the three most populated counties in the Pittsburgh MSA: Allegheny, Washington and Westmoreland. Approximately 99% of the residents of these three counties are admitted to hospitals within the 3-county area. Approximately 27-30% of all hospital admissions in the Pittsburgh MSA were determined to be of circulatory or respiratory origin. A custom dataset of total admissions or admissions for residents of each Pennsylvania county within the Pittsburgh MSA by specific ICD 9/10 code(s) will be obtained in Excel or SPSS format for a minimal cost outlay.

Table 62: Resident admissions to hospitals in the Pittsburgh MSA by county of residence and hospitals, 1999-2004.

Hospitals by County	County of Residence							Total MSA
	Allegheny	**Armstrong**	**Beaver**	**Butler**	**Fayette**	**Washington**	**Westmoreland**	
Allegheny County								
Alle Kiski Medical Center	21363	9513		9553			28156	68585
Allegheny General Hospital	91203	3344	7019	9523	3905	8011	7738	130743
Childrens Hospital of Pittsburgh	32736	1420	2565	3633	2260	3848	5972	52434
Forbes Regional Hospital	58721	578	139	270	343	200	25245	85496
Jefferson Regional Medical Center	78907		53	79	5077	7725	2979	94820

Hospitals by County	**County of Residence**							**Total MSA**
	Allegheny	**Armstrong**	**Beaver**	**Butler**	**Fayette**	**Washington**	**Westmoreland**	
Magee Womens Hospital of UPMC	86205	865	1995	6639	2222	5094	8058	111078
Mercy Hospital of Pittsburgh	94521	425	1540	2256	4756	5977	3974	113449
Mercy Providence Hospital	17024		155	67		354	113	17713
Ohio Valley General Hospital	24401		615	144	76	1218	96	26550
Sewickley Valley Hospital	31046		26750	1138	52	852	146	59984
St. Clair Memorial Hospital	76340		185	167	288	12004	238	89222
Suburban General Hospital	22787		228	359		56		23430
UPMC Braddock	36643	66	313	58	192	181	937	38390
UPMC McKeesport	53178				123	261	2693	56255
UPMC Passavant	42999	79	1435	15384	57	134	275	60363
UPMC Presbyterian/Shadyside	184891	4001	6044	7297	10418	10427	25244	248322
UPMC South Side	31856		108	105	102	279	258	32708
UPMC St. Margaret	52093	1085	303	1876	279	451	6665	62752
Western Pennsylvania Hospital	93647	2054	1639	4354	2033	1992	9039	114758
St. Francis Medical Center (now closed)	43323	478	1137	2228	580	882	3056	51684
St. Francis Central (now closed)	4064		149	331	135	244	96	5019
Allegheny County hospital subtotal	**1177948**	**23908**	**52372**	**65461**	**32898**	**60190**	**130978**	**1543755**
Armstrong County								
Armstrong County Memorial Hosp.	303	27327		3966			1760	33356
Armstrong County hospitals subtotal	**303**	**27327**		**3966**			**1760**	**33356**
Beaver County								
Aliquippa Community Hospital	890		17445			102		18437
Medical Center Beaver	2109		83191	698		186		86184
Beaver County hospitals subtotal	**2999**		**100636**	**698**		**288**		**104621**
Butler County								
Butler Memorial Hospital	1257	2390	528	60807		75	322	65379
UPMC Passavant Cranberry (St. Francis)	782		397	2406				3585
Butler County hospitals subtotal	**2039**	**2390**	**925**	**63213**		**75**	**322**	**68964**

Hospitals by County	County of Residence							Total MSA
	Allegheny	Armstrong	Beaver	Butler	Fayette	Washington	Westmoreland	
Fayette County								
Brownsville General Hospital					9484	2326	84	11894
Highlands Hospital					14451	88	832	15371
Uniontown Hospital	102				57498	645	373	58618
Fayette County hospitals subtotal	**102**				**81433**	**3059**	**1289**	**85883**
Washington County								
Canonsburg General Hospital	1845				128	20550		22523
Monongahela Valley Hospital	1548				15607	38580	10899	66634
Washington Hospital	1674		99	58	1900	75260	292	79283
Washington County hospitals subtotal	**5067**		**99**	**58**	**17635**	**134390**	**11191**	**168440**
Westmoreland County								
Frick Hospital	95				13030	153	19767	33045
Latrobe Area Hospital	270	187			969	89	56385	57900
Mercy Jeannette Hospital	1210				795	122	33610	35737
Monsour Medical Center	736	53			789	226	9051	10855
Westmoreland Regional Hospital	1031	76			5085	237	75446	81875
Citizens General (now closed)	1492	763		103			6331	8689
Westmoreland County hospitals subtotal	**4834**	**1079**		**103**	**20668**	**827**	**200590**	**228101**
County of residence subtotal	**1193292**	**54704**	**154032**	**133499**	**152634**	**198829**	**346130**	**2233120**

Daily Emergency Department Visits

Unlike hospital admission data, clinical emergency department (ED) visit data for population-based epidemiological assessments are not captured by a single centralized agency in Pennsylvania. Therefore ED data will need to be accessible in electronic format from hospital systems or individual hospitals in the Pittsburgh MSA. If records are not electronically formatted, manual data abstraction of medical charts would be required to assemble the necessary variables, a process that is costly, time-consuming, and most likely prohibited under the Health Insurance Portability and Accountability Act (HIPPA) of 1996. **Table 45 (Section 3)** presents a summary of the availability of electronic records for ED visits at 36 hospitals in the Pittsburgh MSA for the 1999-2004 time-period. Allegheny County has the largest number of hospitals (20), followed by Westmoreland (4), Washington and Fayette (3 each) and Beaver and Butler (2 each). In Allegheny County, most hospitals are components of one of four major health systems: University of Pittsburgh Health Care System, West-Penn Allegheny Health System, Pittsburgh Mercy Health System, and the Pittsburgh VA Health System. These systems are more likely to have ED data in electronic format than smaller, independent hospitals located in the other counties in the Pittsburgh MSA. However, many smaller hospitals have recently merged with one of the 4-primary systems, allowing the information to be more readily accessed.

Task 2.5 – Technical Review and QA/QC of the Individual Health Outcomes Datasets

Mortality data will be obtained from the Pennsylvania Department of Health Bureau of Health Statistics and Research and verified using National Center for Health Statistics (NCHS) Division of Vital Statistics. Recent quality analysis comparing these electronic datasets to death certificates suggests that the error rate is 2% or less. Hospitalization data is collected by the Pennsylvania Health Care Cost Containment Council (PHC4). The data are processed using a series of validation rules before being finalized and made available for further analysis and public release. PHC4 edits the data and provides error reports for each data source. The health care facility will make error corrections and provide PHC4 with corrected information. Compliance across health care institutions in Pennsylvania approaches 100% (99% in recently released 2006 reports). Emergency department (ED) data will be acquired from individual hospitals/hospital systems through directed agreements. If necessary, the investigators will utilize an “honest broker” system to acquire identified ED data from hospitals for use in the study. Verification of the accuracy and integrity of the ED and other data will be conducted by the data research associate and will include ID verification, data range, type verification, and duplicate entry checks. Additional data editing and report generation will be performed to assure data integrity and completeness.

Task 2.6 – Construction of the Master Individual Level and Aggregated Daily Health Outcomes Datasets

All data will be organized into electronic data files (PostgreSQL database). Due to the confidential nature of some of the individual data elements, the files will be maintained on secure computer servers. If necessary, we will use an “honest broker” to de-identify the files such that any individual level information (name, street address, SSN, etc.) will be removed and replaced by a study ID code before access by study investigators. The resulting dataset is defined by HIPPA as a “limited dataset.” The limited data set can include the following (potentially identifying) information:

- Admission, discharge, and service dates;

- Dates of birth and, if applicable, death;
- Age (including age 90 or over); and
- Five-digit ZIP code or
- Any other geographic subdivision, such as state, county, city, precinct and their equivalent geocodes (except street address).

Separate databases for individual level health outcomes and aggregated daily outcome data will be created. Each health outcome of interest (deaths, hospitalizations, ED visits) will be represented by two distinct datasets. One dataset will maintain individual level data listed by date of admission or visit, and the other will contain aggregated daily counts of each health outcome by date of death, admission or ED visit as appropriate by ICD-9/10 diagnostic codes for primary and secondary diagnoses.

Task 2.7 – Construct Multiple Linked Health Outcomes-$PM_{2.5}$ Datasets for Analysis of the Impact of $PM_{2.5}$ and Its Speciated Components on Mortality and Morbidity Health Endpoints

Final formatting and preparation of health outcomes dataset for linkage to exposure datasets by date of death, admission, or ED visit will be completed by the end of the second project year. The health outcomes datasets (individual record level and aggregated) will be linked to exposure data in a manner consistent with the stated objectives of conducting both times series and case crossover analyses.

Task 3 – Statistical Analysis and Modeling to Characterize the Relationship between Various Health Outcomes and $PM_{2.5}$

The main goal of the PITT-PM study is to characterize the relationship between human health outcomes and ambient airborne fine particles ($PM_{2.5}$) from coal-fired power plants and other emission sources while accounting for the effects of other confounding factors (such as gaseous co-pollutants and meteorology, among others). Once all of the air monitoring/exposure data are quality assured, calibrated where necessary, and organized in a raw database, some further processing will be required. In assembling the daily air monitoring/exposure daily database in Task 1, a latent variable multivariate source apportionment model will be constructed based on the $PM_{2.5}$ speciation data and used to characterize the daily contributions from coal-fired power plants and include this information in the database. It then will be necessary to combine all the spatial and temporal data available for a given parameter and optimally estimate a daily value (using geostatistical/spatial statistical methods) for each day of the study time period and for each ZIP code area. A time series model (for each health outcome of interest) will be constructed using this information, along with information on numerous potentially confounding factors such as gaseous co-pollutants, meteorologic factors, day of week effect, and ZIP code (represented by indicator variables or as a random effect). The method of generalized linear autoregressive moving average (GLARMA) time series analysis will be the primary approach used to characterize the relationship. In addition, the results from the GLARMA analysis will be compared with other non-time series methods that have been typically used such as generalized linear models (GAMs), generalized additive models (GLMs), and a case-crossover analysis. Each of these methods have their strengths and weaknesses and similar outcomes from all the methods would lend credence to the results.

Task 3.1 – Compile Descriptive and Summary Data Analyses for Health Outcomes and Exposure Monitoring

Before more advanced statistical analyses are attempted, it will be necessary to describe and summarize appropriately the data from the health outcomes and exposure monitoring networks.

A series of descriptive analyses will be conducted and tables will be generated to consider both the quantity of cardiopulmonary admission and control disease data by age and gender as well as the distribution by time (month, year, day of week) and specific diagnosis.   Specific descriptive analyses will include, but not be limited to, the following for circulatory diseases, respiratory diseases, and the control disease, respectively:

1. Number and distribution of hospital admissions by year, month, and day of week,
2. Average annual admission numbers and rates by age and gender,
3. Descriptive statistics (mean, SD, median, minimum, maximum) of daily hospital admissions and graphs of daily admissions over time, and
4. Correlation between hospitals for daily admissions.

In addition, these analyses will be performed with data stratified by whether or not the admission was emergent (admitted through the Emergency Department) and by relevant disease categories (e.g., asthma, acute myocardial infarction.)

Naïve spatial maps will be made for the exposure monitoring parameters. Monthly summary statistics will be compiled and daily time series will be plotted for both the air monitoring parameters and the health outcomes. These summaries will be useful for ensuring data quality and will provide a foundation for the more complex statistical modeling efforts.

Task 3.2 – Develop Source Apportionment Using a Latent Variable Multivariate Receptor Model

The monitoring networks collect $PM_{2.5}$ particles without directly identifiable sources. Only limited source profile information is available. A *latent variable* multivariate receptor *model* (LVM) will be used to model the correlation among the observed $PM_{2.5}$ components. The parameters of the LVM characterize the distributions of the latent sources and the relationship among the sources and their relationships to the measured components. The latent sources explain the observed component correlations. The observed correlation matrix of the $PM_{2.5}$ components will be modeled as a function of these unknown parameters. Enough source profile information will be used to construct an identifiable and interpretable LVM. In addition, all known constraints on the parameters will be included (for example, that all source contributions be positive and that the sum of the percentages of all components for a particular source be less than or equal to 100%). Finally, to account for the autocorrelation in the observed exposure time series measurements, nested block bootstrap sampling will be used to construct robust confidence interval estimates for each unknown parameter.

Task 3.3 – Conduct Geostatistical Modeling for the Determination of Optimal Exposure Estimates

The monitoring networks collect information for each exposure parameter at discrete points irregularly distributed spatially and somewhat more regularly distributed temporally. The area actually sampled (as is typical with all forms of sampling) is a tiny fraction of the total region. Optimal exposure estimates will be constructed for each parameter for each day and each ZIP code area. The geostatistical (i.e., spatial statistical) methodology of kriging will be used. This method produces an optimally weighted average of measurements available spatially and temporally. The appropriate weights will be estimated from the space-time correlation structure derived from the variographic analysis of the observed measurements. When estimating a particular ZIP code area for a given day, typically measurements closest to the area spatially and closest to the particular day are given more weight. For example, any measurement inside the ZIP code area on the particular day being estimated would tend to be given a higher weight than measurements outside the area or on preceding or following days. (The exact weighting is also tempered by the clustering of measurements in that any measurements that are close together tend to have smaller individual weights given that they tend to have less statistically independent information.) The optimal weighting depends on the observed space-time correlation structure (which must be estimated and modeled). For each exposure parameter (including $PM_{2.5}$ components and source contributions) a space-time variogram will be estimated and modeled. The resulting models will be used to produce optimal daily estimates for each exposure parameter.

Task 3.4 – Model Health Outcome Time Series and Conduct Diagnostic Checks of Model Assumptions

The final daily database will contain all the air monitoring/exposure estimates along with all the health outcomes for each day in the study period and each ZIP code area. Each health outcome time series will be modeled as a function of the air monitoring/exposure estimates, trend, seasonality, day-of-week, and ZIP code area (represented either as a random effect or a fixed effect via indicator variables). Three types of models will be used: 1) generalized autoregressive moving average (GLARMA) time series models, 2) generalized additive non-time series models (GAMs) and generalized linear non-time series models (GLMs), and 3) case-crossover analysis.

Task 3.4.1 – Fit Generalized Linear Autoregressive Moving Average Time Series Models to Health Outcomes

If the outcomes were continuous and Normally distributed, various time series models such as autoregressive integrated moving average (ARIMA) models could be potentially useful. This type of model in certain cases can successfully handle the autocorrelation in the time series response. The importance of accounting for time series autocorrelation in the response variable cannot be overstated as not taking account of the autocorrelation typically produces standard errors for parameter estimates that are substantially biased downward which produces confidence intervals that are too narrow and overstates statistical significance. (Extensions to this methodology include handling long memory dependence via fractionally integrated models, and heterogeneous variances through autoregressive conditionally heterogeneous (ARCH) and generalized autoregressive conditionally heterogeneous (GARCH) models.) If the health outcomes (which are essentially counts) are high enough, a Normal approximation to the Poisson distribution could be reasonable and the use of ARIMA methods could be feasible. The generalized autoregressive moving average (GLARMA) model (a new extension to time series models for Normal continuous responses) can directly handle discrete Poisson count responses. Poisson GLARMA models will be constructed for each health outcome and used to characterize the relationship between the health outcome and $PM_{2.5}$ while adjusting for confounding exposure factors.

For example, a model for daily hospital admissions cardiovascular disease would include $PM_{2.5}$ components, gaseous co-pollutants, meteorological factors (such as temperature and relative humidity), allowance for trend and seasonality and day-of-week effects, and ZIP code area. The primary interest are the estimated GLARMA parameters for $PM_{2.5}$ components that describe the relationship with hospital admissions. The other factors are included to remove confounding effects given that they also influence admissions. The inclusion of ZIP code area either as an indicator variable for a fixed effect or otherwise as a random effect would help assess the spatial heterogeneity and account for differences in exposure over the region. A large number of models will be constructed by changing the particular health outcome and by looking at different sets of $PM_{2.5}$ components, $PM_{2.5}$ mass instead of components, or instead latent $PM_{2.5}$ factors from the LVM source apportionment.

Standard diagnostic analyses of model residuals will be employed to assess how well the models fit the observed data. The inclusion of many ZIP code areas will help account for and allow the examination of the degree of spatial heterogeneity in the health outcomes. To account for the expected spatial correlation among the ZIP code areas, spatial bootstrapping will also be used to provide more realistic standard errors and confidence intervals for each estimated model parameter. It should also be noted that less information is available for estimating separate ZIP codes when dealing with $PM_{2.5}$ components compared to $PM_{2.5}$ mass due to the more limited speciation data.

Task 3.4.2 – Fit Non-Time Series Generalized Additive Models and Generalized Linear Models

Many air pollution studies have relied on the use of generalized additive models (GAMs) and generalized linear models (GLMs) for modeling health outcomes. The apparent advantage of GAMs over GLMs has to do with the nonparametric handling of meteorological factors in GAMs. These models, however, do not necessarily or typically account for the autocorrelation in the health outcome response. Even when many plausible covariates are included in the model, the model residuals can still be substantially autocorrelated thus throwing into doubt the reasonableness of the estimated standard errors and confidence intervals for the parameters. GAMs and/or GLMs will also be constructed and compared to the GLARMA models. Standard diagnostic analyses of model residuals will be employed to assess how well the models fit the observed data. It should be possible to use nested block bootstrap sampling to produce more realistic standard error estimates and confidence intervals for the GAM/GLM models.

Task 3.5 – Perform Case-Crossover Analyses

Time series models have been successfully applied in air pollution studies (and in many other areas) so their general usefulness cannot be questioned. It can, however, be difficult to construct appropriate models if only because of the necessity of handling trends and seasonality. Also, time series modeling may not handle factors dealing with individual differences (although including ZIP code area as a factor in the time series analysis should help counteract this limitation). In the case-crossover design, only cases are involved and the exposure of each case during an at-risk “hazard period” just before the event is compared with the exposure levels during one or more reference periods when the event did not occur. Cases serve as their own controls. Therefore, time-invariant confounding factors, such as individual characteristics (e.g., age, gender, race, socioeconomic status, etc.) are controlled by design rather than by statistical adjustment. In comparison with time series analysis, the case-crossover approach has a few significant strengths. First, it avoids complex mathematical modeling and adjusting for seasonality because this approach controls some confounding factors such as long-term trend, seasonality and day of week by design rather than by modeling. Moreover, personal characteristics and other time-invariant variables are also controlled by the design. However, the drawback of the case crossover design is that the efficiency of case-crossover design estimators has been shown to be lower than that of time series analysis A case-crossover analysis will be performed for each health outcome and compared to the results from the GLARMA modeling and the GAM/GLM results

Task 3.6 – Summarize, Interpret and Compare Results from Statistical Modeling and Analysis

The results of the statistical modeling efforts will be summarized and the results will be interpreted in terms of the implications for human health. The degree of concordance of the results from the various models will determine the credibility of the evidence for or against health effects.

Task 4 – Write Annual Progress Reports and Final Study Report

Technical progress reports will be produced at the end of the first and second project years. A final comprehensive scientific report will be written that summarizes, describes and explains the results and conclusions of the study.

Estimated Budget

The budget shown in **Table 63** assumes a three year time period to complete the study with a principal investigator, three co-investigators, two graduate student researchers, data manager, project director, and resource manager. The total cost (both direct and indirect) for the study is estimated to be about $1.96 million.

Table 63: Estimated budget for PITT-PM Study by year and task.Includes both direct and indirect costs.

Task	Year 1	Year 2	Year 3	Total
1 Total	***$314,791***	***$296,308***		***$611,099***
1.1	$29,166			$29,166
1.2	$118,672			$118,672
1.3	$129,356	$258,711		$388,067
1.4	$37,597	$37,597		$75,194
2 Total	***$317,689***	***$332,626***	***$21,217***	***$671,533***
2.1	$79,458			$79,458
2.2	$102,570			$102,570
2.3	$79,880	$26,627		$106,507
2.4	$27,890	$55,781		$83,671
2.5	$27,890	$55,781		$83,671
2.6		$109,568		$109,568
2.7		$84,870	$21,217	$106,087
3 Total	***$147,072***	***$194,754***	***$211,522***	***$553,348***
3.1	$90,304	$47,072	$16,768	$154,145
3.2		$33,536	$16,768	$50,304
3.3	$56,768	$47,072	$16,768	$120,609
3.4		$16,768	$28,536	$45,304
3.4.1		$16,768	$28,536	$45,304
3.4.2		$16,768	$28,536	$45,304
3.5		$16,768	$28,536	$45,304
3.6			$47,072	$47,072
4 Total	***$26,519***	***$26,519***	***$70,716***	***$123,754***
Total	**$806,071**	**$850,207**	**$303,456**	**$1,959,733**

APPENDIX A – Checklist for Air Monitoring Data Inventories

Air Monitoring Data

- $PM_{2.5}$ Total Mass
  - Location(s) where Data were Collected
  - Inventory of Data at Each Location
    - Period of collection – start date, end date
    - Frequency of collection (e.g., daily, 1-in-3 days, etc.)
    - Time resolution of collection (e.g., continuous sampling, 24-hr integrated sampling, etc.)
    - Identify periods/days of missing or invalid data
    - Method of collection (e.g., FRM sampler, 30ºC TEOM, 50ºC TEOM, etc.)
    - QA/QC protocol used to validate results
  - Availability of results for use in an epidemiology study (e.g., are data publicly available?)
    - Preferably, obtain data now for use in development of statistical methods, exploratory data analysis.
- $PM_{2.5}$ Speciation
  - Location(s) where Data were Collected
  - Inventory of Data at Each Location
    - Which species were determined? (e.g., sulfate, nitrate, ammonium, elemental carbon, organic carbon, trace elements – which ones?, etc.)
    - Period of collection for each species – start date, end date
    - Frequency of collection for each species (e.g., daily, 1-in-3 days, etc.)
    - Time resolution of collection for each species (e.g., continuous sampling, 24-hr integrated sampling, etc.)

- Identify periods/days of missing or invalid data for each species
- Sampling Method(s)
    - Type of sampler (e.g., Speciation sampler, FRM sampler, Hi-Vol sampler, Continuous carbon analyzer, Continuous sulfate/nitrate analyzer, etc.) – make, model, type of inlet, presence of denuder, etc.
    - Type of filters (e.g., Teflon, quartz, nylon, polycarbonate, etc.)
- Analytical Method(s)
    - Ions (e.g., ion chromatography, continuous sampler, etc.)
    - Carbon (e.g., TOT, TOR, continuous sampler, etc.)
    - Trace Elements (e.g., XRF, low-res ICP-MS, high-resolution ICP-MS, DRC ICP-MS, PIXE, INAA, etc.)
- QA/QC protocol used to validate results
    - e.g., data flagging protocol, was blank subtraction performed during data reduction?

- Availability of results for use in an epidemiology study (e.g., are data publicly available?)
    - Preferably, obtain data now for use in development of statistical methods, exploratory data analysis.

- Gaseous Pollutants
    - Location(s) where Data were Collected
    - Inventory of Data at Each Location
        - Which gases were measured (e.g., CO, SO_2, $NO/NO_2/NO_x$, O_3, etc.)
        - Period of data collection for each gas – start date, end date
        - Frequency of data collection for each gas (e.g., continuous samplers operating every day?)
        - Time resolution of data collection for each gas (e.g., hourly averages, daily averages, etc.)

        - Identify periods/days of missing or invalid data for each gas
        - Sampling methods (i.e., instrument make, model)
        - QA/QC protocol used to validate results
            - e.g., were continuous monitoring data corrected by interpolation between daily calibrations?
    - Availability of results for use in an epidemiology study (e.g., are data publicly available?)
        - Preferably, obtain data now for use in development of statistical methods, exploratory data analysis.
- Weather Data
    - Location(s) where Data were Collected
    - Inventory of Data at Each Location
        - Which variables were measured (e.g., wind speed, wind direction, temperature, barometric pressure, relative humidity, etc.)
        - Period of data collection for each variable – start date, end date
        - Frequency of data collection for each variable (e.g., continuous samplers operating every day?)
        - Time resolution of data collection for each variable (e.g., 15-minute averages, hourly averages, daily averages, hourly max/min, daily max/min)
        - Identify periods/days of missing or invalid data
        - Method of collection (e.g., 10-m tower, etc.)
        - QA/QC protocol used to validate results
    - Availability of results for use in an epidemiology study (e.g., are data publicly available?)
        - Preferably, obtain data now for use in development of statistical methods, exploratory data analysis.
- Other Data of Potential Interest
    - Examples: PM_{10} mass, speciated PM_{10} mass, $PM_{2.5}$ number concentration, etc.

  - Obtain general information regarding quantity, quality, availability.

Archived $PM_{2.5}$ Filters

- Obtain an inventory of all archived $PM_{2.5}$ filter-based samples, including:
  - Number of archived filter-based samples
  - Location(s) where samples were collected
  - Dates on which samples were collected at each location
  - Type of sampler used to collect samples at each location (e.g., Speciation sampler, FRM sampler, Hi-Vol sampler, etc.) – make, model, type of inlet, presence of denuder, etc.
  - Type(s) of filter(s) on which samples were collected (e.g., Teflon, quartz, nylon, polycarbonate, etc.)
  - Way in which filters are being stored (e.g., Are they in well-sealed containers? Are they being refrigerated? Are they being stored in dark or well-lit areas?, etc.)
- Would the filters be available for use in a retrospective epidemiology study if they are needed?
  - Would they be available for analysis using non-destructive techniques? (e.g., XRF, etc.)
  - Would they be available for analysis using destructive techniques? (e.g., IC, ICP-MS, etc.)

APPENDIX B – Database Design for Air Monitoring Data Inventory Results

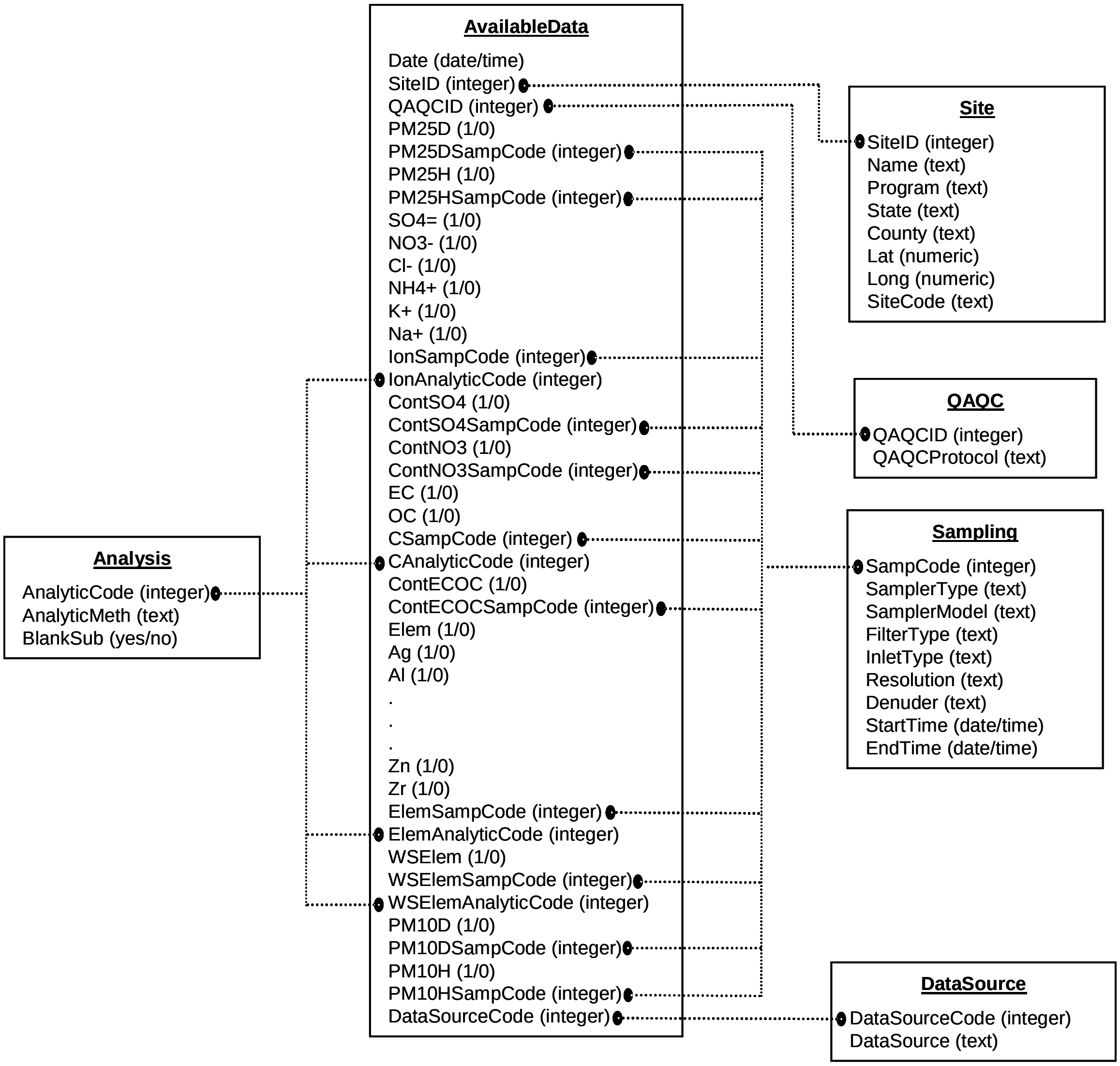


Figure 86: Database design for air monitoring data inventory results.

APPENDIX C – Available Filters

Timelines for archived filter availability are shown in Figures 87-90.

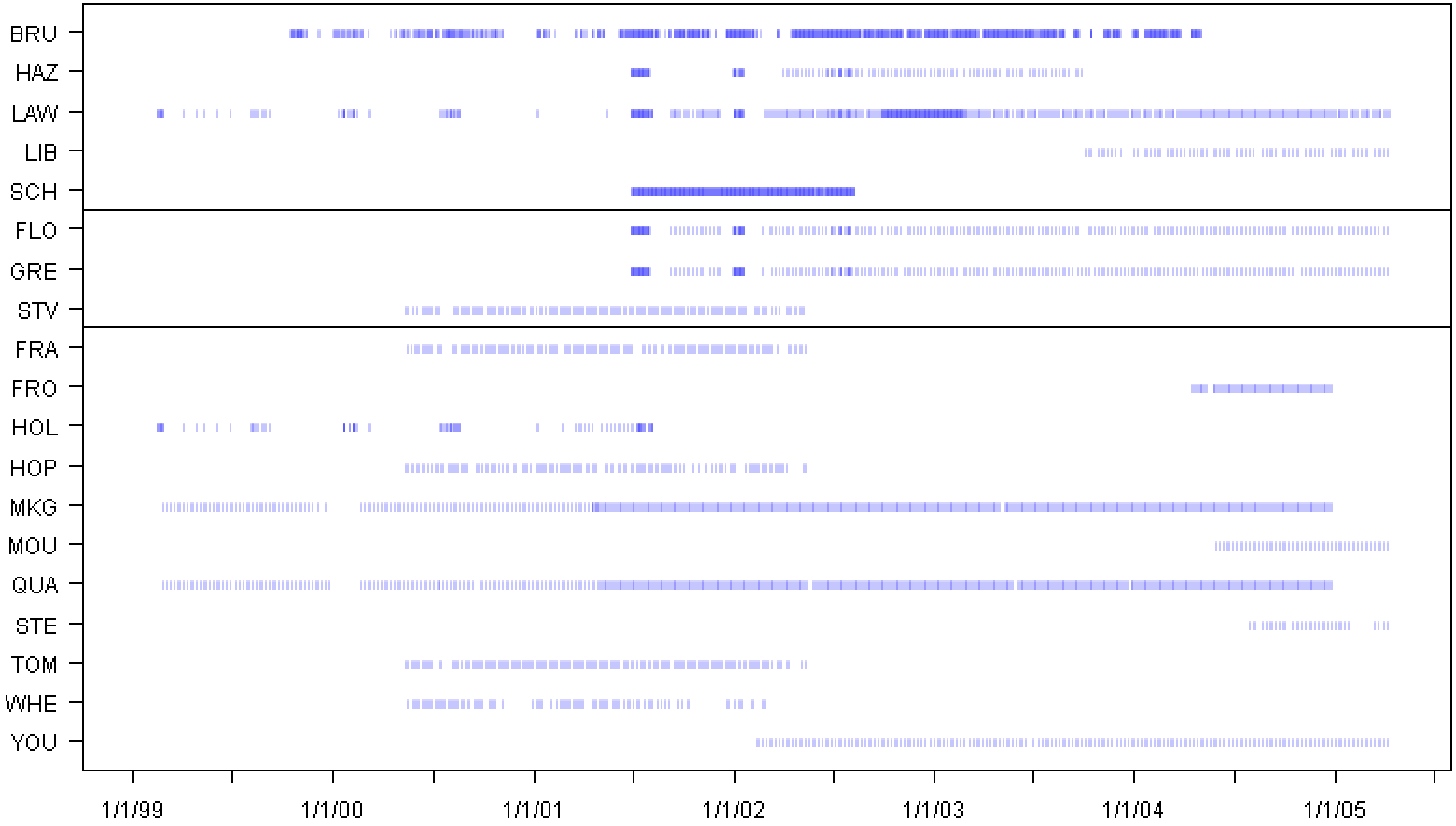


Figure 87: Timeline showing the days for which sulfate data are available from the sites in the 35-county greater Pittsburgh region that monitored for $PM_{2.5}$ speciation between 1999 and 2005. Sites in the top portion of the plot are located in Allegheny County; sites in the middle portion are located in the Pittsburgh MSA, and sites in the lower portion are located outside of the Pittsburgh MSA.

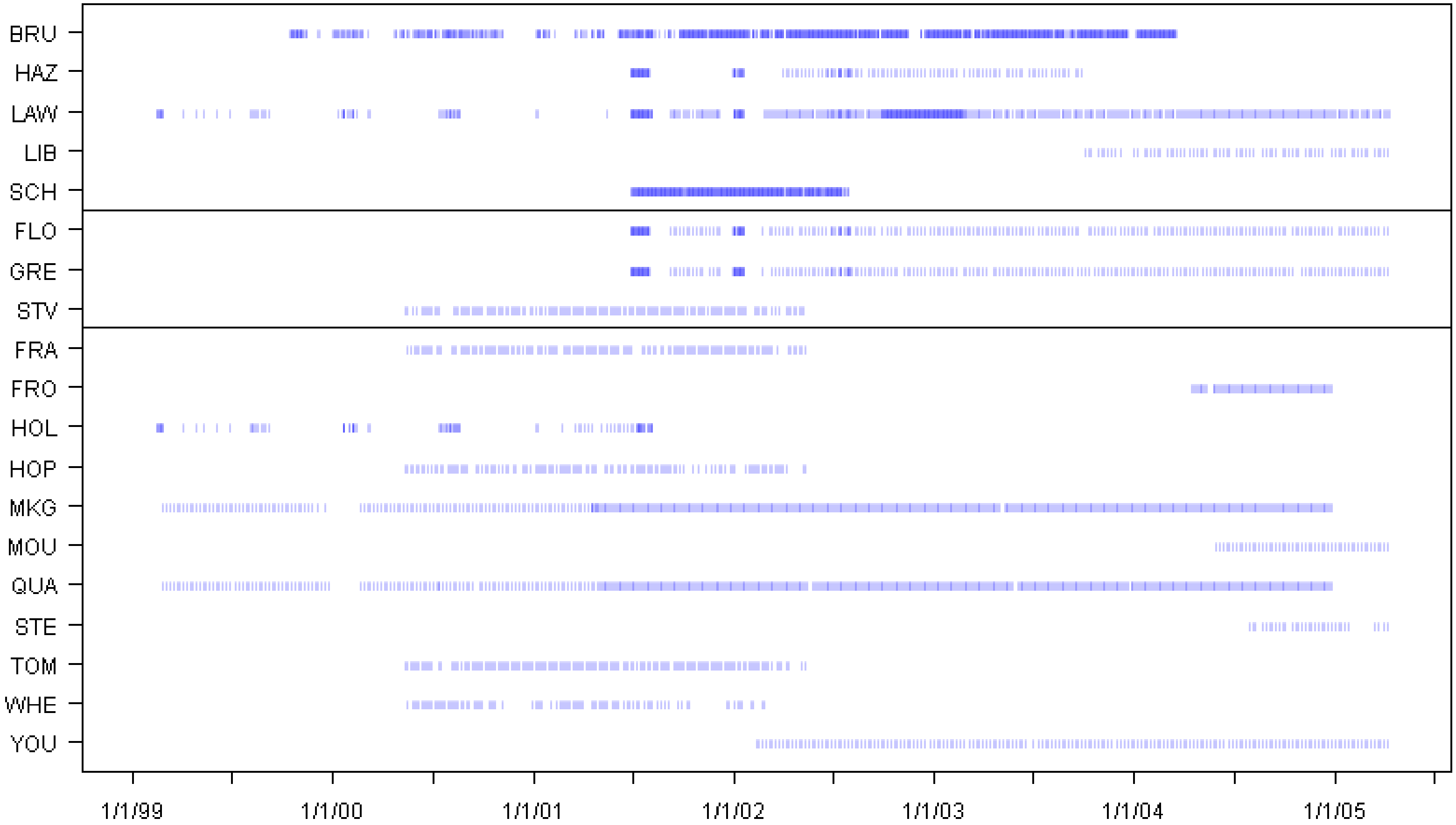


Figure 88: Timeline showing the days for which nitrate data are available from the sites in the 35-county greater Pittsburgh region that monitored for $PM_{2.5}$ speciation between 1999 and 2005. Sites in the top portion of the plot are located in Allegheny County; sites in the middle portion are located in the Pittsburgh MSA, and sites in the lower portion are located outside of the Pittsburgh MSA.

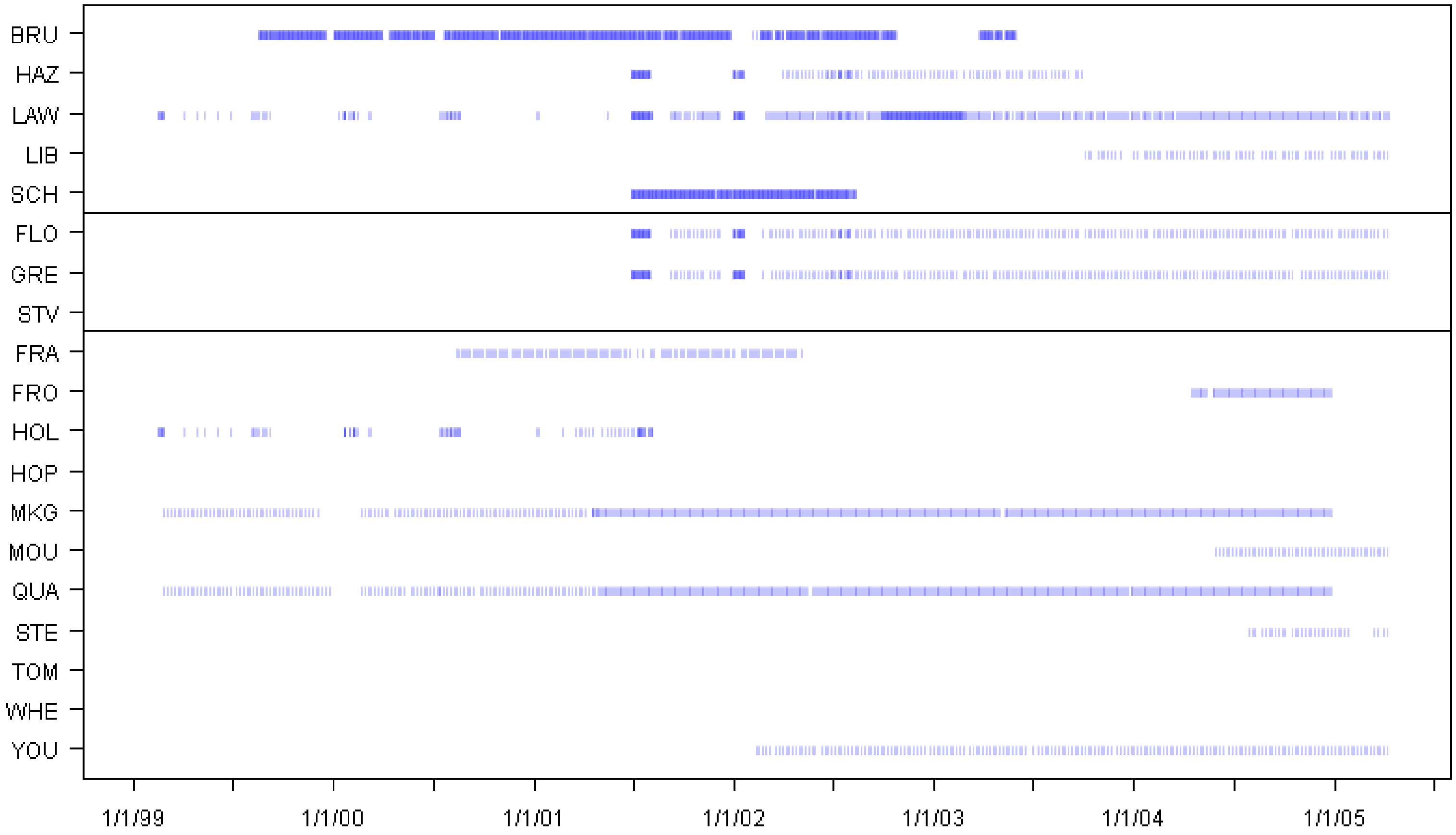


Figure 89: Timeline showing the days for which elemental and organic carbon data are available from the sites in the 35-county greater Pittsburgh region that monitored for PM$_{2.5}$ speciation between 1999 and 2005. Sites in the top portion of the plot are located in Allegheny County; sites in the middle portion are located in the Pittsburgh MSA, and sites in the lower portion are located outside of the Pittsburgh MSA.

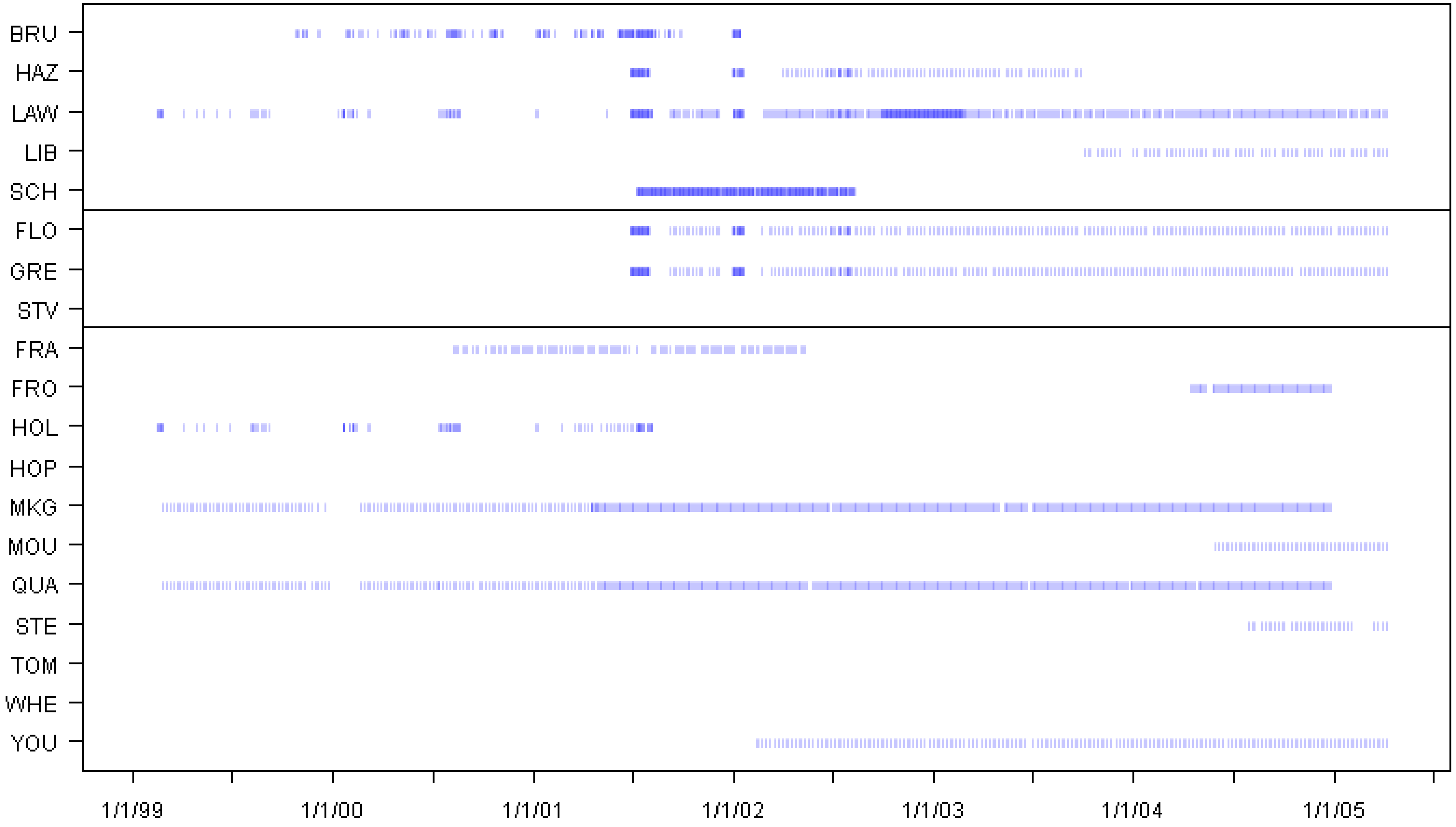


Figure 90: Timeline showing the days for which trace and crustal element data are available from the sites in the 35-county greater Pittsburgh region that monitored for $PM_{2.5}$ speciation between 1999 and 2005. Sites in the top portion of the plot are located in Allegheny County; sites in the middle portion are located in the Pittsburgh MSA, and sites in the lower portion are located outside of the Pittsburgh MSA.

APPENDIX D – Database Design for Archived $PM_{2.5}$ Sample Inventory Results

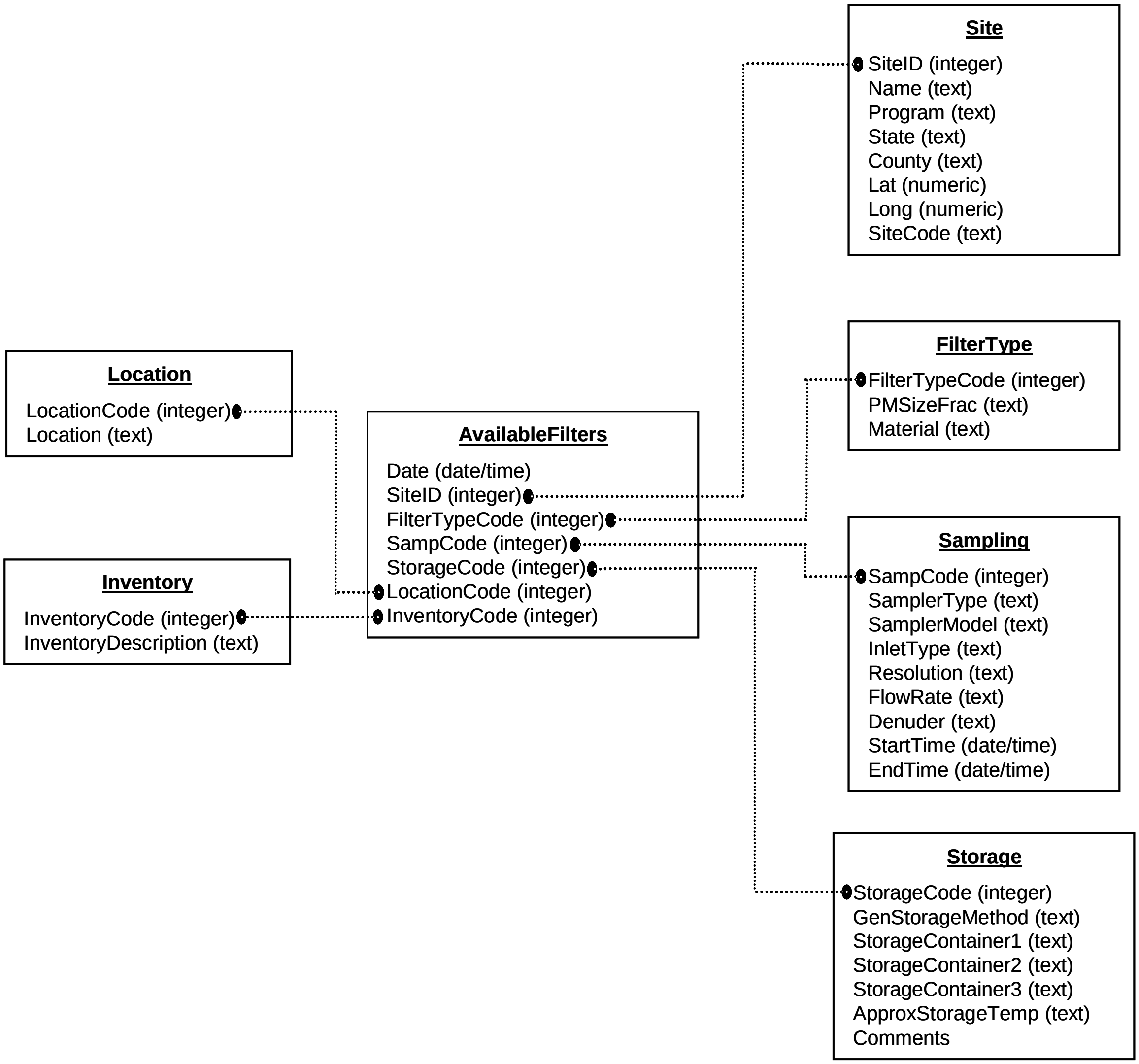


Figure 91: Flowchart for database design for archived $PM_{2.5}$ sample inventory results.

APPENDIX E – Checklist for Health Outcomes Data Inventories

- Deaths, Hospital Admissions, Emergency Room Visits, Physician Visits, etc.
    - Location(s) where Data were Collected
    - Inventory of Data at Each Location
    - Mortality
        - Period of collection (start date, end date)
            - Identify periods/days of missing or invalid data
            - Time resolution of data collection
            - Identify Data format (electronic, hard copy, microfiche, etc)
            - Death certificate availability
            - Date of death
            - Cause of death
            - Secondary causes
            - Street Address of residence available
            - 9-digit or 5-digit Zip Codes of residence available
            - Other data recorded in database
        - QA/QC protocol used to validate results
    - Morbidity
        - Period of collection – start date, end date
            - Identify periods/days of missing or invalid data
            - Time resolution of data collection
            - Identify Data format (electronic, hard copy, microfiche, etc)
            - Determine if aggregate data or individual records available

      - Unique subject identifier recorded
      - Age, gender, ethnicity recorded
      - Street address of residence available
      - 9-digit or 5-digit Zip Codes of residence available
      - Chief complaint
      - ICD 9/10 Codes recorded (primary and secondary)
      - Dates of admission/discharge
      - Other data recorded in database
    - QA/QC protocol used to validate results
    - Availability of data for use in an epidemiology study
      - Are data publicly available?
      - Or is there a cost to obtain the data?
      - Procedure to obtain the data
      - Preferably, obtain data now for use in development of statistical methods, exploratory data analysis

- Pharmaceutical sales
  - Location(s) where Data were Collected
  - Inventory of Data at Each Location
    - Period of collection (start date, end date)
    - Identify periods/days of missing or invalid data
    - Time resolution of data collection
    - Determine if aggregate data or individual records are available
    - Date of sale
    - Medication name provided or class of drug, etc.

        - Street Address of residence available
        - 9-digit or 5-digit Zip Codes of residence available
        - Chief complaint
    - QA/QC protocol used to validate results
    - Availability of data for use in an epidemiology study
        - Are data publicly available?
        - Or is there a cost to obtain the data?
        - Procedure to obtain the data
        - Preferably, obtain data now for use in development of statistical methods, exploratory data analysis
- Cohort Analysis
    - Location(s) and size of available cohorts for longer term retrospective analyses
    - Inventory of data available within each cohort
        - Date of cohort initiation (ongoing?)
        - Residential history (e. g municipalities, street addresses, 9-digit or 5-digit Zip Codes, etc.)
        - Demographics (e.g. age, gender, ethnicity)
        - Medical history
        - Clinical results (laboratory tests, EKGs, heart scans, pulmonary function, etc)
        - Lifestyle factors (e.g. smoking, etc.)
    - Availability of data for use in an epidemiology study
        - Are data publicly available?
        - Or is there a cost to obtain the data
        - Procedure to obtain the data

APPENDIX F – Metadata Matrix for Health Outcomes Datasets

Below is Table 64 which provides information on relevant elements of the important health outcome datasets to be used in the proposed study. Included in the table are the following – time period, availability of electronic data, cost to obtain the data, geographic resolution, and selected variables.

Table 64: Metadata matrix for health outcomes datasets. Source of information: Survey by Allegheny County Health Department and University of Pittsburgh Graduate School of Public Health.

DATA TYPE AND SOURCE(S)	TIME PERIOD AVAILABLE/Data Format (E=electronic; HC=hard copy)						COST TO DATA?	GEOGRAPHIC RESOLUTION			SELECTED VARIABLES IN ELECTRONIC FORMAT					
	1999	2000	2001	2002	2003	2004		Address	Zip9	Zip5	Unique Subj. ID	ICD Death/ Disease	Age	Gender	Race// Ethnicity	Date of Death or Admis-sion
A. Mortality Data: PA Department of Health	E	E	E	E	E	E	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
B. Hospital In-Patient: PA Health Cost Care Containment Council	E	E	E	E	E	E	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
C. Emergency Department Visits:																
1. UPMC MARS System (UPMC hospitals)	E	E	E	E	E	E	Yes	Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
2. Data from individual (non-UPMC) hospitals	SEE BELOW						Unknown	SEE BELOW			SEE BELOW					
Allegheny County Hospitals																
Alle Kiski Medical Center	HC	HC	HC	HC	HC	HC		-		-	-	-	-	-	-	-
Allegheny General Hospital	E	E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Childrens Hospital of Pittsburgh	HC	HC	HC	E	E	E		Yes	Unk.	Yes	No	Yes	Yes	Yes	Yes	Yes
Forbes Regional Hospital	E	E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Jefferson Regional Medical Center	HC	HC	HC	HC	HC	HC		-	-	-	-	-	-	-	-	-
Magee Womens Hospital of UPMC	not obtained							not obtain-ed			not obtain-ed					
Mercy Hospital of Pittsburgh	HC	HC	HC	HC	HC	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	No	Yes
Mercy Providence Hospital	merged with Mercy Hospital 1/2004							-	-	-	-	-	-	-	-	-
Ohio Valley General Hospital	HC	HC	HC	HC	HC	HC		-	-	-	-	-	-	-	-	-
Sewickley Valley Hospital	HC	HC	E	E	E	E		No	Unk.	No	Yes	No	Yes	No	No	Yes
St. Clair Memorial Hospital	E	E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	No	Yes
Suburban General Hospital	HC	HC	HC	HC	HC	HC		-	-	-	-	-	-	-	-	-
UPMC Braddock	E	E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes

DATA TYPE AND SOURCE(S)	TIME PERIOD AVAILABLE/Data Format (E=electronic; HC=hard copy)						COST TO DATA?	GEOGRAPHIC RESOLUTION			SELECTED VARIABLES IN ELECTRONIC FORMAT					
	1999	2000	2001	2002	2003	2004		Address	Zip9	Zip5	Unique Subj. ID	ICD Death/ Disease	Age	Gender	Race// Ethnicity	Date of Death or Admis-sion
UPMC McKeesport	E	E	E	E	E	E		No	Unk.	No	Yes	No	Yes	No	No	Yes
UPMC Passavant	E	E	E	E	E	E		No	Unk.	Yes	Yes	Yes	Yes	Yes	No	Yes
UPMC Presbyterian/Shadyside	E	E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
UPMC South Side	HC	HC/ E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
UPMC St. Margaret	HC	HC	E	E	E	E		No	Unk.	No	Yes	Yes	No	No	Yes	Yes
Western Pennsylvania Hospital	E	E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
VA Healthcare (Federal)	E	E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Armstrong County Hospitals																
Armstrong County Memorial Hosp.	E	E	E	E	E	E		Yes	Unk.	Yes	No	Yes	Yes	Yes	Yes	Yes
Beaver County																
Aliquippa Community Hospital	HC	HC	HC	HC	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Medical Center Beaver	HC	HC	E	E	E	E		Yes	Unk.	Yes	Yes	No	Yes	Yes	Yes	Yes
Butler County																
Butler Memorial Hospital	E	E	E	E	E	E		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
UPMC Passavant Cranberry	not obtained							not obtain-ed			not obtain-ed					
Fayette County																
Brownsville General Hospital	HC	HC	HC	HC	HC	HC		-	-	-	-	-	-	-	-	-
Highlands Hospital	HC	HC	HC	HC	HC	p		-	-	-	-	-	-	-	-	-
Uniontown Hospital	HC	HC	HC	HC/ E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Washington County																
Canonsburg General Hospital	HC	HC	HC	HC	HC	HC		-	-	-	-	-	-	-	-	-
Monongahela Valley Hospital	E	E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Washington Hospital	E	E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes

DATA TYPE AND SOURCE(S)	TIME PERIOD AVAILABLE/Data Format (E=electronic; HC=hard copy)						COST TO DATA?	GEOGRAPHIC RESOLUTION			SELECTED VARIABLES IN ELECTRONIC FORMAT					
	1999	2000	2001	2002	2003	2004		Address	Zip9	Zip5	Unique Subj. ID	ICD Death/ Disease	Age	Gender	Race// Ethnicity	Date of Death or Admis-sion
Westmoreland County																
Frick Hospital	HC	HC	HC	HC/ E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Latrobe Area Hospital	HC	E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mercy Jeannette Hospital	E	E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Monsour Medical Center	HC	HC	HC	HC	HC	HC		-	-	-	-	-	-	-	-	-
Westmoreland Regional Hospital	E	E	E	E	E	E		Yes	Unk.	Yes	Yes	Yes	Yes	Yes	Yes	Yes

APPENDIX G – ICD 9/10 Code Comparability

ICD-10/ICD-9 Comparability Ratios for 113 Selected Causes of Death

How to Use This Table

To compare deaths coded with ICD-10 (used for deaths that occurred in 1999 and later) to deaths coded with ICD-9 (used for deaths that occurred between 1979 and 1998), multiply the number of deaths from earlier years by the comparability ratio for the specific cause of death.

For example, the ratio for Septicemia is 1.19, indicating that 19 percent more deaths were coded to this cause in ICD-10 than in ICD-9 solely because of the revision in the ICD. To analyze the trend in Septicemia deaths, multiply the number of Septicemia deaths in 1998 (317) by 1.19 to make it comparable to the 1999 Septicemia number. Comparing the 1998 number (modified to make it comparable = 379) to the 1999 number (376), we see that the number of deaths from this cause did not change appreciably from 1998 to 1999.

The standard error can be used to obtain a more definitive estimate of whether the numbers for ICD-10 years are different from those for ICD-9 years. The following example (again, for Septicemia) demonstrates how to use the standard error.

1. Multiply:
   Comparability ratio X standard error X 1.96 X the comparability-modified 1998 number.
   1.19 X 0.0042 X 1.96 X 379
   = 3.7127

2. Add and subtract this number to/from the comparability-modified 1998 number:
   379 – 3.7127 = 375.2873 (low point of the range);
   379 + 3.7127 = 382.7127 (high point of the range).

When these two numbers are rounded, the range (or confidence interval) is 375.3 to 382.7. The actual number of Septicemia deaths for 1999 (376) falls within this range. Therefore, there is a 95% probability that mortality due to Septicemia in 1999 was not substantially different from mortality due to Septicemia in 1998.

The National Center for Health Statistics (NCHS) provided the comparability ratios and standard errors used in the table.

ICD-10/ICD-9 Comparability Ratio Table **Page 2 of 9**

CAUSE OF DEATH (Based on the Tenth Revision, International Classification of Diseases)	1999	1998		Compar- ability Ratio	ICD-10 Codes	ICD-9 Codes	Standard Error
	Actual Number	Compara- bility Modified	Actual Number				
Salmonella infections	**0**	+	**1**	0.81	A01-A02	002-003	0.0644
Shigellosis and amebiasis	**0**	+	**0**	*	A03,A06	004,006	*
Certain other intestinal infections	**20**	+	**24**	0.60	A04,A07-A09	007-009	0.0248
Tuberculosis	**3**	+	**15**	0.85	A16-A19	010-018	0.0172
Respiratory tuberculosis	**2**	+	**9**	0.91	A16	010-012	0.0201
Other tuberculosis	**1**	+	**6**	0.70	A17-A19	013-018	0.0407
Whooping cough	**0**	+	**0**	*	A37	033	*
Scarlet fever and erysipelas	**0**	+	**0**	*	A38,A46	034.1-035	*
Meningococcal infection	**3**	+	**5**	1.00	A39	036	0.0149
Septicemia	**376**	**379**	**317**	1.19	A40-A41	038	0.0042
Syphilis	**0**	+	**1**	0.64	A50-A53	090-097	0.1184
Acute poliomyelitis	**0**	+	**0**	*	A80	045	*
Arthropod-borne viral encephalitis	**0**	+	**0**	*	A83-A84,A85.2	062-064	*
Measles	**0**	+	**0**	*	B05	055	*
Viral hepatitis	**15**	+	**34**	0.83	B15-B19	070	0.0120
Human immunodeficiency virus (HIV) disease	**62**	+	**60**	1.06	B20-B24	042-044	0.0018
Malaria	**0**	+	**0**	*	B50-B54	084	*
Other and unspecified infectious and parasitic diseases and their sequelae	**141**	**122**	**111**	1.10	A00,A05,A20-A36,A42-A44, A48-A49,A54-A79, A81-A82,A85.0-A85.1, A85.8,A86-B04,B06-B09, B25-B49,B55-B99	001,005,020-032, 037, 039-041,046-054, 056-061, 065-066,071-083, 085-088,098-134, 136-139,771.3	0.0154
Malignant neoplasms	**10,731**	**10,960**	**10,886**	1.01	C00-C97	140-208	0.0002
Malignant neoplasms of lip, oral cavity and pharynx	**142**	**169**	**176**	0.96	C00-C14	140-149	0.0040
Malignant neoplasm of esophagus	**266**	**273**	**273**	1.00	C15	150	0.0020

+ Comparability-modified ICD-9 deaths are not calculated if the actual number of ICD-9 deaths was less than 100

* Figure does not meet standards of reliability or precision.

--- Category not applicable

(Continued)

ICD-10/ICD-9 Comparability Ratio Table **Page 3 of 9**

CAUSE OF DEATH (Based on the Tenth Revision, International Classification of Diseases)	1999	1998		Compar- ability Ratio	ICD-10 Codes	ICD-9 Codes	Standard Error
	Actual Number	Compara- bility Modified	Actual Number				
Malignant neoplasm of stomach	**203**	**219**	**218**	1.01	C16	151	0.0019
Malignant neoplasms of colon, rectum and anus	**1,169**	**1,128**	**1,128**	1.00	C18-C21	153-154	0.0009
Malignant neoplasms of liver and intrahepatic bile ducts	**214**	**222**	**230**	0.96	C22	155	0.0023
Malignant neoplasm of pancreas	**576**	**576**	**576**	1.00	C25	157	0.0009
Malignant neoplasm of larynx	**69**	+	**63**	1.00	C32	161	0.0053
Malignant neoplasms of trachea, bronchus and lung	**2,642**	**2,753**	**2,799**	0.98	C33-C34	162	0.0005
Malignant melanoma of skin	**142**	**138**	**143**	0.97	C43	172	0.0032
Malignant neoplasm of breast	**834**	**869**	**864**	1.01	C50	174-175	0.0010
Malignant neoplasm of cervix uteri	**67**	**125**	**127**	0.99	C53	180	0.0034
Malignant neoplasms of corpus uteri and uterus, part unspecified	**141**	+	**59**	1.03	C54-C55	179,182	0.0040
Malignant neoplasm of ovary	**250**	**288**	**288**	1.00	C56	183.0	0.0016
Malignant neoplasm of prostate	**681**	**761**	**751**	1.01	C61	185	0.0015
Malignant neoplasms of kidney and renal pelvis	**246**	**227**	**227**	1.00	C64-C65	189.0,189.1	0.0022
Malignant neoplasm of bladder	**241**	**242**	**242**	1.00	C67	188	0.0026
Malignant neoplasms of meninges, brain and other parts of central nervous system	**252**	**277**	**286**	0.97	C70-C72	191-192	0.0025
Malignant neoplasms of lymphoid, hematopoietic and related tissue	**1,243**	**1,253**	**1,253**	1.00	C81-C96	200-208	0.0012
Hodgkin's disease	**22**	+	**40**	0.99	C81	201	0.0089
Non-Hodgkin's lymphoma	**519**	**519**	**531**	0.98	C82-C85	200,202	0.0018
Leukemia	**469**	**473**	**467**	1.01	C91-C95	204-208	0.0019
Multiple myeloma and immunoproliferative neoplasms	**231**	**223**	**215**	1.04	C88,C90	203	0.0030

+ Comparability-modified ICD-9 deaths are not calculated if the actual number of ICD-9 deaths was less than 100
* Figure does not meet standards of reliability or precision.
--- Category not applicable

(Continued)

ICD-10/ICD-9 Comparability Ratio Table **Page 4 of 9**

CAUSE OF DEATH (Based on the Tenth Revision, International Classification of Diseases)	1999	1998		Compar-ability Ratio	ICD-10 Codes	ICD-9 Codes	Standard Error
	Actual Number	Compara-bility Modified	Actual Number				
Other and unspecified malignant neoplasms of lymphoid, hematopoietic and related tissue	**2**	+	**0**	*	C96	---	*
All other and unspecified malignant neoplasms	**1,353**	**1,331**	**1,183**	1.13	C17,C23-C24,C26-C31, C37-C41,C44-C49, C51-C52, C57-C60, C62-C63, C66, C68-C69,C73-C80,C97	152,156,158-160, 163-171,173,181, 183.2-184,186-187, 189.2-190,193-199	0.0021
In situ neoplasms, benign neoplasms and neoplasms of uncertain or unknown behavior	**321**	**352**	**210**	1.67	D00-D48	210-239	0.0164
Anemias	**83**	+	**97**	0.96	D50-D64	280-285	0.0077
Diabetes mellitus	**1,273**	**1,284**	**1,274**	1.01	E10-E14	250	0.0011
Nutritional deficiencies	**81**	**144**	**124**	1.16	E40-E64	260-269	0.0165
Malnutrition	**57**	**118**	**121**	0.98	E40-E46	260-263	0.0151
Other nutritional deficiencies	**24**	+	**3**	6.20	E50-E64	264-269	0.5961
Meningitis	**17**	+	**9**	1.01	G00,G03	320-322	0.0136
Parkinson's disease	**340**	**311**	**311**	1.00	G20-G21	332	0.0028
Alzheimer's disease	**1,169**	**797**	**513**	1.55	G30	331.0	0.0071
Major cardiovascular diseases	**18,745**	**18,416**	**18,416**	1.00	I00-I78	390-434,436-448	0.0002
Diseases of heart	**13,802**	**13,479**	**13,673**	0.99	I00-I09,I11,I13,I20-I51	390-398,402,404,410-429	0.0002
Acute rheumatic fever and chronic rheumatic heart diseases	**78**	**97**	**118**	0.82	I00-I09	390-398	0.0089
Hypertensive heart disease	**434**	**315**	**392**	0.80	I11	402	0.0028
Hypertensive heart and renal disease	**47**	+	**32**	1.07	I13	404	0.0160
Ischemic heart diseases	**9,615**	**9,602**	**9,602**	1.00	I20-I25	410-414,429.2	0.0002
Acute myocardial infarction	**4,388**	**4,406**	**4,457**	0.99	I21-I22	410	0.0003

+ Comparability-modified ICD-9 deaths are not calculated if the actual number of ICD-9 deaths was less than 100
* Figure does not meet standards of reliability or precision.
--- Category not applicable

(Continued)

ICD-10/ICD-9 Comparability Ratio Table

Page 5 of 9

CAUSE OF DEATH (Based on the Tenth Revision, International Classification of Diseases)	1999	1998		Compar- ability Ratio	ICD-10 Codes	ICD-9 Codes	Standard Error
	Actual Number	Compara- bility Modified	Actual Number				
Other acute ischemic heart diseases	**27**	+	**11**	1.01	I24	411	0.0117
Other forms of chronic ischemic heart disease	**5,200**	**5,162**	**5,134**	1.01	I20,I25	412-414,429.2	0.0004
Atherosclerotic cardiovascular disease, so described	**693**	**703**	**670**	1.05	I25.0	429.2	0.0016
All other forms of chronic ischemic heart disease	**4,507**	**4,435**	**4,464**	0.99	I20,I25.1-I25.9	412-414	0.0004
Other heart diseases	**3,628**	**3,429**	**3,529**	0.97	I26-I51	415-429.1,429.3-429.9	0.0010
Acute and subacute endocarditis	**20**	+	**14**	1.00	I33	421	0.0137
Diseases of pericardium and acute myocarditis	**14**	+	**7**	1.03	I30-I31,I40	420,422-423	0.0160
Heart failure	**1,333**	**1,430**	**1,374**	1.04	I50	428	0.0013
All other forms of heart disease	**2,261**	**2,000**	**2,134**	0.94	I26-I28,I34-I38,I42-I49,I51	415-417,424-427, 429.0-429.1,429.3-429.9	0.0014
Essential (primary) hypertension and hypertensive renal disease	**246**	**223**	**199**	1.12	I10,I12	401,403	0.0050
Cerebrovascular diseases	**3,849**	**3,851**	**3,637**	1.06	I60-I69	430-434,436-438	0.0008
Atherosclerosis	**243**	**252**	**262**	0.96	I70	440	0.0025
Other diseases of circulatory system	**605**	**610**	**645**	0.95	I71-I78	441-448	0.0021
Aortic aneurysm and dissection	**366**	**399**	**399**	1.00	I71	441	0.0010
Other diseases of arteries, arterioles and capillaries	**239**	**209**	**246**	0.85	I72-I78	442-448	0.0053
Other disorders of circulatory system	**77**	+	**88**	1.03	I80-I99	451-459	0.0172
Influenza and pneumonia	**1,423**	**1,333**	**1,909**	0.70	J10-J18	480-487	0.0018
Influenza	**117**	**104**	**103**	1.01	J10-J11	487	0.0073
Pneumonia	**1,306**	**1,256**	**1,806**	0.70	J12-J18	480-486	0.0018
Other acute lower respiratory infections	**13**	+	**19**	0.97	J20-J22	466	0.0392

+ Comparability-modified ICD-9 deaths are not calculated if the actual number of ICD-9 deaths was less than 100

* Figure does not meet standards of reliability or precision.

--- Category not applicable

(Continued)

ICD-10/ICD-9 Comparability Ratio Table **Page 6 of 9**

CAUSE OF DEATH (Based on the Tenth Revision, International Classification of Diseases)	1999	1998					
	Actual Number	Compara-bility Modified	Actual Number	Compar-ability Ratio	ICD-10 Codes	ICD-9 Codes	Standard Error
Acute bronchitis and bronchiolitis	**7**	+	**19**	0.75	J20-J21	466	0.0264
Unspecified acute lower respiratory infection	**6**	+	**0**	*	J22	---	*
Chronic lower respiratory diseases	**2,266**	**2,153**	**2,055**	1.05	J40-J47	490-494,496	0.0009
Bronchitis, chronic and unspecified	**19**	+	**45**	0.39	J40-J42	490-491	0.0107
Emphysema	**329**	**304**	**313**	0.97	J43	492	0.0031
Asthma	**85**	+	**93**	0.89	J45-J46	493	0.0061
Other chronic lower respiratory diseases	**1,833**	**1,760**	**1,604**	1.10	J44,J47	494,496	0.0014
Pneumoconioses and chemical effects	**13**	+	**12**	1.02	J60-J66,J68	500-506	0.0099
Pneumonitis due to solids and liquids	**359**	**267**	**239**	1.12	J69	507	0.0048
Other diseases of respiratory system	**411**	**437**	**374**	1.17	J00-J06,J30-J39,J67,J70-J98	034.0,460-465, 470-478, 495,508-519	0.0052
Peptic ulcer	**100**	+	**83**	0.97	K25-K28	531-534	0.0045
Diseases of appendix	**8**	+	**7**	1.03	K35-K38	540-543	0.0242
Hernia	**35**	+	**31**	1.04	K40-K46	550-553	0.0154
Chronic liver disease and cirrhosis	**402**	**402**	**388**	1.04	K70,K73-K74	571	0.0027
Alcoholic liver disease	**199**	**198**	**194**	1.02	K70	571.0-571.3	0.0050
Other chronic liver disease and cirrhosis	**203**	**204**	**194**	1.05	K73-K74	571.4-571.9	0.0041
Cholelithiasis and other disorders of gallbladder	**71**	+	**53**	0.96	K80-K82	574-575	0.0060
Nephritis, nephrotic syndrome and nephrosis	**677**	**710**	**576**	1.23	N00-N07,N17-N19,N25-N27	580-589	0.0044
Acute and rapidly progressive nephritic and nephrotic syndrome	**8**	+	**3**	0.65	N00-N01,N04	580-581	0.0342
Chronic glomerulonephritis, nephritis and nephritis not specified as acute or chronic, and renal sclerosis unspecified	**7**	+	**28**	0.39	N02-N03,N05-N07,N26	582-583,587	0.0144

+ Comparability-modified ICD-9 deaths are not calculated if the actual number of ICD-9 deaths was less than 100

* Figure does not meet standards of reliability or precision.

--- Category not applicable

(Continued)

ICD-10/ICD-9 Comparability Ratio Table **Page 7 of 9**

CAUSE OF DEATH (Based on the Tenth Revision, International Classification of Diseases)	1999	1998		Compar-ability Ratio	ICD-10 Codes	ICD-9 Codes	Standard Error
	Actual Number	Compara-bility Modified	Actual Number				
Renal failure	**661**	**704**	**544**	1.29	N17-N19	584-586	0.0050
Other disorders of kidney	**1**	+	**1**	0.91	N25,N27	588-589	0.0867
Infections of kidney	**21**	+	**15**	1.01	N10-N12,N13.6,N15.1	590	0.0144
Hyperplasia of prostate	**12**	+	**12**	1.00	N40	600	0.0159
Inflammatory diseases of female pelvic organs	**5**	+	**2**	0.98	N70-N76	614-616	0.0410
Pregnancy, childbirth and the puerperium	**4**	+	**6**	*	O00-099	630-676	*
Certain conditions originating in the perinatal period	**213**	**234**	**220**	1.07	P00-P96	760-771.2,771.4-779	0.0033
Congenital malformations, deformations and chromosomal abnormalities	**201**	**202**	**238**	0.85	Q00-Q99	740-759	0.0055
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	**198**	**154**	**161**	0.96	R00-R99	780-799	0.0034
Sudden infant death syndrome	**55**	+	**45**	1.04	R95	798.0	0.0040
All other diseases (Residual)	**3,838**	**3,812**	**4,237**	0.90	Residual	Residual	0.0015
Accidents	**1,943**	**1,927**	**1,870**	1.03	V01-X59,Y85-Y86	E800-E869,E880-E929	0.0014
Transport accidents	**824**	**813**	**813**	1.00	V01-V99,Y85	E800-E848, E929.0,E929.1	0.0006
Motor vehicle accidents	**761**	**642**	**753**	0.85	V02-V04, V09.0, V09.2, V12-V14,V19.0-V19.2, V19.4-V19.6,V20-V79, V80.3-V80.5,V81.0-V81.1, V82.0-V82.1,V83-V86, V87.0-V87.8,V88.0-V88.8, V89.0,V89.2	E810-E825	0.0027

+ Comparability-modified ICD-9 deaths are not calculated if the actual number of ICD-9 deaths was less than 100

* Figure does not meet standards of reliability or precision.

--- Category not applicable

(Continued)

ICD-10/ICD-9 Comparability Ratio Table **Page 8 of 9**

CAUSE OF DEATH (Based on the Tenth Revision, International Classification of Diseases)	1999	1998		Compar-ability Ratio	ICD-10 Codes	ICD-9 Codes	Standard Error
	Actual Number	Compara-bility Modified	Actual Number				
Other land transport accidents	**32**	+	**9**	14.13	V01,V05-V06, V09.1, V09.3-V09.9, V10-V11, V15-V18, V19.3, V19.8-V19.9, V80.0-V80.2, V80.6-V80.9,V81.2-V81.9, V82.2-V82.9, V87.9, V88.9,V89.1,V89.3,V89.9	E800-E807,E826-E829	0.9952
Water, air and space, and other and unspecified transport accidents and their sequelae	**31**	**0**	**51**	1.01	V90-V99,Y85	E830-E848, E929.0,E929.1	0.0209
Nontransport accidents	**1,119**	**1,138**	**1,057**	1.08	W00-X59,Y86	E850-E869,E880-E928,E929.2-E929.9	0.0035
Falls	**595**	**437**	**520**	0.84	W00-W19	E880-E888	0.0049
Accidental discharge of firearms	**14**	+	**65**	1.06	W32-W34	E922	0.0127
Accidental drowning and submersion	**65**	+	**67**	1.00	W65-W74	E910	0.0127
Accidental exposure to smoke, fire and flames	**52**	+	**65**	0.97	X00-X09	E890-E899	0.0089
Accidental poisoning and exposure to noxious Substances	**139**	**127**	**143**	0.89	X40-X49	E850-E869,E924.1	0.0191
Other and unspecified nontransport accidents and their sequelae	**254**	**349**	**246**	1.42	W20-W31, W35-W64, W75-W99,X10-X39, X50-X59,Y86	E900-E909, E911-E921, E923-E924.0, E924.8-E928, E929.2-E929.9	0.0123
Intentional self-harm (suicide)	**598**	**593**	**593**	1.00	X60-X84,Y87.0	E950-E959	0.0005
Intentional self-harm (suicide) by discharge of Firearms	**301**	**312**	**312**	1.00	X72-X74	E955.0-E955.4	0.0007
Intentional self-harm (suicide) by other and unspecified means and their sequelae	**297**	**281**	**281**	1.00	X60-X71,X75-X84,Y87.0	E950-E954, E955.5-E959	0.0023
Assault (homicide)	**203**	**188**	**188**	1.00	X85-Y09,Y87.1	E960-E969	0.0006
Assault (homicide) by discharge of firearms	**129**	**114**	**114**	1.00	X93-X95	E965.0-E965.4	0.0008

+ Comparability-modified ICD-9 deaths are not calculated if the actual number of ICD-9 deaths was less than 100

* Figure does not meet standards of reliability or precision.

--- Category not applicable

(Continued)

ICD-10/ICD-9 Comparability Ratio Table **Page 9 of 9**

CAUSE OF DEATH (Based on the Tenth Revision, International Classification of Diseases)	1999	1998					
	Actual Number	Compara-bility Modified	Actual Number	Compar-ability Ratio	ICD-10 Codes	ICD-9 Codes	Standard Error
Assault (homicide) by other and unspecified means and their sequelae	**74**	+	**74**	1.00	X85-X92,X96-Y09,Y87.1	E960-E964, E965.5-E969	0.0024
Legal intervention	**4**	+	**5**	*	Y35,Y89.0	E970-E978	*
Events of undetermined intent	**48**	+	**49**	*	Y10-Y34,Y87.2,Y89.9	E980-E989	*
Discharge of firearms, undetermined intent	**3**	+	**3**	*	Y22-Y24	E985.0-E985.4	*
Other and unspecified events of undetermined intent and their sequelae	**45**	+	**46**	*	Y10-Y21,Y25-Y34,Y87.2,Y89.9	E980-E984, E985.5-E989	*
Operations of war and their sequelae	**0**	+	**0**	*	Y36,Y89.1	E990-E999	*
Complications of medical and surgical care	**48**	+	**52**	*	Y40-Y84,Y88	E870-E879,E930-E949	*
Injury by firearms	**451**	**450**	**450**	1.00	W32-W34,X72-X74, X93-X95, Y22-Y24,Y35.0	E922, E955.0-E955.4, E965.0-E965.4, E970, E985.0-E985.4	0.0006
Drug-induced deaths	**239**	**272**	**228**	1.20	F11.0-F11.5,F11.7-F11.9, F12.0-F12.5,F12.7-F12.9, F13.0-F13.5, F13.7-F13.9, F14.0-F14.5, F14.7-F14.9, F15.0-F15.5, F15.7-F15.9, F16.0-F16.5, F16.7-F16.9, F17.0, F17.3-F17.5, F17.7-F17.9, F18.0-F18.5, F18.7-F18.9, F19.0-F19.5, F19.7-F19.9, X40-X44, X60-X64, X85, Y10-Y14	292,304, 305.2-305.9, E850-E858, E950.0-E950.5,E962.0, E980.0-E980.5	0.0225
Alcohol-induced deaths	**348**	**370**	**382**	0.97	F10,G31.2,G62.1,I42.6, K29.2,K70,R78.0,X45,X65 ,Y15	291,303,305.0,357.5, 425.5,535.3,571.0-571.3,790.3,E860	0.0025

Source: Bureau of Health Information, Division of Health Care Financing, Department of Health and Family Services. The comparability ratio for each cause of death is from the National Center for Health Statistics, December 5, 2000.

+ Comparability-modified ICD-9 deaths are not calculated if the actual number of ICD-9 deaths was less than 100
* Figure does not meet standards of reliability or precision.
--- Category not applicable

APPENDIX H – PHC4 Data Layouts

PHC4 Data Layout: 1999-2002

Special Requests Public Inpatient File Layout and Supporting Documentation, 1994 – 2002

DATA FILE

FIELD NAME	DATA ELEMENT DESCRIPTION	DATA FILE LOCATION	* FIELD FORMAT	NOTES
	Record Identifier			
SYSID	System assigned unique (for quarter) record sequence number	001-007	X(7)	Unique number for each record for a given facility for each quarter
YEAR	Processing Year	008-011	X(4)	
QUARTER	Processing Quarter	012	X(1)	Quarter of year
	Facility Identification			
PAF	Facility Number (PAF)	013-016	X(4)	Numeric portion of PHC4 assigned facility #
HREGION	Facility Region Code	017	X(1)	Standard PHC4 Region Code (1 through 9)
MAID	MAID (Pa. Medical Assistance Identifier)	018-025	X(8)	No values available effective 2003
	Patient Data			
PTSEX	Sex Code	026	X(1)	
ETHNIC	Hispanic/Latino Origin or Descent	027	X(1)	
RACE	Race Code	028	X(1)	
PSEUDOID	Pseudo Patient Identifier	029-038	X(10)	PHC4 assigned unique patient identifier
AGE	Patient Age in Years	039-041	9(3)	Zero if less than 1 year or unknown
AGECAT	Patient Age in Days (if less than 1 year old)	042-043	X(2)	
PRIVZIP	Home Zip Code	044-048	X(5)	
MKTSHARE	Home Market Share Area Code	049-051	X(3)	Future - to be developed by PHC4
COUNTY	Patient Home County Code	052-054	X(3)	Pa. Counties coded as 1 through 67
STATE	State Code	055-056	X(2)	USPS standard state code
	Admission Data			
ADTYPE	Admission Type	057	X(1)	
ADSOURCE	Admission Source	058	X(1)	
ADHOUR	Admission Hour	059-060	X(2)	Military time (24 hour clock)
ADMDX	Admitting Diagnosis	061-066	X(6)	
ADDOW	Admission Day of Week	067	X(1)	Code for day (1 = Sunday)
	Discharge Data			
DCSTATUS	Discharge Status	068-069	X(2)	
LOS	Length of Stay	070-074	9(5)	
DCHOUR	Discharge Hour	075-076	X(2)	
DCDOW	Discharge Day of Week	077	X(1)	Code for day (1 = Sunday)
	Diagnosis Codes			

***Field Format**
x = text
9 = numeric

Special Requests Public Inpatient File Layout and Supporting Documentation, 1994 – 2002

DATA FILE

FIELD NAME	DATA ELEMENT DESCRIPTION	DATA FILE LOCATION	* FIELD FORMAT	NOTES
ECODE	E-code (External Cause of Injury Code)	078-083	X(6)	
PDX	Principal Diagnosis Code	084-089	X(6)	
SDX1	Secondary Diagnosis Code (1)	090-095	X(6)	
SDX2	Secondary Diagnosis Code (2)	096-101	X(6)	
SDX3	Secondary Diagnosis Code (3)	102-107	X(6)	
SDX4	Secondary Diagnosis Code (4)	108-113	X(6)	
SDX5	Secondary Diagnosis Code (5)	114-119	X(6)	
SDX6	Secondary Diagnosis Code (6)	120-125	X(6)	
SDX7	Secondary Diagnosis Code (7)	126-131	X(6)	
SDX8	Secondary Diagnosis Code (8)	132-137	X(6)	
	Procedure Codes			
PPX	Principal Procedure Code	138-144	X(7)	
SPX1	Secondary Procedure Code (1)	145-151	X(7)	
SPX2	Secondary Procedure Code (2)	152-158	X(7)	
SPX3	Secondary Procedure Code (3)	159-165	X(7)	
SPX4	Secondary Procedure Code (4)	166-172	X(7)	
SPX5	Secondary Procedure Code (5)	173-179	X(7)	
	Procedure Day of Week			
PPXDOW	Principal Procedure	180	X(1)	Code for day of procedure (1 = Sunday)
SPX1DOW	Secondary Procedure Code (1)	181	X(1)	
SPX2DOW	Secondary Procedure Code (2)	182	X(1)	
SPX3DOW	Secondary Procedure Code (3)	183	X(1)	
SPX4DOW	Secondary Procedure Code (4)	184	X(1)	
SPX5DOW	Secondary Procedure Code (5)	185	X(1)	
	Pennsylvania State License Number			
REFID	Referring Physician	186-194	X(9)	
ATTID	Attending Physician	195-203	X(9)	
OPERID	Operating Physician	204-212	X(9)	
	Payor Identification			
PAYTYPE1	One	213-214	X(2)	
PAYTYPE2	Two	215-216	X(2)	
PAYTYPE3	Three	217-218	X(2)	

*__Field Format__
x = text
9 = numeric

Special Requests Public Inpatient File Layout and Supporting Documentation, 1994 – 2002

DATA FILE

FIELD NAME	DATA ELEMENT DESCRIPTION	DATA FILE LOCATION	* FIELD FORMAT	NOTES
ESTPAYER	Estimated Payor Code	219-220	X(2)	Not filled after Q3 1998
NAIC	NAIC	221-227	X(7)	
BILLTYPE	Type of Bill	228-230	X(3)	
DRGHOSP	DRG - Diagnosis Related Group – Hospital Submitted	231-233	X(3)	
PCMU	Procedure Coding Method Used	234	X(1)	
DRGHC4	DRG - Diagnosis Related Group – PHC4	235-237	X(3)	
	Cancer ID			
CANCER1	1	238	X(1)	
CANCER2	2	239	X(1)	
MDCHC4	MDC	240-241	X(2)	
MQSEV	Severity – MediQual	242	X(1)	
MQNRSP	Non-Responder – MediQual	243	X(1)	This data element is no longer available effective 2002Q4.
	Summary Charges			
PROFCHG	Professional Fees	244-254	9(11)	Leading sign character (- if minus); seven whole numbers with two for cents
TOTALCHG	Total Of Other Charges (exclusive of Prof. Fees)	255-265	9(11)	
NONCVCHG	Non-covered Charges	266-276	9(11)	
ROOMCHG	Room & Board Charges	277-287	9(11)	
ANCLRCHG	Ancillary Charges	288-298	9(11)	
DRUGCHG	Drug Charges	299-309	9(11)	
EQUIPCHG	Equipment Charges	310-320	9(11)	
SPECLCHG	Specialty Charges	321-331	9(11)	
MISCCHG	Miscellaneous Charges	332-342	9(11)	
	APR Grouper Data			
APRMDC	APR MDC	343-344	X(2)	No values available for all APR elements effective 2003
APRDRG	APR DRG	345-347	X(3)	
APRSOI	APR Severity of Illness Subclass	348	X(1)	
APRROM	APR Risk of Mortality Subclass	349	X(1)	
	MediQual			
MQGCLUST	MQ Grouper Score – Cluster	350-353	X(4)	
MQGCELL	MQ Grouper Score – Cell	354	X(1)	

*__Field Format__
x = text
9 = numeric

Special Requests Public Inpatient File Layout and Supporting Documentation, 1994 – 2002

FACILITY PROFILE

Quote/comma delimited format

DATA ELEMENT NAME	FORMAT	NOTES
Facility Number (1234)	string - 4	Standard PHC4 facility identifier
Facility Name	string - 35	
Facility Type	string - 3	
Facility bed count	numeric - 6	Actual bed count of facility
PHC4 Region Code	string - 1	
Facility County Code	string - 3	Standard Pa. County codes 1 through 67
Unit ID1	string - 7	
Unit Type1	string - 3	
Unit ID2	string - 7	
Unit Type2	string - 3	
Unit ID3	string - 7	
Unit Type3	string - 3	
Unit ID4	string - 7	
Unit Type4	string - 3	
Unit ID5	string - 7	
Unit Type5	string - 3	
IPDisch	numeric - 10	This field will be null prior to quarter being finalized.
OPDisch	numeric - 10	This field will be null prior to quarter being finalized.

PHYSICIAN PROFILE FILE

Quote/comma delimited format

DATA ELEMENT NAME	FORMAT	NOTES
Physician License Number	String - 9	Pa. Assigned License Number; reference UB-92 Data Collection Manual
Physician Last Name	String – 12	
Physician Initials	String - 2	

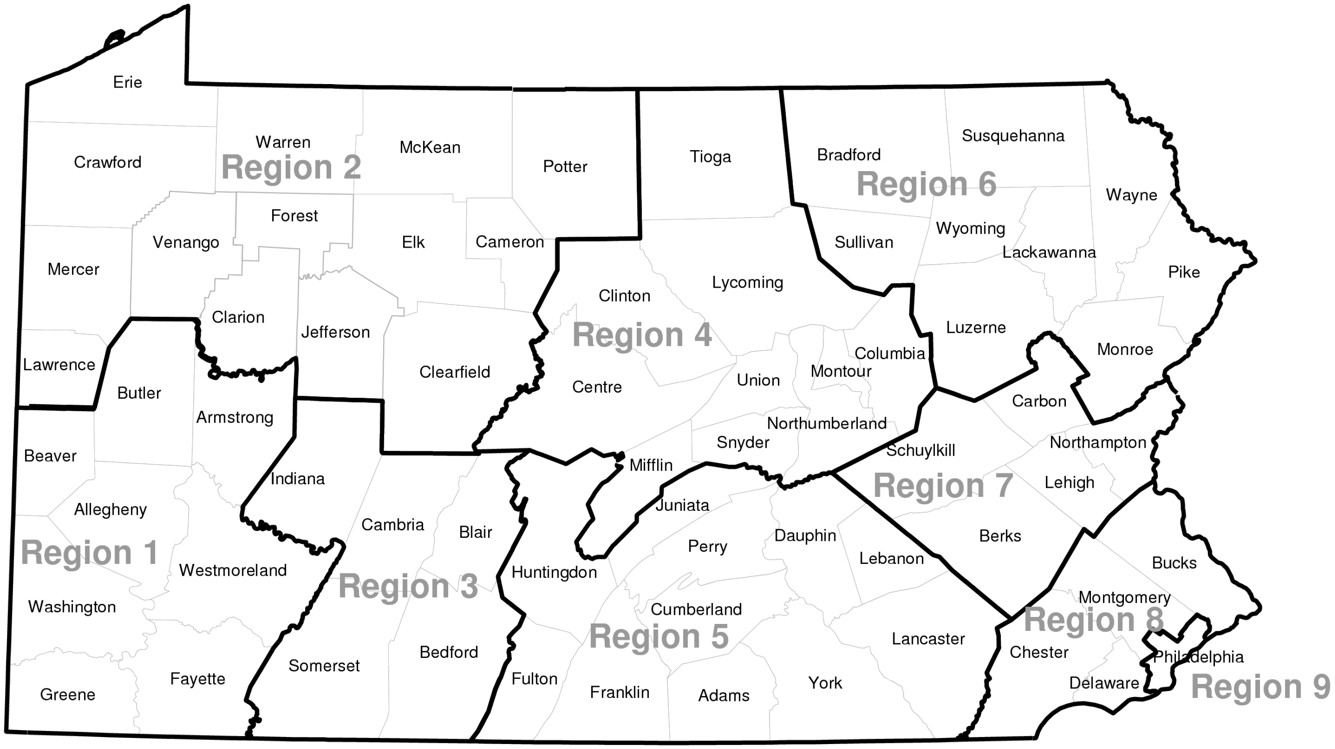
Special Requests Public Inpatient File Layout and Supporting Documentation, 1994 – 2002

STATE & COUNTY CODES

AL – Alabama
AK – Alaska
AZ – Arizona
AR – Arkansas
CA – California
CO – Colorado
CT – Connecticut
DE – Delaware
DC – District of Columbia
FL – Florida
GA – Georgia
HI – Hawaii
ID – Idaho
IL – Illinois
IN – Indiana
IA – Iowa
KS – Kansas
KY – Kentucky
LA – Louisiana
ME – Maine
MD – Maryland
MA – Massachusetts
MI – Michigan
MN – Minnesota
MS – Mississippi
MO – Missouri
MT – Montana
NE – Nebraska
NV – Nevada
NH – New Hampshire
NJ – New Jersey
NM – New Mexico
NY – New York
NC – North Carolina
ND – North Dakota
OH – Ohio
OK – Oklahoma
OR – Oregon
PA – Pennsylvania
PR – Puerto Rico
RI – Rhode Island
SC – South Carolina
SD – South Dakota
TN – Tennessee
TX – Texas
UT – Utah
VT – Vermont
VI – Virgin Islands
VA – Virginia
WA – Washington
WV – West Virginia
WI – Wisconsin
WY – Wyoming

PENNSYLVANIA COUNTY CODES

01 – Adams (Region 5)
02 – Allegheny (Region 1)
03 – Armstrong (Region 1)
04 – Beaver (Region 1)
05 – Bedford (Region 3)
06 – Berks (Region 7)
07 – Blair (Region 3)
08 – Bradford (Region 6)
09 – Bucks (Region 8)
10 – Butler (Region 1)
11 – Cambria (Region 3)
12 – Cameron (Region 2)
13 – Carbon (Region 7)
14 – Centre (Region 4)
15 – Chester (Region 8)
16 – Clarion (Region 2)
17 – Clearfield (Region 2)
18 – Clinton (Region 4)
19 – Columbia (Region 4)
20 – Crawford (Region 2)
21 – Cumberland (Region 5)
22 – Dauphin (Region 5)
23 – Delaware (Region 8)
24 – Elk (Region 2)
25 – Erie (Region 2)
26 – Fayette (Region 1)
27 – Forest (Region 2)
28 – Franklin (Region 5)
29 – Fulton (Region 5)
30 – Greene (Region 1)
31 – Huntingdon (Region 5)
32 – Indiana (Region 3)
33 – Jefferson (Region 2)
34 – Juniata (Region 5)
35 – Lackawanna (Region 6)
36 – Lancaster (Region 5)
37 – Lawrence (Region 2)
38 – Lebanon (Region 5)
39 – Lehigh (Region 7)
40 – Luzerne (Region 6)
41 – Lycoming (Region 4)
42 – McKean (Region 2)
43 – Mercer (Region 2)
44 – Mifflin (Region 4)
45 – Monroe (Region 6)
46 – Montgomery (Region 8)
47 – Montour (Region 4)
48 – Northampton (Region 7)
49 – Northumberland (Region 4)
50 – Perry (Region 5)
51 – Philadelphia (Region 9)
52 – Pike (Region 6)
53 – Potter (Region 2)
54 – Schuylkill (Region 7)
55 – Snyder (Region 4)
56 – Somerset (Region 3)
57 – Sullivan (Region 6)
58 – Susquehanna (Region 6)
59 – Tioga (Region 4)
60 – Union (Region 4)
61 – Venango (Region 2)
62 – Warren (Region 2)
63 – Washington (Region 1)
64 – Wayne (Region 6)
65 – Westmoreland (Region 1)
66 – Wyoming (Region 6)
67 – York (Region 5)


Erie
Crawford
Warren
Region 2
McKean
Potter
Tioga
Bradford
Susquehanna
Region 6
Wayne
Forest
Venango
Elk
Cameron
Mercer
Sullivan
Wyoming
Lackawanna
Pike
Clarion
Clinton
Lycoming
Jefferson
Region 4
Luzerne
Lawrence
Columbia
Monroe
Clearfield
Montour
Union
Butler
Centre
Carbon
Armstrong
Northumberland
Beaver
Snyder
Northampton
Schuylkill
Mifflin
Indiana
Region 7
Lehigh
Juniata
Allegheny
Cambria
Blair
Dauphin
Region 1
Perry
Berks
Lebanon
Bucks
Westmoreland
Huntingdon
Region 3
Montgomery
Washington
Cumberland
Region 8
Lancaster
Region 5
Chester
Bedford
Philadelphia
Somerset
Fayette
Fulton
Delaware
York
Region 9
Greene
Franklin
Adams

PHC4 Data Layout: 2003-2005

Special Requests Public Inpatient File Layout and Supporting Documentation, 2003 - 2005

DATA FILE

FIELD NAME	DATA ELEMENT DESCRIPTION	DATA FILE LOCATION	* FIELD FORMAT	NOTES
	Record Identifier			
SYSID	System assigned unique (for quarter) record sequence number	001-007	X(7)	Unique number for each record for a given facility for each quarter
YEAR	Processing Year	008-011	X(4)	
QUARTER	Processing Quarter	012	X(1)	Quarter of year
	Facility Identification			
PAF	Facility Number (PAF)	013-016	X(4)	Numeric portion of PHC4 assigned facility #
HREGION	Facility Region Code	017	X(1)	Standard PHC4 Region Code (1 through 9)
	Patient Data			
PTSEX	Sex Code	018	X(1)	
ETHNIC	Hispanic/Latino Origin or Descent	019	X(1)	
RACE	Race Code	020	X(1)	
PSEUDOID	Pseudo Patient Identifier	021-030	X(10)	PHC4 assigned unique patient identifier
AGE	Patient Age in Years	031-033	9(3)	Zero if less than 1 year or unknown
AGECAT	Patient Age in Days (if less than 1 year old)	034-035	X(2)	
ZIP	Home Zip Code	036-040	X(5)	Effective 2004Q3 the zip code will no longer be privatized.
MKTSHARE	Home Market Share Area Code	041-043	X(3)	Future - to be developed by PHC4
COUNTY	Patient Home County Code	044-046	X(3)	Pa. Counties coded as 1 through 67
STATE	State Code	047-048	X(2)	USPS standard state code
	Admission Data			
ADTYPE	Admission Type	049	X(1)	
ADSOURCE	Admission Source	050	X(1)	
ADHOUR	Admission Hour	051-052	X(2)	Military time (24 hour clock)
ADMDX	Admitting Diagnosis	053-058	X(6)	
ADDOW	Admission Day of Week	059	X(1)	Code for day (1 = Sunday)
	Discharge Data			
DCSTATUS	Discharge Status	060-061	X(2)	
LOS	Length of Stay	062-066	9(5)	
DCHOUR	Discharge Hour	067-068	X(2)	
DCDOW	Discharge Day of Week	069	X(1)	Code for day (1 = Sunday)

***Field Format**
x = text
9 = numeric

PA Health Care Cost Containment Council
Special Requests Unit

1

Special Requests Public Inpatient File Layout and Supporting Documentation, 2003 - 2005

DATA FILE

FIELD NAME	DATA ELEMENT DESCRIPTION	DATA FILE LOCATION	* FIELD FORMAT	NOTES
	Diagnosis Codes			
ECODE	E-code (External Cause of Injury Code)	070-075	X(6)	
PDX	Principal Diagnosis Code	076-081	X(6)	
SDX1	Secondary Diagnosis Code (1)	082-087	X(6)	
SDX2	Secondary Diagnosis Code (2)	088-093	X(6)	
SDX3	Secondary Diagnosis Code (3)	094-099	X(6)	
SDX4	Secondary Diagnosis Code (4)	100-105	X(6)	
SDX5	Secondary Diagnosis Code (5)	106-111	X(6)	
SDX6	Secondary Diagnosis Code (6)	112-117	X(6)	
SDX7	Secondary Diagnosis Code (7)	118-123	X(6)	
SDX8	Secondary Diagnosis Code (8)	124-129	X(6)	
	Procedure Codes			
PPX	Principal Procedure Code	130-136	X(7)	
SPX1	Secondary Procedure Code (1)	137-143	X(7)	
SPX2	Secondary Procedure Code (2)	144-150	X(7)	
SPX3	Secondary Procedure Code (3)	151-157	X(7)	
SPX4	Secondary Procedure Code (4)	158-164	X(7)	
SPX5	Secondary Procedure Code (5)	165-171	X(7)	
	Procedure Day of Week			
PPXDOW	Principal Procedure	172	X(1)	Code for day of procedure (1 = Sunday)
SPX1DOW	Secondary Procedure Code (1)	173	X(1)	
SPX2DOW	Secondary Procedure Code (2)	174	X(1)	
SPX3DOW	Secondary Procedure Code (3)	175	X(1)	
SPX4DOW	Secondary Procedure Code (4)	176	X(1)	
SPX5DOW	Secondary Procedure Code (5)	177	X(1)	
	Pennsylvania State License Number			
REFID	Referring Physician	178-186	X(9)	
ATTID	Attending Physician	187-195	X(9)	
OPERID	Operating Physician	196-204	X(9)	
	Payor Identification			
PAYTYPE1	One	205-206	X(2)	
PAYTYPE2	Two	207-208	X(2)	

*__Field Format__
x = text
9 = numeric

Special Requests Public Inpatient File Layout and Supporting Documentation, 2003 - 2005

DATA FILE

FIELD NAME	DATA ELEMENT DESCRIPTION	DATA FILE LOCATION	* FIELD FORMAT	NOTES
PAYTYPE3	Three	209-210	X(2)	
ESTPAYER	Estimated Payor Code	211-212	X(2)	Not filled after Q3 1998
NAIC	NAIC	213-219	X(7)	
BILLTYPE	Type of Bill	220-222	X(3)	
DRGHOSP	DRG - Diagnosis Related Group – Hospital Submitted	223-225	X(3)	
PCMU	Procedure Coding Method Used	226	X(1)	
DRGHC4	DRG - Diagnosis Related Group – PHC4	227-229	X(3)	
	Cancer ID			
CANCER1	1	230	X(1)	
CANCER2	2	231	X(1)	
MDCHC4	MDC	232-233	X(2)	
MQSEV	Severity – MediQual	234	X(1)	
MQNRSP	Non-Responder – MediQual	235	X(1)	This data element is no longer available effective 2002Q4.
	Summary Charges			
PROFCHG	Professional Fees	236-246	9(11)	Leading sign character (- if minus); seven whole numbers with two for cents
TOTALCHG	Total Of Other Charges (exclusive of Prof. Fees)	247-257	9(11)	
NONCVCHG	Non-covered Charges	258-268	9(11)	
ROOMCHG	Room & Board Charges	269-279	9(11)	
ANCLRCHG	Ancillary Charges	280-290	9(11)	
DRUGCHG	Drug Charges	291-301	9(11)	
EQUIPCHG	Equipment Charges	302-312	9(11)	
SPECLCHG	Specialty Charges	313-323	9(11)	
MISCCHG	Miscellaneous Charges	324-334	9(11)	
	MediQual			
MQGCLUST	MQ Grouper Score – Cluster	335-338	X(4)	
MQGCELL	MQ Grouper Score – Cell	339	X(1)	

*Field Format
x = text
9 = numeric

Special Requests Public Inpatient File Layout and Supporting Documentation, 2003 - 2005

FACILITY PROFILE FILE – EFFECTIVE THROUGH 2003Q3

Quote/comma delimited format

DATA ELEMENT NAME	FORMAT	NOTES
Facility Number (1234)	string - 4	Standard PHC4 facility identifier
Facility Name	string - 35	
Facility Type	string - 3	
Facility bed count	numeric - 6	Actual bed count of facility
PHC4 Region Code	string - 1	
Facility County Code	string - 3	Standard Pa. County codes 1 through 67
Unit ID1	string - 7	
Unit Type1	string - 3	
Unit ID2	string - 7	
Unit Type2	string - 3	
Unit ID3	string - 7	
Unit Type3	string - 3	
Unit ID4	string - 7	
Unit Type4	string - 3	
Unit ID5	string - 7	
Unit Type5	string - 3	
IPDisch	numeric - 10	This field will be null prior to quarter being finalized.
OPDisch	numeric - 10	This field will be null prior to quarter being finalized.

FACILITY PROFILE FILE – EFFECTIVE 2003Q4 THROUGH PRESENT

Quote/comma delimited format

DATA ELEMENT NAME	FORMAT	NOTES
Facility Number (1234)	string - 4	Standard PHC4 facility identifier
Facility Name	string - 35	
Facility Type	string - 3	
Facility bed count	numeric - 6	Actual bed count of facility
PHC4 Region Code	string - 1	
Facility County Code	string - 3	Standard Pa. County codes 1 through 67
MAID	string - 7	
IPDisch	numeric - 10	
OPDisch	numeric - 10	

Special Requests Public Inpatient File Layout and Supporting Documentation, 2003 - 2005

PHYSICIAN PROFILE FILE

Quote/comma delimited format

DATA ELEMENT NAME	FORMAT	NOTES
Physician License Number	string - 9	Pa. Assigned License Number; reference UB-92 Data Collection Manual
Physician Last Name	string - 12	
Physician Initials	string - 2	

Special Requests Public Inpatient File Layout and Supporting Documentation, 2003 - 2005

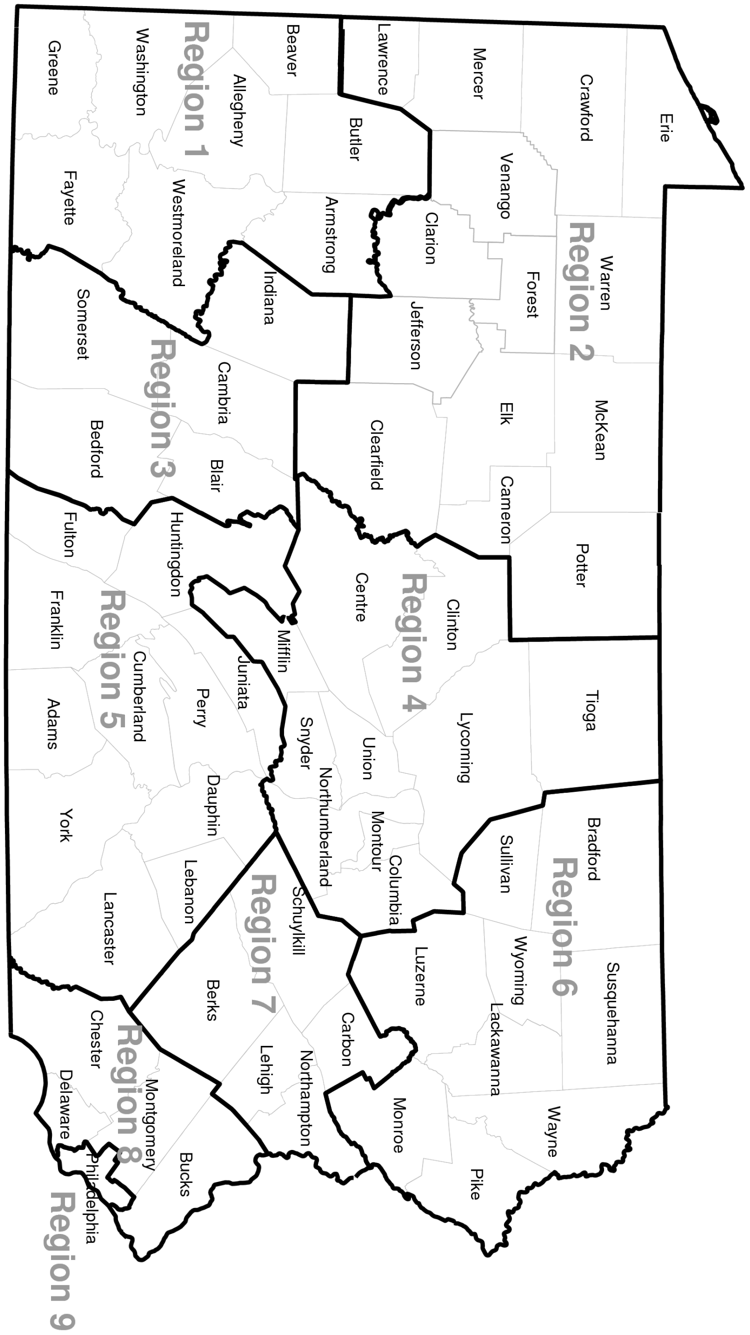
STATE & COUNTY CODES

AL – Alabama
AK – Alaska
AZ – Arizona
AR – Arkansas
CA – California
CO – Colorado
CT – Connecticut
DE – Delaware
DC – District of Columbia
FL – Florida
GA – Georgia
HI – Hawaii
ID – Idaho
IL – Illinois
IN – Indiana
IA – Iowa
KS – Kansas
KY – Kentucky
LA – Louisiana
ME – Maine
MD – Maryland
MA – Massachusetts
MI – Michigan
MN – Minnesota
MS – Mississippi
MO – Missouri
MT – Montana
NE – Nebraska
NV – Nevada
NH – New Hampshire
NJ – New Jersey
NM – New Mexico
NY – New York
NC – North Carolina
ND – North Dakota
OH – Ohio
OK – Oklahoma
OR – Oregon
PA – Pennsylvania
PR – Puerto Rico
RI – Rhode Island
SC – South Carolina
SD – South Dakota
TN – Tennessee
TX – Texas
UT – Utah
VT – Vermont
VI – Virgin Islands
VA – Virginia
WA – Washington
WV – West Virginia
WI – Wisconsin
WY – Wyoming

PENNSYLVANIA COUNTY CODES

01 – Adams (Region 5)
02 – Allegheny (Region 1)
03 – Armstrong (Region 1)
04 – Beaver (Region 1)
05 – Bedford (Region 3)
06 – Berks (Region 7)
07 – Blair (Region 3)
08 – Bradford (Region 6)
09 – Bucks (Region 8)
10 – Butler (Region 1)
11 – Cambria (Region 3)
12 – Cameron (Region 2)
13 – Carbon (Region 7)
14 – Centre (Region 4)
15 – Chester (Region 8)
16 – Clarion (Region 2)
17 – Clearfield (Region 2)
18 – Clinton (Region 4)
19 – Columbia (Region 4)
20 – Crawford (Region 2)
21 – Cumberland (Region 5)
22 – Dauphin (Region 5)
23 – Delaware (Region 8)
24 – Elk (Region 2)
25 – Erie (Region 2)
26 – Fayette (Region 1)
27 – Forest (Region 2)
28 – Franklin (Region 5)
29 – Fulton (Region 5)
30 – Greene (Region 1)
31 – Huntingdon (Region 5)
32 – Indiana (Region 3)
33 – Jefferson (Region 2)
34 – Juniata (Region 5)
35 – Lackawanna (Region 6)
36 – Lancaster (Region 5)
37 – Lawrence (Region 2)
38 – Lebanon (Region 5)
39 – Lehigh (Region 7)
40 – Luzerne (Region 6)
41 – Lycoming (Region 4)
42 – McKean (Region 2)
43 – Mercer (Region 2)
44 – Mifflin (Region 4)
45 – Monroe (Region 6)
46 – Montgomery (Region 8)
47 – Montour (Region 4)
48 – Northampton (Region 7)
49 – Northumberland (Region 4)
50 – Perry (Region 5)
51 – Philadelphia (Region 9)
52 – Pike (Region 6)
53 – Potter (Region 2)
54 – Schuylkill (Region 7)
55 – Snyder (Region 4)
56 – Somerset (Region 3)
57 – Sullivan (Region 6)
58 – Susquehanna (Region 6)
59 – Tioga (Region 4)
60 – Union (Region 4)
61 – Venango (Region 2)
62 – Warren (Region 2)
63 – Washington (Region 1)
64 – Wayne (Region 6)
65 – Westmoreland (Region 1)
66 – Wyoming (Region 6)
67 – York (Region 5)

Special Requests Public Inpatient File Layout and Supporting Documentation, 2003 - 2005



PA Health Care Cost Containment Council
Special Requests Unit

7

APPENDIX I – ACHD Emergency Room Visit Data Archival Survey (2005)

The Allegheny County Health Department (ACHD) Air Quality Program and Office of Epidemiology/Biostatistics are collaborating with the Department of Energy (DOE) and National Energy Technology Laboratory (NETL) on a project entitled "Design and Feasibility of a Retrospective Epidemiological Study of Coal-Fired Plant Emissions and Health Effects in the Pittsburgh, Pennsylvania Region". We are requesting information from all hospitals within the Pittsburgh greater metropolitan area to determine the type and extent of archived emergency room (ER) visit data that are readily accessible, particularly in electronic format in the 1999-2004 time period. **Please answer the following questions:**

***Facility name:*____________________ *Facility Contact:* ___________________*Phone*: ________________**

1. How are you currently archiving ER visit data collected from 1999-2004? (Please check)

Electronic

Hard copy only

Both electronic and hard copy

Unknown

2. If data are stored *electronically*, what format(s) are you currently using to archive your ER visit data (i.e., Oracle, SQL, etc). Format type: __________________________________

a,As of what year did you begin collecting ER data electronically? __________________

b. Please check all variables that are included in your electronic ER data archival system.

Facility identifier

Date of admission

Time of admission

Admitting point location (i.e. hospital, clinic, urgent center, etc)

Age

Date of birth

Ethnic group

Gender

Zip code

Admitting diagnosis (ICD-9/10 code)

Admitting diagnosis description (i.e. chest pain, shortness of breath, traumatic injury, etc)

Principal (discharge) diagnosis (ICD-9/10 code)

Secondary (discharge) diagnosis 1 (ICD-9/10 code)

Secondary (discharge) diagnosis 2 (ICD-9/10 code)

Secondary (discharge) diagnosis 3 (ICD-9/10 code)

Discharge status (admitted to hospital, transferred, discharged, deceased, etc.)

Discharge date

3. If archived in *hard-copy format,* where are the data housed? Please check:

In-house

Off-site

Other ____________________

a. How is the ER hard-copy data stored (i.e. chart format, Microfiche, etc.)? ______________

b. How far back do you maintain your ER hard-copy data? Year: __________

c. Do you intend to convert to electronic format? Yes ___ No ____ ***If yes***, when? Year ________

Thank you for taking the time to fill out this survey. Please return the survey in the enclosed envelope to the ACHD by **July 15, 2005** ***or*** if you prefer to fax the completed survey, please fax to Dr. Samuel Schlosberg, at 412-578-8325.

Version: May 2, 2005

APPENDIX J – MARS Databases

The Medical Archival Retrieval System (MARS) at the University of Pittsburgh Medical Center (UPMC) aggregates office visit and other data for physicians affiliated with UPMC. The available datasets are detailed in the following tables.

Table 65: UPMC MARS databases - (As of 07/01/05) - https://mars.mars-systems.com.

CLINICAL DATABASES	**Frequency**	**BMC**	**CHP**	**BRH**	**HHG**	**HHS**	**LEE**	**MCH**	**MWH**	**PAS**	**PUH**	**QST**	**REH**	**SHY**	**SMH**	**SSH**	**WPIC**
Autopsy Reports	hourly	Jul-00	Jun-99	Jul-99	Dec-98	Dec-98	Sep-00	May-99	Nov-01	Jul-04	Jan-81			Dec-97	Feb-99	Mar-97	
Cardiac Cath. Lab Reports	daily									Sep-04	Jan-90			Apr-04	Jan-05	Jun-04	
Case Log for Procedures in OR	daily			May-05				Jun-04	Jul-04		Jul-96				Jun-04	Apr-03	
Cytology Reports	hourly	Jul-00	Jun-99	Jul-99	Dec-98	Dec-98	Sep-00	May-99	Nov-01	Jul-04	Jan-81			Dec-97	Feb-99	Mar-97	
Discharge Summaries	hourly		Jun-00	Jan-01				Mar-01	Jun-05	Sep-04	Jan-90		Nov-03	Jan-98	Feb-00	Jun-03	Jan-92
Echocardiogram Reports	hourly									Sep-04	Apr-91 +					Jun-04	
Electroencephalogram/Electromyogram	daily			Nov-02							Jan-93						
Electrocardiogram	hourly							Oct-04			Feb-96			Aug-03			Feb-96
Emergency Room Notes	hourly			Jan-01 **						Nov-04	Jan-95				Feb-99		
EPIC Outpatient PhysicianProgress Notes	daily										Jul-00		Jul-00				
History & Physicals	hourly		Jun-00	Jan-01				Mar-01	Jun-05	Sep-04	Jan-90		Nov-03	Jan-98	Feb-00	Jun-03	Jan-92
Laboratory Results	every 30 minutes		Mar-92	Jul-99	Dec-98	Dec-98		May-99	Apr-05	Sep-05	Oct-90	May-00		Jan-99	Feb-99	Mar-97	Oct-90
Nuclear Cardiology Reports	daily									Sep-04	Jan-92					Jun-04	
Operative Notes	hourly		Jun-00	Jan-01				Mar-01	Jun-05	Sep-04	Jan-90			Jan-98	Feb-00	Jun-03	
PV Lab Reports	daily										1/1/2002 - 02/05			Sep-03			
Pharmacy Discharge Summaries	daily								Jun-04		Apr-92					Mar-97	Jan-92
Pharmacy Orders	hourly			Nov-03	Nov-03	Nov-03		Nov-03	Jun-04		Jan-97		Nov-03		Nov-03	Mar-97	Jan-97
Progress Notes	daily		Jun-00								Jan-90		Feb-04	Jan-98	Feb-00		Jan-96
Pulmonary Function Tests	hourly			Nov-02							Jan-98						
Radiology Reports	every 15 minutes		Jul-99	Jul-99	Dec-98	Dec-98		May-99	Nov-00 **	Sep-04	Jan-90			Dec-97	Feb-99	Mar-97	
Referral Letters	hourly										Jan-90				Feb-00		
Sleep Lab Tests	hourly							Sep-04			Mar-99				Mar-04		
Surgical Pathology Reports	hourly	Jul-00	Jun-99	Jul-99	Dec-98	Dec-98	Sep-00	May-99	Nov-01	Jul-04	Jan-81			Dec-97	Feb-99	Mar-97	
FINANCIAL DATABASES		BMC		BRH	HHG	HHS	LEE	MCH	MWH	PAS	PUH	QST	REH	SHY	SMH	SSH	WPIC
Inpatient Charges	daily			Jul-99	Dec-98	Dec-98		May-99		Aug-05	Jul-92		Oct-99	Jul-99	Feb-99	Mar-97	Jul-94
Medical Record Discharge Abstracts	monthly			Jul-99	Dec-98	Dec-98		May-99			Jul-92		Oct-99	Nov-01	Feb-99	Mar-97	Jul-98
Outpatient Charges	daily			Jul-99	Dec-98	Dec-98		May-99		Aug-05	Jul-92		Oct-99	Jul-99	Feb-99	Mar-97	Jul-94
Payment Transactions	daily			Jul-99	Dec-98	Dec-98		May-99			Jul-92		Oct-99	Nov-01	Feb-99	Mar-97	Jul-94
Ratio of Cost to Charge Table	annually			X	X	X		X			X		X	X	X	X	
AUXILIARY DATABASES		BMC		BRH	HHG	HHS	LEE	MCH	MWH	PAS	PUH	QST	REH	SHY	SMH	SSH	WPIC
Admission/Discharge/Transfer	hourly			X	X	X		X			X		X	X	X	X	

CLINICAL DATABASES	**Frequency**	**BMC**	**CHP**	**BRH**	**HHG**	**HHS**	**LEE**	**MCH**	**MWH**	**PAS**	**PUH**	**QST**	**REH**	**SHY**	**SMH**	**SSH**	**WPIC**
Charge Description Master	weekly			X	X	X		X			X		X	X	X	X	X
Data Dictionary- Map tables	monthly			X	X	X		X			X		X	X	X	X	X
Inpatient Census	hourly		X	X	X	X		X			X		X	X	X	X	X
Patient Demographics	daily			X	X	X		X			X		X	X	X	X	
Physician Information	weekly			X	X	X		X			X		X	X	X	X	X

Table 66: Hospital codes for Table 65.

UPMC Bedford	BMC	*Notes:	
UPMC Braddock	BRH		Date
Childrens	CHP		
UPMC Horizon - Greenville Campus	HHG		For SHY- Incomplete History and Physicals, Discharge Summaries, Operative Notes and Progress Notes from 7/99-9/99
UPMC Horizon - Shenango Campus	HHS		When historical data was loaded, variable data elements may exist within a data feed
UPMC Lee Regional	LEE		** not all sites within hospital provide report type
UPMC McKeesport	MCH		
Magee Womens Hospital	MWH		+ missing reports from Nov 11, 2002 - October 21, 2003
UPMC Presbyterian/Montefiore/EEI	PUH		
UPMC Passavant	PAS		
Quest Diagnostics	QST		
UPMC Rehabilitation Hospital	REH - merged with SSH 7/05		prepared by Melissa Saul
UPMC St. Margaret	SMH		
UPMC South Side	SSH		
Western Psychiatric Institute and Clinic	WPIC		

Table 67: MARS clinical databases.

Dataset name	**Description**	**Freq. Update**	Hospital*	Number of Reports
AUT	Autopsy records- includes final anatomical diagnosis, cause of death, summary of hospital course, anatomical descriptions	hourly	PUH, WPIC, SSH, SHY, HHG, HHS, SMH, BRH, MCH,MWH, LEE, BMC,PAS,CHP	*included in SP record count
CATH	Cardiac catherization lab records- includes indication, procedure type, medications used in procedure, and complications	daily	PUH,SHY,SMH, SSH,PAS	62,621
CYTO	Cytology records- includes specimen type, description and diagnosis	hourly	PUH, WPIC, SSH, SHY, HHG, HHS, SMH, BRH,MCH,MWH, LEE,BMC,PAS,CHP	*included in SP record count
DEROY	Completed OR case log - includes all procedures done in the operating room, includes surgeons, assistants, time-in, time-out, and description of operation	daily	PUH,SSH,MCH,MWH, SMH,BRH	366,056
DRUGS	Pharmacy real-time orders - includes dosage, frequency, route, and allergy information	hourly	PUH,WPIC,SSH,HHG, HHS,SMH,BRH,MCH, REH,MWH	18,765,388
DS	Discharge summaries – includes summary of present illness, discharge medications, discharge location	hourly	PUH,WPIC,SSH, SHY, SMH,MCH,BRH,REH, PAS,MWH,CHP	948,311
ECHO	Echocardiograph records- includes echocardiographic dimensions, clinical impression	hourly	PUH,SSH,PAS	116,115
EEG	Neurodiagnostics EEG records	daily	PUH,WPIC, SHY,SSH, MCH	32,306
EKG	Electrocardiogram records	hourly	PUH,WPIC,SMH,MCH,SHY	492,995
EMG	Neurodiagnostics EMG records	daily	PUH,WPIC,SHY	28,137
ER	Emergency room notes- includes chief complaint, past medical history and assessment and plan	hourly	PUH,WPIC,SSH,SHY, SMH,MCH,BRH,PAS, MWH,CHP	10,56,688
HP	History and Physical- includes history of present illness, social history, medications on admission, physical exam, clinical assessment and plan	hourly	PUH,WPIC,SSH,SHY, SMH,MCH,BRH,REH, PAS,MWH,CHP	1,136,870
LAB	Laboratory results- includes chemistry, hematology, microbiology, immunopathology, blood bank tests	hourly	PUH,WPIC,SSH, SHY,HHG, HHS, SMH,MCH,BRH,MWH, QST,CHP	50,685,565
LETT	Referral letters	hourly	PUH,WPIC,SSH,SHY	745,264
NUCLEAR	Nuclear cardiology study	daily	PUH,SSH	50,862
OP	Operative notes- includes title of procedure, event date, preoperative and postoperative diagnoses, OR description	hourly	PUH,SSH,SHY, SMH,MCH,BRH, PAS,MWH,CHP	1,088,238
PGN	Progress notes, includes outpatient clinic notes, ambulatory EMR summaries and inpatient physician progress notes	daily	PUH,WPIC,SSH,SHY SMH,MCH, BRH, REH,PAS,MWH,CHP	5,606,336
PSG	Polysomnogram Sleep Report	hourly	PUH,BRH	8,512
PULM	Pulmonary function tests- includes predicted spirometry and lung volumes, clinical impression	hourly	PUH,MCH,BRH	27407
RAD	Radiology records - includes diagnostic radiology, CT, MRI, PET, vascular studies, special procedures	every 15 minutes	PUH,WPIC, SSH,SHY, HHG, HHS, SMH, MCH, BRH,MWH,PAS,CHP	8,948,260

Dataset name	**Description**	**Freq. Update**	Hospital*	Number of Reports
SP	Surgical pathology record- includes date of event, specimen type, final diagnosis, and gross description	hourly	PUH,WPIC,SSH,SHY, HHG, HHS, SMH, MCH, BRH,MWH,LEE, BMC,PAS,CHP	206,0097

Table 68: MARS financial databases.

Dataset name	**Description**	**Freq. Update**	Hospital*	Number of Reports
CDM	Charge description master	weekly	PUH,WPIC,SHY, SSH, HHG, HHS, SMH,MCH,BRH, REH	2,517,377
CHGIN	Inpatient charges- includes patient name, account number, date of service, transaction code, quantity, charge amount, physician	daily	PUH,WPIC,SHY,SSH, HHG, HHS, SMH,MCH,BRH, REH	184,560,715
CHGOUT	Outpatient charges- includes same data as CHGIN	daily	PUH,WPIC,SHY,SSH, HHG, HHS, SMH,MCH,BRH, REH	66,793,018
MPAX	Medical record discharge abstracts - includes social security number, medical record number, hospital location, admission and discharge dates, date of birth, sex, race, marital status, financial class, current address, visit type, discharge disposition, attending physician, admitting service, admit source, admitting diagnosis, admitting chief complaint and onset date and time, primary diagnosis, final diagnoses (25 fields), procedures (25 fields), DRG	monthly	PUH,WPIC,SHY,SSH, HHG, HHS, SMH,MCH, BRH, REH	10,299,088
PAY	Patient payments- (see CHGIN; also includes payor name)	daily	PUH,WPIC,SHY,SSH, HHG,HHS,SMH,MCH, BRH, REH	33,198,515

Table 69: MARS auxiliary databases.

Dataset name	**Description**	**Freq. Update**	Hospital*
ADT	Admission, discharge, and transfer transactions	hourly	PUH, WPIC,SHY, SSH, HHG, HHS, SMH, MCH,BRH, REH
DOCS	Physician information, including office address, phone, and hospital privileges	weekly	PUH, WPIC, SSH, SHY, HHG, HHS, SMH, MCH, BRH,REH
CENSUS	Inpatient bed census, admit date, chief complaint, admitting diagnosis	hourly	PUH, WPIC, SSH, SHY, HHG, HHS, SMH, MCH,BRH,REH
MAP	Data Dictionary containing admission, discharge, and service types; census, medipac, and docs database fields; state, county, CPT, ICD-9, DRG, race, financial class and codes; UPMC Cost Centers codes	as needed	PUH, WPIC, SSH, SHY, HHG, HHS, SMH, MCH,BRH
PATIENTS	Patient demographic information, including address, phone, DOB, and emergency contact	daily	PUH, SSH, SHY, HHG, HHS, SMH,MCH, BRH, REH

Hospital codes for **Tables 67, 68, and 69**:

- PUH - Presbyterian/Montefiore/Eye and Ear Institute
- WPIC -Western Psychiatric Institute and Clinic
- SSH - South Side
- SHY - Shadyside
- HHS - Horizon(Shenango)
- HHG - Horizon(Greenville)
- SMH-St. Margaret's
- MCH – Mckeesport
- BRH – Braddock
- BMC – Bedford Memorial
- MWH – Magee Womens
- CHP – Childrens'
- REH- Rehabilitation Hospital(07/05 end)
- LEE – Lee(08/05 end)
- QST – Quest Diagnostics
- PAS – Passavant

APPENDIX K – General Health Plan Data Sharing Information

Contact Sheet for HMOs, PPOs and POS Plans for Pitt-PM Study: April 10, 2006

Name of Health Plan: ________________Initial Phone Number:________________

HMO/PPO or POS____

Contact person:_______________ Phone#:________________Email:____________

1. ***Year of plan inception (in Pittsburgh area) ______________________________***
2. ***Geographic coverage area ____***
3. ***Local enrollment (Pgh. SMA) _____________Total enrollment (PA)__________***
4. ***Is electronic data available? _________________Years Available____________***
5. ***Is there a specific category for unscheduled office visits in the electronic database?***

   ____
6. ***Is it possible to search on chief complaint or billing code for office visit to delineate cardiovascular or respiratory disease exacerbations?________________________***

7) Procedure for requesting access to data from you health plan for research purposes:

- ***g.** Application process (packet or written information via email or regular mail)*

  ____

- ***h.** What are the IRB requirements? Through University of Pittsburgh IRB or through external entity associated with the health plan?___________________*

i. ***Is there a cost to obtain the data or access to the data?*__________________**
***Approximate cost:*______________________________**

j. ***Is de-identified (aggregated) data available?*______________**

k. ***Can we obtain a HIPPA- compliant limited dataset to include dates of office visits, city/municipality and Zip Code of residence?*__________________________**

l. ***Can we obtain street addresses for geocoding if the Health Plan acts as an "honest broker"?* ______________________________**

<u>Comments or other information</u>

____
____
____
____

References

[Adamson et al., 2000] Adamson, I. Y., Prieditis, H., Hedgecock, C., and Vincent, R. (2000). Zinc is the toxic factor in the lung response to an atmospheric particulate sample. *Toxicology and Applied Pharmacology*, 166: 111–119.

[Allen et al., 1997] Allen, G., Sioutas, C., Koutrakis, P., Reiss, R., Lurmann, F. W., and Roberts, P. T. (1997). Evaluation of the TEOM method for measurement of ambient particulate mass in urban areas. *Journal of the Air & Waste Management Association (1995)*, 47(6): 682–9.

[Arena et al., 2006] Arena, V., Mazumdar, S., Zborowski, J., Talbott, E., He, S., Chuang, Y.-H., and Schwerha, J. (2006). Allegheny County air pollution study (ACAPS): A retrospective investigation of cardiopulmonary outcomes and exposure to air pollutants. *Journal of Occupational and Environmental Medicine*, 48(1): 38–47.

[Basu et al., 2005] Basu, R., Dominici, F., and Samet, J. M. (2005). Temperature and mortality among the elderly in the united states: a comparison of epidemiologic methods. *Epidemiology (Cambridge, Mass.)*, 16(1): 58–66.

[Bateson and Schwartz, 1999] Bateson, T. F. and Schwartz, J. (1999). Control for seasonal variation and time trend in case-crossover studies of acute effects of environmental exposures. *Epidemiology (Cambridge, Mass.)*, 10(5): 539–44.

[Bland and Altman, 1986] Bland, J. M. and Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1: 307–310.

[Carter et al., 1997] Carter, J. D., Ghio, A. J., Samet, J. M., and Devlin, R. B. (1997). Cytokine production by human airway epithelial cells after exposure to an after exposure to an air pollution particle is metal-dependent. *Toxicology and Applied Pharmacology*, 146: 180–188.

[Chock et al., 2000] Chock, D. P., Winkler, S. L., and Chen, C. (2000). A study of the association between daily mortality and ambient air pollutant concentrations in Pittsburgh, Pennsylvania. *Journal of Air & Waste Management Association*, 50: 1481–1500.

[Chow and Watson, 1998] Chow, J. C. and Watson, J. G. (1998). Guideline on speciated particulate monitoring. *Desert Research Institute, Reno, NV. Prepared for US EPA, Research Triangle Park, NC.*

[Chow et al., 2004] Chow, J. C., Watson, J. G., Chen, L. W. A., Arnott, W. P., Moosmüller, H., and Fung, K. (2004). Equivalence of elemental carbon by thermal/optical reflectance and transmittance with different temperature protocols. *Environmental Science & Technology*, 38(16): 4414–22.

[Christensen and Sain, 2002] Christensen, W. F. and Sain, S. R. (2002). Accounting for dependence in a flexible multivariate receptor model. *Technometrics*, 44(4): 328–337.

[Connell et al., 2005a] Connell, D. P., Winter, S. E., and Conrad, V. B. (2005a). Characterizaiton of $PM_{2.5}$ trace elements in Steubenville, Ohio, using dynamic reaction cell ICP-MS. In *Proceedings of Air Quality V, September 19-21*, Arlington, VA.

[Connell et al., 2005b] Connell, D. P., Withum, J. A., Winter, S. E., Statnick, R. M., and Bilonick, R. A. (2005b). The Steubenville comprehensive air monitoring program (SCAMP): Overview and statistical considerations. *Journal of the Air & Waste Management Association*, 55(4): 467–480.

[Connell et al., 2005c] Connell, D. P., Withum, J. A., Winter, S. E., Statnick, R. M., and Bilonick, R. A. (2005c). The Steubenville Comprehensive Air Monitoring Program (SCAMP): Analysis of short-term and episodic variations in $PM_{2.5}$ concentrations using hourly air monitoring data. *Journal of the Air & Waste Management Association*, 55(5): 559–73.

[Coutant and Stetzer, 2001] Coutant, B. and Stetzer, S. (2001). Evaluation of $PM_{2.5}$ speciation sampler performance and related sample collection and stability issues. Technical Report EPA-454/R-01-008, U. S. Environmental Protection Agency, Research Triangle Park, NC.

[Cressie and Huang, 1999] Cressie, N. and Huang, H. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94: 1330–1340.

[Davis et al., 2003] Davis, R. A., Dunsmuir, W. T. M., and Streett, S. B. (2003). Observation driven models for poisson counts. *Biometrika*, 90: 777–790.

[Davis et al., 2005] Davis, R. A., Dunsmuir, W. T. M., and Streett, S. B. (2005). Maximum likelihood estimation for an observation driven model for Poisson counts. *Methodology, Computing and Applied Probability*, 7(2): 149–159.

[Davis et al., 1999] Davis, R. A., Dunsmuir, W. T. M., and Wang, Y. (1999). Modelling time series of count data. In Ghosh, S., editor, *Asymptotics, Nonparametrics and Time Series*, pages 63–114. Marcel Decker.

[Davison and Hinkley, 1997] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

[De Cesare et al., 2001a] De Cesare, L., Myers, D. E., and Posa, D. (2001a). Estimating and modelling space-time correlation structures. *Statistical Probability Letters*, 51(1): 9–14.

[De Cesare et al., 2001b] De Cesare, L., Myers, D. E., and Posa, D. (2001b). Product-sum covariance for space-time modeling: an environmental application. *Environmetrics*, 12: 11–23.

[D'Ippoliti et al., 2003] D'Ippoliti, D., Forastiere, F., Ancona, C., Agabiti, N., Fusco, D., Michelozzi, P., and Perucci, C. A. (2003). Air pollution and myocardial infarction in rome: A case-crossover analysis. *Epidemiology (Cambridge, Mass.)*, 14(5): 528–35.

[Drescher, 2005] Drescher, D. (2005). Alternative distributions for observation driven count series models. http: //opus.zbw-kiel.de/volltexte/2005/3197/pdf/EWP-2005-11.pdf.

[Drescher, 2006] Drescher, D. (2006). R code for estimating GLARMA model parameters. Personal Communication.

[Eatough et al., 1993] Eatough, D. J., Wadsworth, A., Eatough, D. A., Crawford, J. W., Hansen, L. D., and Lewis, E. A. (1993). A multiple-system, multi-channel diffusion denuder sampler for the determination of fine-particulate organic material in the atmosphere. *Atmospheric Environment; Part A, General Topics*, 27(8): 1213–1219.

[Frank, 2006] Frank, N. H. (2006). Retained nitrate, hydrated sulfates, and carbonaceous mass in federal reference method fine particulate matter for six eastern u.s. cities. *Journal of the Air & Waste Management Association (1995)*, 56(4): 500–11.

[Fung et al., 2003] Fung, K. Y., Krewski, D., Chen, Y., Burnett, R., and Cakmak, S. (2003). Comparison of time series and case-crossover analyses of air pollution and hospital admission data. *International Journal of Epidemiology*, 32(6): 1064–1070.

[Greenland, 1996] Greenland, S. (1996). Confounding and exposure trends in case-crossover and case-time-control designs. *Epidemiology (Cambridge, Mass.)*, 7(3): 231–9.

[Gupta et al., 2004] Gupta, D., Saul, M., and Gilbertson, J. (2004). Evaluation of a deidentification (de-id) software engine to share pathology reports and clinical documents for research. *American Journal of Clinical Pathology*, 121(2): 176–86.

[Jaech, 1985] Jaech, J. L. (1985). *Statistical Analysis of Measurement Errors*. John Wiley & Sons, New York.

[Jansen et al., 2002] Jansen, J. J., Edgerton, E. S., Hansen, D. A., and Hartsell, B. E. (2002). Sampling artifacts in the federal reference method for $PM_{2.5}$. In *Proceedings of the International Conference on Air Quality III*.

[Kim et al., 2005] Kim, E., Hopke, P. K., and Qin, Y. (2005). Estimation of organic carbon blank values and error structures of the speciation trends network data for source apportionment. *Journal of the Air & Waste Management Association (1995)*, 55(8): 1190–9.

[Kline, 1998] Kline, R. B. (1998). *Structural Equation Modeling*. Guilford Press, New York.

[Künzli et al., 2001] Künzli, N., Medina, S., Kaiser, R., Quénel, P., Horak, F., and Studnicka, M. (2001). Assessment of deaths attributable to air pollution: Should we use risk estimates based on time series or on cohort studies? *American Journal of Epidemiology*, 153(11): 1050–5.

[Kwon et al., 2001] Kwon, H. J., Cho, S. H., Nyberg, F., and Pershagen, G. (2001). Effects of ambient air pollution on daily mortality in a cohort of patients with congestive heart failure. *Epidemiology (Cambridge, Mass.)*, 12(4): 413–9.

[Lahiri, 2003] Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer Series in Statistics. Springer-Verlag, New York.

[Lipfert and Wyzga, 1997] Lipfert, F. W. and Wyzga, R. E. (1997). Air pollution and mortality: The implications of uncertainties in regression modeling and exposure measurement. *Journal of the Air & Waste Management Association*, 47: 517–523.

[Lipfert et al., 2006] Lipfert, F. W., Wyzga, R. E., Baty, J. D., and Miller, J. P. (2006). Traffic density as a surrogate measure of environmental exposures in studies of air pollution health effects: Long-term mortality in a cohort of US veterans. *Atmospheric Environment*, 40: 154–169.

[Lu and Zeger, 2006] Lu, Y. and Zeger, S. L. (2006). On the equivalence of case-crossover and time series methods in environmental epidemiology. *Biostatistics*.

[Maclure, 1991] Maclure, M. (1991). The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*, 133: 144–153.

[Mar et al., 2000] Mar, T. F., Norris, G. A., Koenig, J. Q., and Larson, T. V. (2000). Associations between air pollution and mortality in Phoenix. *Environmental Health Perspectives*, 108: 347–353.

[Maranche, 2006] Maranche, J. (2006). Allegheny County $PM_{2.5}$ source apportionment results using the positive matrix factorization model (PMF Version 1.1), model timeframe: July 2003 through August 2005: Air quality program report. Technical report, Allegheny County Health Department.

[Marcouilides and Moustaki, 2002] Marcouilides, G. A. and Moustaki, I. (2002). *Latent Variable and Latent Structure Models*. Lawrence Erlbaum Associates, London.

[Martello et al., 2006] Martello, D. V., Pekney, N. J., Anderson, R. R., Davidson, C. I., Hopke, P. K., Kim, E., Christensen, W. F., Mangelson, N. F., and Eatough, D. J. (2006). Apportionment of ambient primary and secondary $PM_{2.5}$ at the NETL Pittsburgh PM characterization site using PMF and PSCF. *Journal of Air & Waste Management Association*. Submitted.

[Modey and Eatough, 2004] Modey, W. and Eatough, D. J. (2004). Twenty four-hour PC-BOSS air-monitoring results from the NETL fine-particulate sampling site in Pittsburgh, Pennsylvania: An annual perspective. *Aerosol Science and Technology*, 38(3): 194–204.

[Moolgavkar, 2003] Moolgavkar, S. H. (2003). Air pollution and daily mortality in two U.S. counties: Season-specific analyses and exposure-response relationships. *Inhalation Toxicology*, 15: 877–907.

[Moolgavkar and Luebeck, 1996a] Moolgavkar, S. H. and Luebeck, E. G. (1996a). A critical review of the evidence on particulate air pollution and mortality. *Epidemiology (Cambridge, Mass.)*, 7(4): 420–8.

[Moolgavkar and Luebeck, 1996b] Moolgavkar, S. H. and Luebeck, E. G. (1996b). Zinc is the toxic factor in the lung response to an atmospheric particulate sample. *Epidemiology*, 7: 420–428.

[Navidi, 1998] Navidi, W. (1998). Bidirectional case-crossover designs for exposures with time trends. *Biometrics*, 54: 596–605.

[Neale et al., 2003] Neale, M. C., Boker, S. M., Xie, G., and Maes, H. H. (2003). *Mx: Statistical Modeling*. Department of Psychiatry, Box 900126, VCU, Richmond, VA 23298, sixth edition.

[Neas et al., 1999] Neas, L., Schwartz, J., and Dockery, D. (1999). A case-crossover analysis of air pollution and mortality in Philadelphia. *Environmental Health Perspectives*, 107(8): 629–631.

[Ostro et al., 2000] Ostro, B. D., Broadwin, R., and Lipsett, M. J. (2000). Coarse and fine particles and daily mortality in the Coachella Valley, CA: A follow-up study. *Journal of Exposure Analysis and Environmental Epidemiology*, 10: 412–419.

[Pekney and Davidson, 2005] Pekney, N. and Davidson, C. (2005). Determination of trace elements in ambient aerosol samples. *Analytica Chimica Acta*, 540(2): 269–277.

[Pekney et al., 2005] Pekney, N. J., Davidson, C. I., Zhou, L., and Hopke, P. K. (2005). Identifying sources of $PM_{2.5}$ in Pittsburgh using PMF and PSCF. Presented at the AAAR Specialty Conference: Particulate Matter, Supersites Program & Related Studies.

[Peng,] Peng, R. D. NMMAPSdata R Package. http: //www.ihapss.jhsph.edu/data/NMMAPS/R/.

[Peng et al., 2004] Peng, R. D., Welty, L. J., and McDermott, A. (2004). The national morbidity, mortality, and air pollution study database in r. Technical Report Working Paper 44, Johns Hopkins University, Dept. of Biostatistics Working Papers.

[Pope, 1999] Pope, C. A. (1999). Mortality and air pollution: associations persist with continued advances in research methodology. *Environmental Health Perspectives*, 107(8): 613–4.

[Pun and Seigneur, 2004] Pun, B. and Seigneur, C. (2004). Creation of an air pollutant database for health effects studies. Technical Report CP190-04-03.

[R Development Core Team, 2006] R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Ripley and Thompson, 1987] Ripley, B. D. and Thompson, M. (1987). Regression techniques for the detection of analytical bias. *Analyst*, 112: 377–383.

[Schauer et al., 2003] Schauer, J. J., Mader, B. T., Deminter, J. T., Heidemann, G., Bae, M. S., Seinfeld, J. H., Flagan, R. C., Cary, R. A., Smith, D., Huebert, B. J., Bertram, T., Howell, S., Kline, J. T., Quinn, P., Bates, T., Turpin, B., Lim, H. J., Yu, J. Z., Yang, H., and Keywood, M. D. (2003). ACE-Asia intercomparison of a thermal-optical method for the determination of particle-phase organic and elemental carbon. *Environmental Science & Technology*, 37(5): 993–1001.

[Schmid et al., 2001] Schmid, H., Laskus, L., Abraham, H. J., Baltensperger, U., Lavanchy, V., Bizjak, M., Burba, P., Cachier, H., Crow, D., Chow, J., Gnauk, T., Even, A., Brink, H. M. T., Giesen, K., Hitzenberger, R., Hueglin, C., Maenhaut, W., Pio, C., Carvalho, A., Putaud, J., Toom-Sauntry, D., and Puxbaum, H. (2001). Results of the "Carbon Conference" International Aerosol Carbon Round Robin Test Stage I. *Atmospheric Environment*, 35(12): 2111–2121.

[Schwartz et al., 1996] Schwartz, J., Dockery, D., and Neas, L. (1996). Is daily mortality associated specifically with fine particles? *Journal of Air & Waste Management Association*, 46(10): 927–939.

[Schwartz and Dockery, 1992a] Schwartz, J. and Dockery, D. W. (1992a). Increased mortality in philadelphia associated with daily air pollution concentrations. *American Review of Respiratory Disease*, 145(3): 600–4.

[Schwartz and Dockery, 1992b] Schwartz, J. and Dockery, D. W. (1992b). Particulate air pollution and daily mortality in Steubenville, Ohio. *American Journal of Epidemiology*, 135(1): 12–9; discussion 20–5.

[Suarez and Ondov, 2002] Suarez, A. and Ondov, J. (2002). Ambient aerosol concentrations of elements resolved by size and source: Contributions of some cytokine-active metals from coal- and oil-fired power plants. *Energy & Fuels*, 16: 562–568.

[Sunyer et al., 2000] Sunyer, J., Schwartz, J., Tobias, A., Macfarlane, D., Garcia, J., and Anto, J. M. (2000). Patients with chronic obstructive pulmonary disease are at increased risk of death associated with urban particle air pollution: A case-crossover analysis. *American Journal of Epidemiology*, 151(1): 50–6.

[Tsai et al., 2006] Tsai, S.-S., Cheng, M.-H., Chiu, H.-F., Wu, T.-N., and Yang, C.-Y. (2006). Air pollution and hospital admissions for asthma in a tropical city: Kaohsiung, Taiwan. *Inhalation Toxicology*, 18(8): 549–54.

[Villeneuve et al., 2003] Villeneuve, P. J., Burnett, R. T., Shi, Y., Krewski, D., Goldberg, M. S., Hertzman, C., Chen, Y., and Brook, J. (2003). Time-series study of air pollution, socioeconomic status, and mortality in Vancouver, Canada. *Journal of Exposure Analysis and Environmental Epidemiology*, 13: 427–435.

[Vineis and Husgafvel-Pursiainen, 2005] Vineis, P. and Husgafvel-Pursiainen, K. (2005). Air pollution and cancer: biomarker studies in human populations. *Carcinogenesis*, 26(11): 1846–55.

[Wade et al., 2006] Wade, K. S., Mulholland, J. A., Marmur, A., Russell, A. G., Hartsell, B., Edgerton, E., Klein, M., Waller, L., Peel, J. L., and Tolbert, P. E. (2006). Effects of instrument precision and spatial variability on the assessment of the temporal variation of ambient air pollution in Atlanta, Georgia. *Journal of the Air & Waste Management Association (1995)*, 56(6): 876–88.

[Zanobetti and Schwartz, 2005] Zanobetti, A. and Schwartz, J. (2005). The effect of particulate air pollution on emergency admissions for myocardial infarction: A multicity case-crossover analysis. *Environmental Health Perspectives*, 113(8): 978–82.

[Zeger et al., 2000] Zeger, S. L., Thomas, D., Dominici, F., Samet, J. M., Schwartz, J., Dockery, D., and Cohen, A. (2000). Exposure measurement error in time-series studies of air pollution: Concepts and consequences. *Environmental Health Perspectives*, 108(5): 419–26.

[Zelikoff et al., 2002] Zelikoff, J. T., Schermerhorn, K. R., Fang, K., Cohen, M., and Schlesinger, R. (2002). A role for associated transition metals in the immunotoxicity of inhaled ambient particulate matter. *Environmenatl Health Perspectives*, 110 suppl. 5: 871–875.