Assessment of the Contribution to Personal Exposures of Air Toxics from Mobile Sources

United States
Environmental Protection
Agency

Assessment of the Contribution to Personal Exposures of Air Toxics from Mobile Sources

Assessment and Standards Division
Office of Transportation and Air Quality
U.S. Environmental Protection Agency

Prepared for EPA by

Clifford P. Weisel, Ph.D
Environmental & Occupational Health Sciences Institute
Robert Wood Johnson Medical School
University of Medicine and Dentistry of New Jersey

EPA Contract No. 68-C-03-149

NOTICE

This technical report does not necessarily represent final EPA decisions or positions. It is intended to present technical analysis of issues using data that are currently available. The purpose in the release of such reports is to facilitate the exchange of technical information and to inform the public of technical developments which may form the basis for a final EPA decision, position, or regulatory action.

United States
Environmental Protection
Agency

EPA420-R-05-025
December 2005

Executive Summary:

To evaluate the role of proximity to mobile source emissions on ambient air surrounding residences, statistical analyses using linear regression models were conducted for selected volatile organic compounds, carbonyls, PM2.5 mass, elemental carbon and organic carbon with mobile emission sources. The log transformed ambient air concentration of individual air toxics measured in Elizabeth, NJ during the Relationship of Indoor, Outdoor and Personal Air (RIOPA) study was used as the dependent variable and inverse distance to roadways, gas stations, and point sources and meteorological parameters as the independent variables in the regression models. The home, roadway, point and area sources in and around Elizabeth, NJ were geocoded using Geographic Information System (GIS) techniques to determine the distance between the homes and potential ambient sources. Meteorological data (wind speed, wind direction, temperature, and atmospheric pressure) were obtained from the NOAA, Weather-Bureau-Army-Navy (WBAN) station in the Newark Liberty International Airport, which is immediately to the north of Elizabeth, and mixing height data from Brookhaven, NY (the closest station to Elizabeth containing that type of data). The meteorological data were averaged over the 48 hour sampling period to provide a single value for each sample. The roads were stratified into six roadway types based on categories used in the EPA Mobile 6 model. Quality assurance steps were taken to confirm the location and each home and location, including direct visits to Elizabeth to verify the address and coordinates. Various regression models (and selection criteria) were used to confirm that repeatable set of associations were obtained.

All target aromatic compounds (benzene, toluene, ethyl benzene, *m,p* xylene, *o* xylene), methyl *tert* butyl ether, PM2.5, and organic carbon were statistically associated with the inverse distance to urban major arterials (FC14) or the interstate highway (FC11); methyl *tert* butyl ether (MTBE), benzene, *m,p* xylene, and *o* xylene were statistically associated with the inverse distance to gasoline stations; the carbonyl compounds (acetaldehyde, acrolein, and formaldehyde) were not associated with the inverse distance to roadways; PM2.5 and elemental carbon were associated with area sources of diesel emissions based on truck or bus depot and idling activity, two PAH compounds (coreonene-gasoline emissions and benzo[ghi]perylene-mobile emissions) were statistically associated with the inverse distance FC11 and PM area sources. Two volatile compounds and two PM constitutes without mobile sources (carbon tetrachloride, tetrachloroethylene, sulfur and selenium), were examined as controls to check for spurious associations, were not associated with distance to roadways.

The regression model had overall r^2 of between 0.16 and 0.67, indicating that between approximate 20% and 70% of the variability in the air concentrations was explained by the model. However, the partial r^2 of the distance terms were less than 10%, as meteorology was a more important factor on controlling the variations in the concentrations than the distance between the home and a mobile emission source. The effect of mobile sources emissions appears to be confined to residences very close to the sources within 200 meters, though within that distance that can cause in several $\mu g/m^3$, dependent upon the sources strength for that compound. Thus, for most homes in Elizabeth, NJ the influence of mobile sources is to raise the general background levels of the compounds emitted with the increase dependent upon the meteorological condition,

especially the atmospheric stability, and there are appears to only be small increases in concentrations as the distance decreases. For homes within very close proximity (no more than several hundred meters) of gas stations and highly trafficked roadways the regression model predict changes in the median ambient air concentration around the homes from the typical background levels by 2 to 10μg/m^3.

BACKGROUND

The Relationship between Indoor, Outdoor, and Personal Air (RIOPA) study was undertaken to determine the influence of outdoor sources on indoor and personal air concentrations of a set of volatile organic compounds (VOCs), aldehydes, and $PM_{2.5}$ mass. Indoor/outdoor polyaromatic hydrocarbons (PAHs), and elemental carbon/organic carbon (EC/OC) concentrations were also measured (Weisel et al 2004a,b). The study collected data on indoor, outdoor, and personal air concentrations for approximately 300 non-smoking homes in Los Angeles (CA), Elizabeth (NJ), and Houston (TX), visited twice from the summer of 1999 to the spring of 2001. Either one or two homes was visited on a single day, though some days had samples collected from three or four homes. Samples were collected throughout the year. This report focuses on the analysis of VOCs, carbonyls and $PM_{2.5}$ associated with mobile source emissions and air samples collected outside residence in Elizabeth, NJ. One dominant source in Elizabeth NJ for aromatic hydrocarbons and $PM_{2.5}$ is mobile sources. Prior to examining the ambient source contributions to indoor/personal VOCs, the association between source emission and ambient VOC concentrations near residences should be established. One approach to this is to evaluate the role of proximity to the potential emission sources and meteorological conditions on ambient concentrations. Precise proximity information between the residences where the samples were collected and potential emission sources are needed along with locally collected meteorological information to evaluate the effect of proximity and meteorological conditions. All roadway classes bisect the city of Elizabeth, NJ, so wide distributions of distances to each roadway type and gasoline stations exist. The home selection criteria included over-sampling homes close to heavily

trafficked roadways and being near gasoline stations. The association of air concentration and proximity to mobile sources was examined by deriving linear regression equations using air concentration as the dependent variable and proximity to roadways, gasoline stations, and point sources and meteorological parameters as the independent variables. Attempts to see such statistical association have met with only minimal success unless the locations of the homes were very close to the roadways. The RIOPA dataset was designed to contain a substantial number of homes within 0.5km of mobile sources thereby allowing for examination of the effect of proximity to mobile source emissions in a northeast urban environment, Elizabeth, NJ.

METHODOLOGY

Construction of the RIOPA Database

The major components in the RIOPA database were sample information, analysis results, and questionnaire responses. The database was implemented in Microsoft Access 97® and upgraded in Microsoft Access 2000®. A decomposition process was used to remove internal duplication in a series of steps without loss of data. Every tabular record was indexed with a unique data-independent primary key. The unique, data-independent primary key enables the linking, indexing, filtering and sorting of records in multiple tables and their components. A normalization process was used to re-organized the data into a streamlined effective tabular structure. For decomposition and normalization, the Access commands 'selection query' and 'make table query' were most frequently used. To find the repetition of the identical record in a table, the Access commands 'find the duplicate query' was used, while 'find unmatched query' was used to determine when

there was the missing data. Establishing relationships between one table and another table by assigning a unique primary key such as identification field was mandatory for the database performance.

Each sampling home and sample was assigned a unique identification number (ID) prior to collecting the sample. Each unique sample number was linked to the home ID so the samples associated with each home could be identified. The home ID was coded to identify the state the sample was collected in using the two letter state abbreviations (CA, TX, and NJ), followed by a three-digit number unique for each home in that state. Among the three digit numbers, the first digit represented whether the visit was the first or a repeat visit (1, 2 respectively), and the second and third digits the chronological order the house was selected in (00~99). A unique five digit sequential number was assigned to each sample as the sample identifier. The first digit of sample ID was reserved to identify sample types while the remaining numbers randomly assign so that the analyst could not determine where the sample came from nor the sample type (indoor, outdoor, personal, blank, duplicate) prior to analysis. The descriptions of the data fields contained in the RIOPA database are listed in Tables 1 to 3.

Quality Assurance of Database

The following quality assurance protocols were followed at each data entry and modification step to find data entry errors and repeated or missing data. All the sampling information, analysis results and questionnaire data were transferred into the database. Quality assurance at the data entry level was performed by having an individual who did not enter the data compare the original written sampling records to the electronic data files. Validation equations were used in Access Query to identify potential data entry

errors, especially for the fields containing calculated values. Access commands were used to find duplicate data entries ("find the duplicate query"), which were deleted from the database and missing data across different tables, ("find the unmatched query").

All detailed information concerning the sample collection, sample analysis and questionnaires were consolidated and compiled into the main database in an organized manner as illustrated in Figure 1 (adopted from Weisel et al. 2004b, RIOPA final report). The final database was reviewed by research associates, experienced in analyzing each specific type of sample. This review included cross checking keyed data entries against the original printed hard copy of the analytical data. The research associate double-checked all the calculations used to transform the analytical data into the reported ambient air concentrations. Finalized data were confirmed by reapplying all of the calculations to the original analytical data. After the research associate completed his or her verification, the initial database was then classified as the preliminary database.

The field teams validated the preliminary database by reviewing the field sampling information and confirming the calculations that incorporated the information from the field sampling sheets. The field teams then made any necessary corrections and noted the change, which was then reported back to the originator for further confirmation of the needed correction. After the field teams made their comments and corrections, the principal investigators randomly checked the data by cross-referencing the electronic data for a subset of samples with the respective original data from the analytical results or sampling information sheets.

Table 1. Components and Data Fields of the Sampling Information in the RIOPA Database

Data Fields	Description
Home ID	Unique identification number with state abbreviation
Source ID, location	PFT source ID number (alpha-numeric) and location (floor-room)
CAT ID	Capillary absorbent tube ID (numeric)
Sample ID	Unique 5 digit number linked to the house ID, identifying contaminant category measured (VOC, DNPH, DNSH, Teflon and Quartz filter for $PM_{2.5}$, PUF)
Sampling date, time	Date (mm/dd/yy) and time (hh:mm) sampling started and ended
Sample duration	Calculated duration of sampling in minutes
Flow rate	Initial, final and average flow rate of pump (cc/min, or L/min)
Sample volume	Calculated volume of sample (L, or m^3)
Pump elapsed time	Pump elapsed time recorded on the pump counter in minutes
Pump recorded volume	Pump recorded volume of air sampled (m^3)
Sample type	Sample type (indoor, outdoor, personal adult, child, duplicate, blank, control)
Equipment ID	Pump, head and battery IDs
Leak test	Leak test check done before and after sampling (yes/no)

Table 2. Components and Data Fields of the Information of the Analysis Results in the RIOPA Database

Data Fields	Description
VOCs	Concentration (ppb, µg/m^3) of 1,3-butadiene, methylene chloride, chloroprene, methyl *tert* butyl ether, carbon tetrachloride, chloroform, benzene, *m,p*-xylene, toluene, trichloroethylene, tetrachloroethylene, ethylbenzene, *o*-xylene, styrene, µ-pinene, µ-pinene, *d*-limonene, 1,4-dichlorobenzene
Carbonyls	Concentration (ppb, µg/m^3) of formaldehyde, acetaldehyde, acetone, acrolein, propionaldehyde, crotonaldehyde, benzaldehyde, hexaldehyde, glyoxal, methylglyoxal
$PM_{2.5}$	$PM_{2.5}$ mass, Concentration (ppb, µg/m^3) of organic carbon (OC) and elemental carbon (EC), elements; Ag, Al, As, Ba, Be, Bi, Br, Ca, Cd, Cl, Co, Cr, Cs, Cu, Fe, Ga, Ge, Hg, In, K, La, Mn, Mo, Ni, P, Pb, Pd, Rb, S, Sb, Se, Si, Sn, Sr, Ti, Tl, U, V, Y, Zn, Zr
PAHs	Concentration (ppb, µg/m^3) of gas/ particle phase polycyclic aromatic hydrocarbons; Dibenzothiophene, Phenanthrene, Anthracene, 2-Methylanthracene, 1-Methylanthracene, 1-Methylphenanthrene, 9-Methylanthracene, 4,5-Methylenephenanthrene, 3,6-Dimethylphenanthrene, 9,10-Dimethylanthracene, Fluoranthene, Pyrene, Benzo[a]fluorene, Retene, Benzo[b]fluorene, Cyclopenta[c,d]pyrene, Benzo[a]anthracene, Chrysene+Triphenylene, Benzo[b]naphtho[2,1-d]thiophene, Benzo[b+k]fluoranthene, Benzo[e]pyrene, Benzo[a]pyrene, Perylene, Indeno[1,2,3-c,d]pyrene, Dibenzo[a,c+a,h]anthracene, Benzo[g,h,i]perylene, Coronene
House Information	Air exchange rate (1/hr) and the volume of house (m^3)
Meteorological Information	Temperature and relative humidity measured inside and outside of house

Table 3. Components and Data Fields of the Questionnaire Data in the RIOPA Database

Data Fields	Description
Technician Walkthrough	Evaluation of the house and its usage and a description of the neighborhood regarding possible sources.
Baseline Survey	Household and participant characteristics; demographics and socioeconomic status; housing characteristics, facilities and usage; personal exposure activities before the study period; and respiratory health status of participant
Activity Questionnaire	A detailed series of questions related to activities, duration and use of consumer products
Time Diary	48-hour activity log listing the time spent in each microenvironment

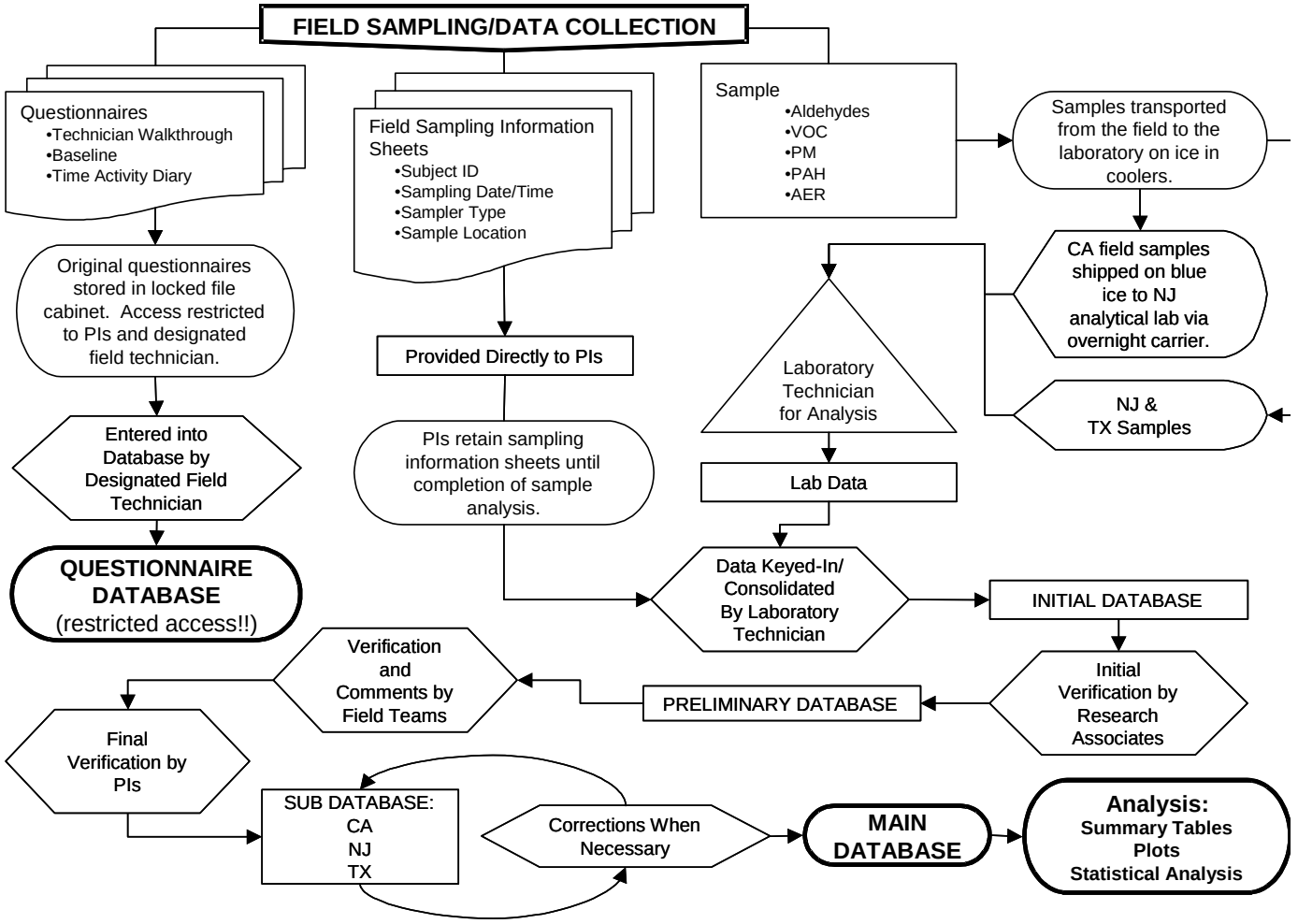

Figure 1. The Flow Diagram of the Transference of Information from the Field Sampling to Database Construction and the Quality Assurance Processes (adopted from Final Report of the RIOPA study)

Data Integration in the RIOPA Database

To expand the utility of the RIOPA database and to facilitate data analysis with meteorological and geographical datasets, different databases in the public domain were either imported into or linked to the RIOPA database. The details of the integration of the databases are illustrated in Figure 2. The databases included were the National Emission Inventory of 1999 (version 3.0 final for HAPs and criteria pollutants, US EPA), National Climatological data obtained from the National Oceanographic and Atmospheric Administration (NOAA), 2000 US Census data, 2000 TIGER/Line data, and Roadway Information & Transportation data obtained from NJ DOT (Table 4)

National Emission Inventory of 1999

The emission data of the states of New Jersey and New York from mobile, area, and point sources were obtained from the 1999 National Emission Inventory (NEI, the final version 3.0 for the hazardous air pollutants, released on Dec 2003; the final version 3.0 for the criteria pollutants, released on Feb 2004). The datasets were divided into four categories (On-road, Non-road, Point, Non-Point) and available from the Technology Transfer Network, Clearinghouse for Inventories and Emission Factors (TTN CHIEF, http://www.epa.gov/ttn/chief/net/1999 inventory.html). The emission sources of compounds collected in the RIOPA study were selected from the inventory datasets of the counties containing or adjacent to the RIOPA study area. The counties were Union, Essex, and Hudson Counties, New Jersey, and Richmond County, New York.

The emissions from on road mobile sources were calculated to evaluate which road types to consider in the regression models. Actual emission rates were not used as

inputs in the models since only statistical associations were examined in this analysis and not a comparison of predicted to measured concentrations. The emissions were calculated by multiplying emission factors (g/mile) estimated by US EPA using MOBILE 6.2 model and vehicle miles traveled (VMT, 10^6 miles). The VMT were estimated from the sampled traffic counts of road segments by Federal Highway Administration (FHWA)'s Highway Statistics 1999 (US EPA, Documents for NEI, 2003). The emission estimates for each county were stratified by road types (6 urban categories of public roads were present in Union County) and by twelve vehicle types. The emission rate per unit length of public road by functional classification was estimated from the total roadway mileages of Union County and the annual total emission from on-road sources in Union County. The emission rates of selected VOCs by roadway class are listed in Table 5. The emission rates calculated from the major roadways (FC11, urban interstate highways; FC12, urban other freeways and expressways; FC14, urban major arterials) were more than 6 to 90 fold higher than the emission rates from the minor classes of roadways (FC16, urban minor arterial; FC17, urban collector; and FC19, urban local). To apportion the annual total amount of emissions from the on-road mobile sources countywide to Elizabeth, the ratio of the roadway mileage in Elizabeth to the roadway mileage in Union County was calculated for each category of functional classification. The public roadway mileages in Elizabeth were 11.5, 3.7, and 22.3 in kilometers for FC11, FC12, and FC14, respectively. The percentage of the major public roadway miles inElizabeth classified as FC11, FC12 and FC14 were 3.5%, 11% and 6.7%, respectively. The proportion of major roadway miles in Elizabeth were larger than that in Union County as a whole (FC11, 1.4%; FC14, 3.6%), in New York Northeast New Jersey (FC11, 1.3%; FC14, 5.4%) and

in the composite of New Jersey urban areas (FC11, 1.2%; FC14, 5.4%). As a result, the proportion of urban local roads (FC19) was lower (64%) than the proportion of local roads of other metropolitan areas mentioned (over 70%) (Table 6). The largest contributions to on-road source emissions in Elizabeth were from roadways of FC14 (about 33%), followed by contributions from roadways of FC11 (about 30%). More than 75% of aromatic compounds and MTBE were emitted from major roadways (FC14, FC11, FC12) according to the emission inventory data and public roadway information of New Jersey.

Emissions from a specific area source were estimated from the annual emission estimate for Elizabeth divided by the total number of area sources in Union County. The population ratio of Elizabeth to Union county was used to apportion the annual emission for specific area sources in Elizabeth. The national emission inventory of point sources provided the annual generation and the coordinates. The daily emission from a point source was estimated by dividing the annual total by 365 days, which assumes that the facility operated everyday. The emission from the non-road mobile sources was ignored because the total number non-road sources (lawn and garden equipment, snowmobiles, snow blowers, construction equipment etc) in the study area an urban center, was considerably lower than the on-road emissions or the off road emissions for the more suburban regions of Union County.

A number of non-point sources for diesel emissions were identified in and near Elizabeth, NJ. These included: a truck depot and bus depot in north-east Elizabeth, the Port Authority-Marine Terminal in East Elizabeth and the Newark Liberty International Airport located north – north east of Elizabeth. All of these locations were north to north

east of the majority of sampling locations, though the truck and bus depots were close to a subset of homes. No residencies exist intermingled with either the seaport or airport.

The Meteorological Data for New Jersey

Surface Observation Data

Meteorological data for Elizabeth, New Jersey, were obtained from NCDC/NOAA (National Climatic Data Center, National Oceanic and Atmospheric Administration). The data are part of the quality assured national climatological database. The datasets contain hourly observation tables, along with daily and monthly summary tables covering the entire period of the RIOPA Study. The hourly observation datasets were used because those could be matched to the exact 48-hour sampling time of individual samples. The ASCII data files were linked to the RIOPA weather database for data extraction. First, the meteorological data were selected from the observation station that was closest to the study area, the Weather-Bureau-Army-Navy (WBAN) station in the Newark Liberty International Airport (EWR, 14734, Latitude; 40.72°, Longitude; -74.17°). Next, a series of the selection queries in Access were used to retrieve the hourly observation dataset corresponding to each individual sample according to the date/time the sampling was started and ended.

Among the meteorological data extracted, the variables considered as possibly influencing the ambient air concentrations were: the dry bulb temperature (°F), relative humidity (%), precipitation (inches), station atmospheric pressure (inHg), resultant wind speed (knots), resultant wind direction (tens of degrees from true north). The English units were converted to the SI units. Meteorological values averaged for individual 48-

hour sampling periods, were wind speed (U, m/s), temperature (K, Kelvin), atmospheric pressure (mmHg), and relative humidity (RH, %). The precipitation was totaled for the 48-hour sampling period.

Mixing Height Data

The mixing height data were obtained from NCDC/NOAA. The mixing height data were computed from source code made available by the US EPA. The dataset was computed using the upper air data of Brookhaven, NY and the surface data of Newark, NJ. Brookhaven, NY, was the closest monitoring station to the RIOPA study site recording the upper level air data. Mixing heights were reported as AM and PM mixing heights. The values were averaged for individual homes according to the corresponding sampling duration of 48-hour.

Atmospheric Pasquill Stability

The Atmospheric Pasquill Stability classes with a time resolution of 3 hours were retrieved from NOAA AIR Resources laboratory's READY (Real-time Environmental Applications and Display system) web site (http://www.arl. noaa.gov/ready.html). The archived datasets were EDAS (Eta Data Assimilation System) meteorological data (80km, 3 hourly, US). The representative coordinates of Elizabeth (Latitude; 40.65°, Longitude; -74.20°) were used as the location. The text results were tabulated and the stability time-series plots were saved for individual sample dates when available. The 48-hour average stability was calculated from the stability time-series classes for each sample. The classification of the atmospheric stability is described in Table 7.

Table 4. Description of Integrated Data from Databases in the Public Domain

Databases	Description
National Climatological Data	
Hourly observations	ASOS; WBAN number, date, time in local standard time, sky conditions, visibility, significant weather types, dry bulb temperature, dew point temperature, wet bulb temperature, relative humidity, wind speed, wind direction, wind characteristic gusts, value for wind character, station pressure, pressure tendency, sea level pressure, report type, precipitation totals in inches
Hourly precipitation	ASOS; WBAN number, date, time, hourly precipitation
Daily table	ASOS; WBAN number, date, temperature (maximum, minimum, average, departure from normal, average dew point, average wet bulb), degree days (heating, cooling), significant weather types, snow/ice depth and water equivalent, precipitation snowfall, pressure (average station and average sea level), resultant wind speed, resultant wind direction, average speed, maximum 5 second, 2 minute speed and direction
Mixing height	Morning and afternoon mixing height (meters) produced from surface air and upper air data by NCDC/NOAA
Atmospheric stability	Atmospheric Pasquill stability class from NOAA AIR resources laboratory
National Emission Inventory Data	
On-road sources	County level estimates are stratified by type of roadways and vehicles; NEI for criteria pollutants and HAPs for year 1999 (version 3 final)
Non-road sources	NEI for criteria pollutants and HAPs for year 1999 (version 3 final)
Point sources	County level estimates from registered point sources; NEI for criteria pollutants and HAPs for year 1999 (version 3 final)
Non-point sources	County level estimates of non-point sources; NEI for criteria pollutants and HAPs for year 1999 (version 3 final)

Geographic Information and the Spatial Data	
Transportation data	Public roadway mileages, functional class of roadways, vehicle miles traveled by stratified vehicle types; NJ DOT
Census 2000 TIGER data	Line features (roadways, railroads, hydrography etc.), municipality from US Census Bureau

Table 5. Estimated Emission Rates (µg/sec·m) of Selected VOCs for Public Roadways of Union County by its Functional Classes. (Estimation based on 1999 NEI v3 Final)

VOCs	FC11	FC12	FC14	FC16	FC17	FC19
Xylene	50.9	69.8	29.2	5.0	3.4	0.9
Toluene	88.2	121.0	50.4	8.7	5.9	1.6
MTBE	44.4	60.9	25.6	4.4	3.0	0.8
Benzene	32.2	44.1	18.0	3.1	2.1	0.5
Ethylbenzene	13.3	18.3	7.7	1.3	0.9	0.2
Formaldehyde	17.6	24.2	11.1	1.92	1.31	0.3
Acetaldehyde	5.12	7.02	3.22	0.56	0.38	0.1
Acrolein	0.70	0.95	0.49	0.09	0.06	0.01

Table 6. Percent Contribution of On-road Source Emission by Roadway Types in the City of Elizabeth, NJ (Estimation based on 1999 NEI v3 Final)

VOCs	FC11	FC12	FC14	FC16	FC17	FC19	Total
Xylene	29.4	13.1	32.6	9.9	5.2	9.8	100
Toluene	29.5	13.1	32.6	9.9	5.2	9.7	100
MTBE	29.2	13.0	32.5	9.9	5.2	10.2	100
Benzene	29.8	13.3	32.4	9.8	5.2	9.4	100
Ethylbenzene	29.2	13.0	32.7	9.9	5.2	9.9	100
Formaldehyde	17.4	14.3	27.4	16.5	7.2	17.1	100
Acetaldehyde	17.4	14.4	27.4	16.5	7.2	17.0	100
Acrolein	16.0	13.2	28.4	17.1	7.5	17.8	100

Table 7. The Description of the Classification of the Atmospheric Pasquill Stability

Pasquill Stability Class	Description	Coded
A	Extremely unstable conditions	1
B	Moderately unstable conditions	2
C	Slightly unstable conditions	3
D	Neutral conditions	4
E	Slightly stable conditions	5
F	Moderately stable conditions	6
G	Extremely stable	7

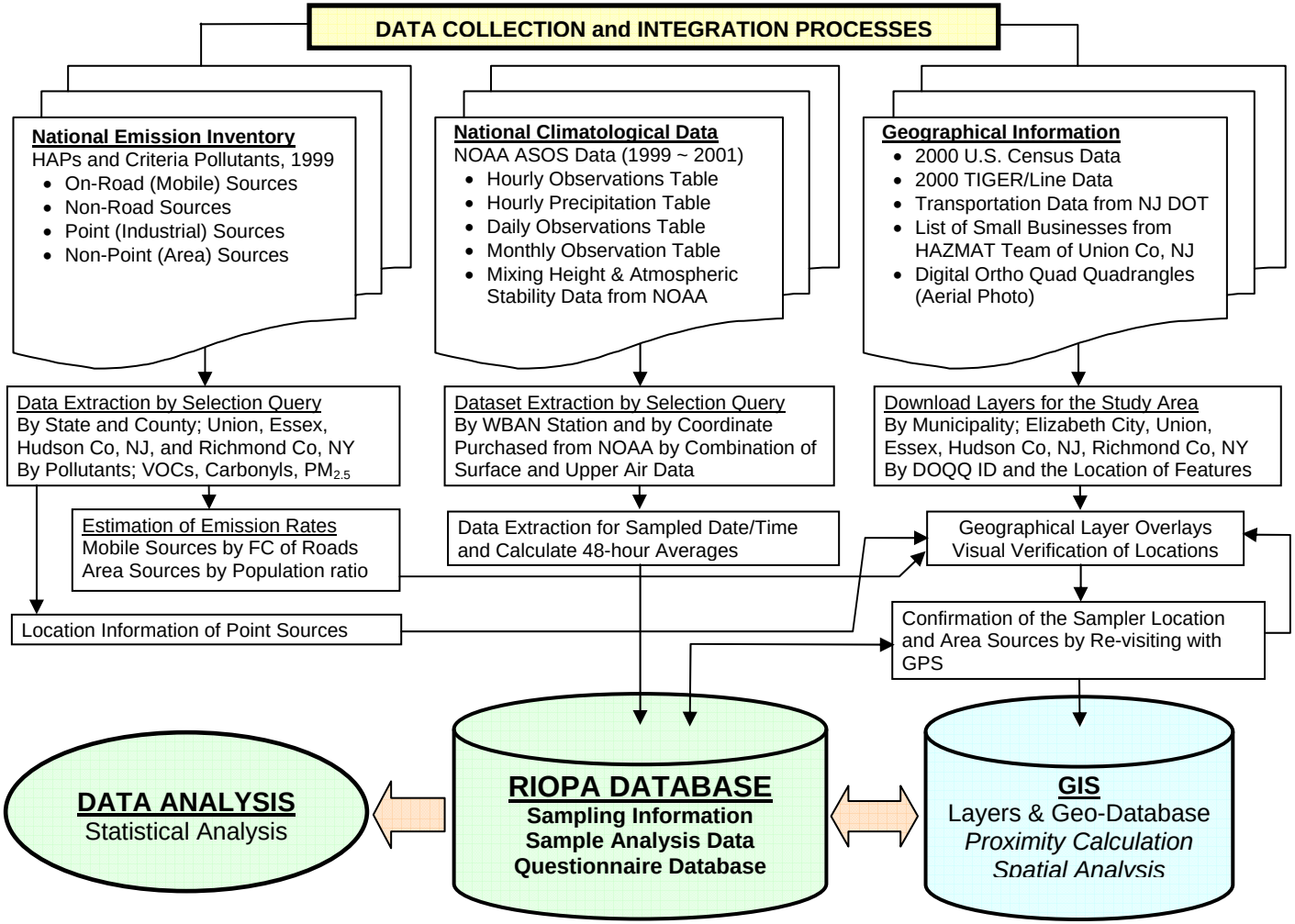

Figure 2. Data Integration Processes of the Public Databases into the RIOPA Database for Data Analysis of New Jersey Site

Geographical Information Systems

ArcView GIS (version 3.1, ESRI, Inc.) was used to build the geographical inputs for statistical analysis. The spatial analyst extension used was for geo-processes such as dissolve, merge, clip, union, spatial join, and select themes. The scripts downloaded were used to measure the distances between geographical locations. For the geographic coordinates of projection, NAD83 (North American Datum 1983), New Jersey State Plane 1983 was used with units of decimal degrees and feet using ArcScript, Addxycoo (ESRI). GIS application itself provided a powerful database tool for integration of datasets by joining and linking databases.

Census 2000 TIGER/Line® Datasets

The Census 2000 TIGER® (Topologically Integrated Geographic Encoding and Referencing system) datasets were downloaded from the Geography Network (US Census Bureau, Geography Division, http://www.census.gov/geo/www/tiger). The line features included were roads, railroads, and hydrography. The polygon features were municipal boundaries such as county, township, and city borderlines. Not only were the spatial data of Union County, NJ included in the resulting map, but also the spatial features of adjacent counties (Essex, Hudson Counties, NJ and Richmond County, NY) since the proximity information and source emissions were also reviewed for these counties (figure 3).

Digital Images

Digital orthoquarter quadrangles (DOQQs) are the combined image of a photograph with geometric qualities of a map. The primary digital orthophotoquad has a 1-meter ground resolution, quarter-quadrangle (3.75-minutes of latitude by 3.75-minutes of longitude) image cast on the Universal Transverse Mercator Projection (UTM) on the North American Datum of 1983 (NAD83). For the RIOPA study area in New Jersey, the corresponding 1997 DOQQs were downloaded from the New Jersey Image Warehouse site of the NJ DEP, Bureau of GIS (http://njgin.nj.gov/OIT_IW/index.jsp). The downloaded DOQQs are listed in Table 8. Figure 4 illustrates the digital image of the City of Elizabeth with municipal borderlines.

New Jersey Road Network

The functional classes of roadways (Table 9) in Elizabeth were obtained from the functional classification map of Union County from the Bureau of Transportation Data and Development in Department of Transportation of New Jersey (http://www.state.nj.us/transportation/refdata). The functional class information was assigned to the appropriate road segments using the roadway line feature layer of ArcView GIS® project file. The Straight Line Diagrams provided a graphical representation of state, toll, and county roads and showed intersecting streets, administrative and geometric characteristics. The Straight Line Diagrams provided the width of the roadways for estimating the general offset distance from the centerline of roadways. The offset distance used was one half the roadway width and was required to specify the location of each home relative to that the roadway centerline. This allowed the home to be placed on the correct side of the

roadway rather than on the center line and to calculate the distance from the home to the center line of the roadway. Offset distances of 20 to 30 meters were used based on the functional classes of the roadways. The customized map of the public roadways in the study area is illustrated in Figure 5.

Location of Area and Point Sources

Lists of the street locations of service stations were obtained from visual observation and written records made during the sampling, from web sites that list gasoline stations by zip code for price comparison (http://www.gaspricewatch.com/USGas_index.asp), and from the yellow pages (http://www.yellowbook.com) for Elizabeth, New Jersey. After combining and comparing the information contained in these lists, it was determined a more reliable compilation was still needed. This was obtained from the Emergency Response/HAZMAT of Union County, Division of Environmental Health and Emergency Management. The list of the actually operating dry cleaning facilities in the City of Elizabeth, NJ, was also obtained from HAZMAT Team of Union County. Figure 6 and Figure 7 are the maps of gas station and dry cleaning facilities identified and located in the study area. The latitude and longitude of the point sources identified in the study area from the emission inventory database were provided with the list used to generate customized maps by making event themes. (Table 10), (Figure 8).

Quality Assurance of Geographical Data

To evaluate the effect of proximity and meteorological conditions simultaneously, the relative locations of sources and sampling sites should be defined precisely. All downloaded geographical layers were overlaid on the New Jersey State Plane of NAD 83. TIGER maps placed road centerlines substantial distances (15 ~ >50 meters) from actual location based on aerial photos (DOQQs). Therefore, to obtain the needed accuracy of the proximity data acquisition, TIGER data were evaluated before geo-coding and calculating the distance between road centerline and receptor location. The errors of 2000 TIGER/Line® data were corrected by following the centerlines of the roadways observed on the overlaid DOQQs as reference themes. The point themes were finalized after correcting the locations based on the street information collected during confirmation trips done by driving to each address listed in the RIOPA dataset, digital orthophoto, the Elizabeth City engineer's map, pictures taken from the sampling, and the GPS readings from the confirmation trip. The GPS unit used to read the coordinates was a GeoStats wearable GeoLogger™. The GPS reading was used solely as an aid to locate the houses during the quality assurance visit to Elizabeth. The values retrieved from the GPS were not used in the data analysis, rather the longitude and latitude obtained from the GIS mapping was used. The corrected point themes were the locations of the outdoor sampler, point sources, gas stations, and the dry cleaning facilities (figures 5 – 8). Approximate receptor (outdoor sampler) locations are given for each residence in figure 9 for illustration purposed to maintain confidentiality of the subjects, actual locations coordinates were used to determine proximity to sources.

Measurement and Calculation of Geographical Data

The location of the residences, point sources, and area sources were determined by the address-matching technique within ArcView on the corrected and quality assured line files from Census 2000 TIGER/Line® as the reference theme using US streets with zones. The spatial coordinates of the point themes, such as residences, point sources, gas stations, and dry cleaning facilities were determined by "Addxycoo", a commonly used ArcScript. The distances from point theme to point theme and the distances, from point theme to line theme were measured by "the nearest features", an extension patch available in ESRI's site for ArcScripts (http://arcscripts.esri.com).

Table 8. The List of Digital Orthoquarter Quadrangles Used in this Study for Quality Assurance (Source: NJ DEP)

QQ Number	QQ Name
514	SE ROSELL NJ
521	NW ELIZABETH NJ-NY
522	NE ELIZABETH NJ-NY
523	SW ELIZABETH NJ-NY
524	SE ELIZABETH NJ-NY

Table 9. The Functional Classification of Public Roadways in Urban Area (Source: NJ DOT)

Functional Class	Description
FC 11	Urban Interstate Highways
FC 12	Urban Other Highways/Freeways
FC 14	Urban Major Arterial
FC 16	Urban Minor Arterial
FC 17	Urban Collector
FC 19	Urban Local

Table 10. The Point Sources of Selected VOCs Used for Data Analysis (Source: 1999 NEI for HAPs version 3 Final)

PS ID	Emissions	Facility/Process	X	Y
Xyl_PS1	1.95	Refinery	-74.22	40.64
Xyl_PS2	0.94	Tanker Terminal	-74.25	40.63
Xyl_PS3	0.91	Industry	-74.19	40.67
Xyl_PS4	0.24	Aviation Service	-74.17	40.70
Tol_PS1	4.05	Refinery	-74.22	40.64
Tol_PS2	3.03	Tanker Terminal	-74.25	40.63
Tol_PS3	2.14	Industry	-74.19	40.69
Tol_PS4	0.50	Industry	-74.22	40.63
Tol_PS5	0.38	Aviation Service	-74.17	40.70
Bzn_PS1	4.55	Refinery	-74.22	40.64
Bzn_PS2	1.73	Tanker Terminal	-74.25	40.63
Bzn_PS3	0.20	Joint Meeting of Essex and Union	-74.20	40.64
Bzn_PS4	0.10	Aviation Service	-74.17	40.70
Ebz_PS1	0.60	Refinery	-74.22	40.64
Ebz_PS2	0.27	Industry	-74.19	40.67
MTBE_PS1	43.50	Refinery	-74.22	40.64
PCE_PS1	1.03	Refinery	-74.22	40.64

PS ID: Point Source ID, Emissions are annual total generations in metric tons, X: Longitude, Y: Latitude.

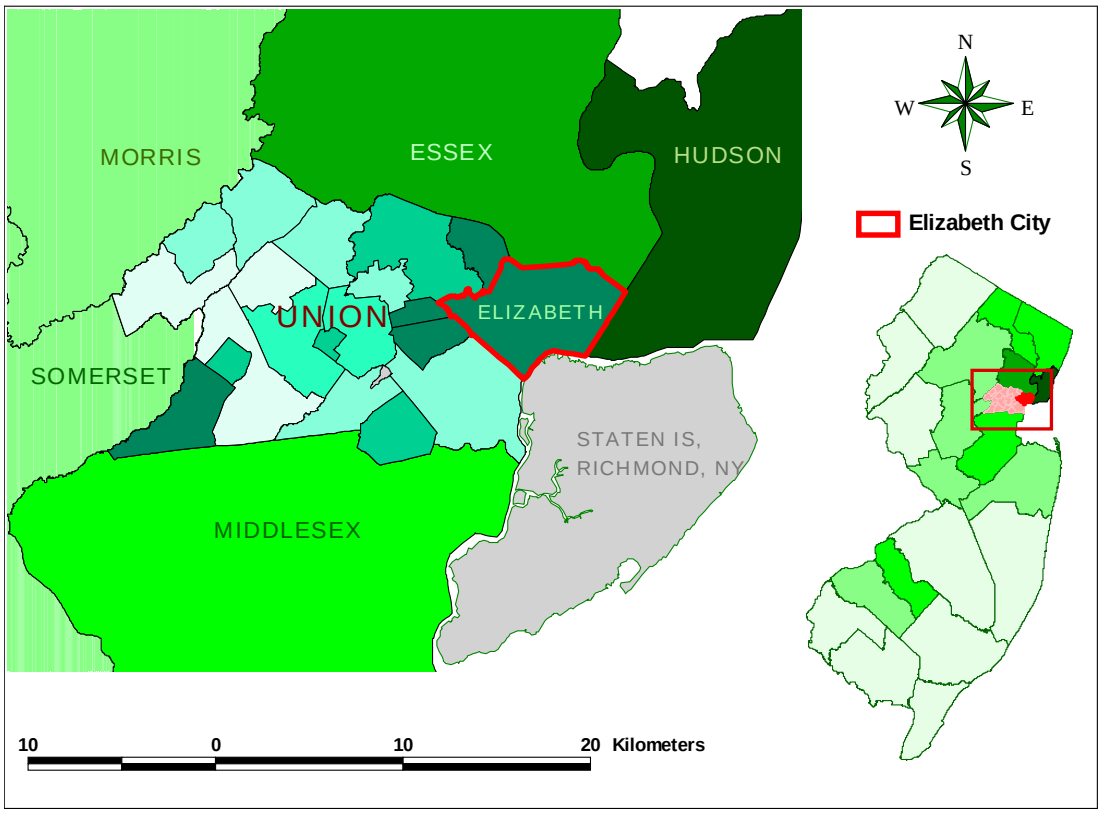

Figure 3. The Location of Union County and City of Elizabeth in New Jersey

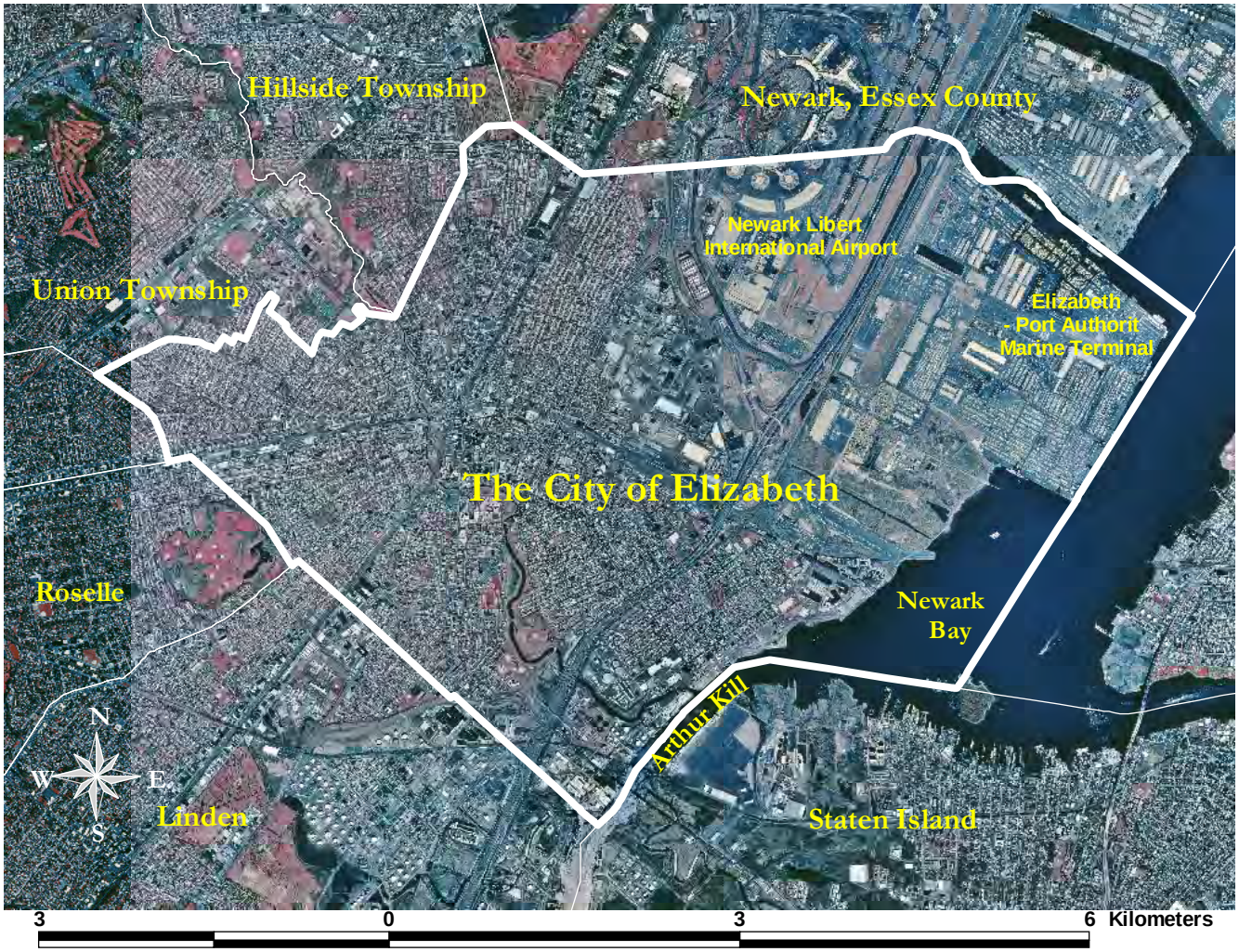

Figure 4. Digital Image of Study Area, the City of Elizabeth, New Jersey (Source of DOQQs: NJ DEP, jpeg97)

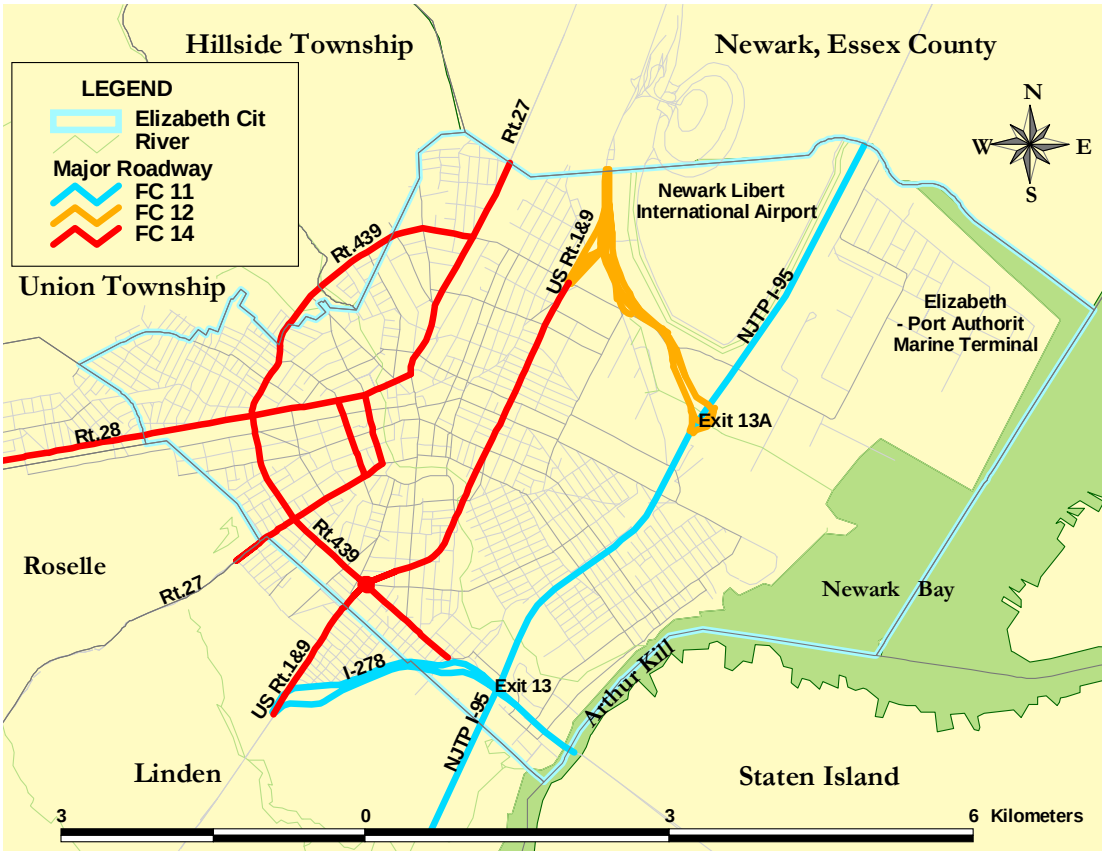

Figure 5. Major Public Roadways in Study Area, the City of Elizabeth, New Jersey

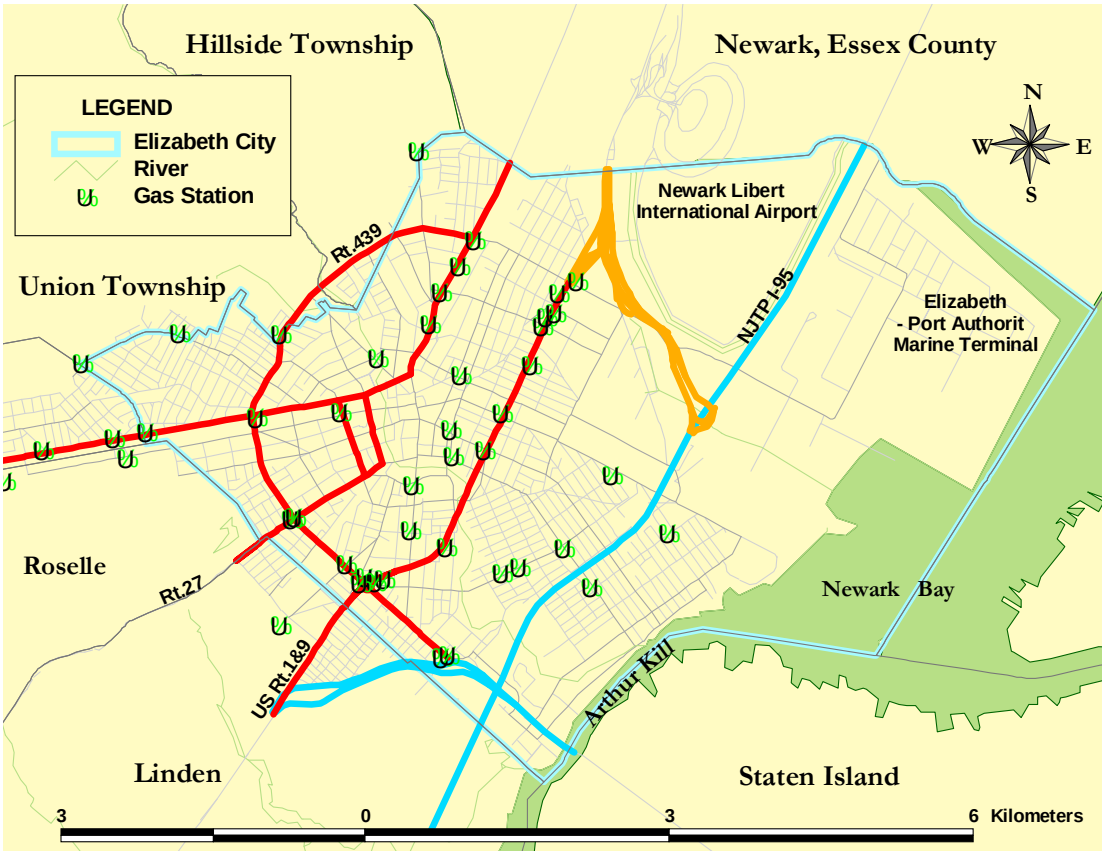

Figure 6. Identified Gas Stations in Study Area, the City of Elizabeth, New Jersey (Source: HAZMAT List of Union County)

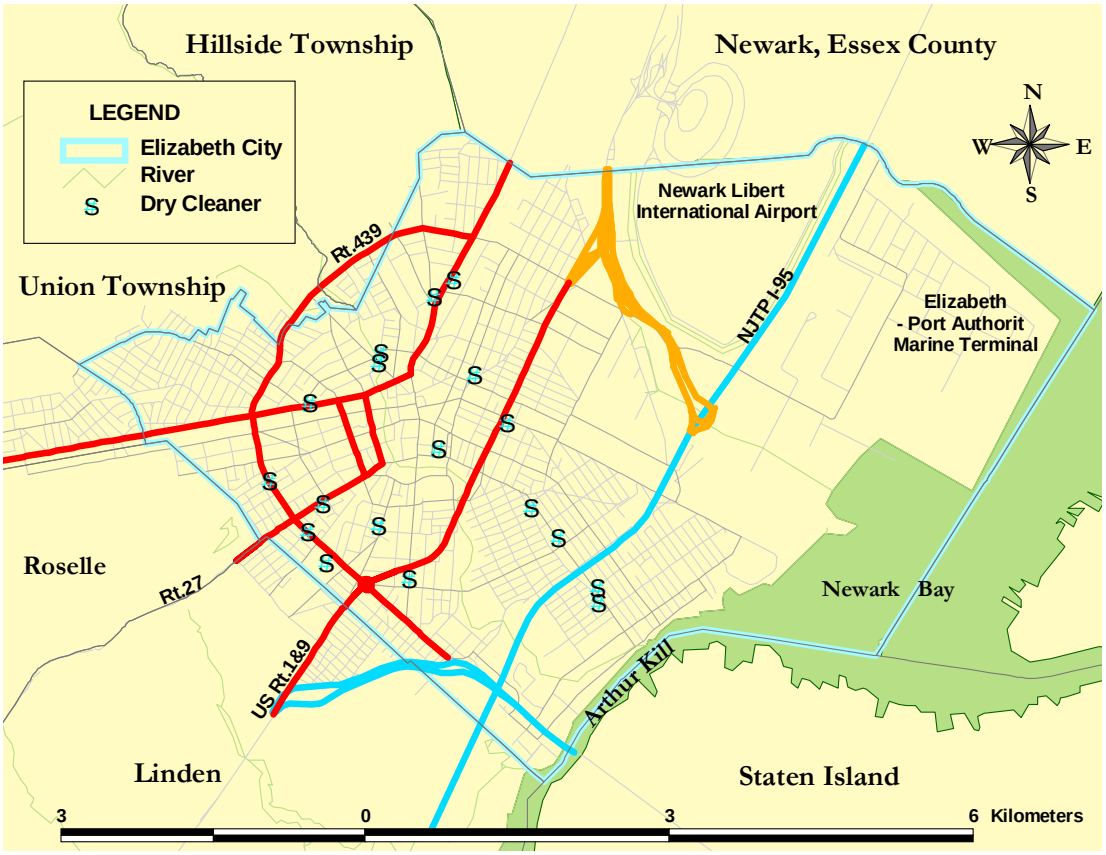

Figure 7. Identified Dry Cleaning Facilities in Study Area, the City of Elizabeth, New Jersey (Source: HAZMAT List of Union County)

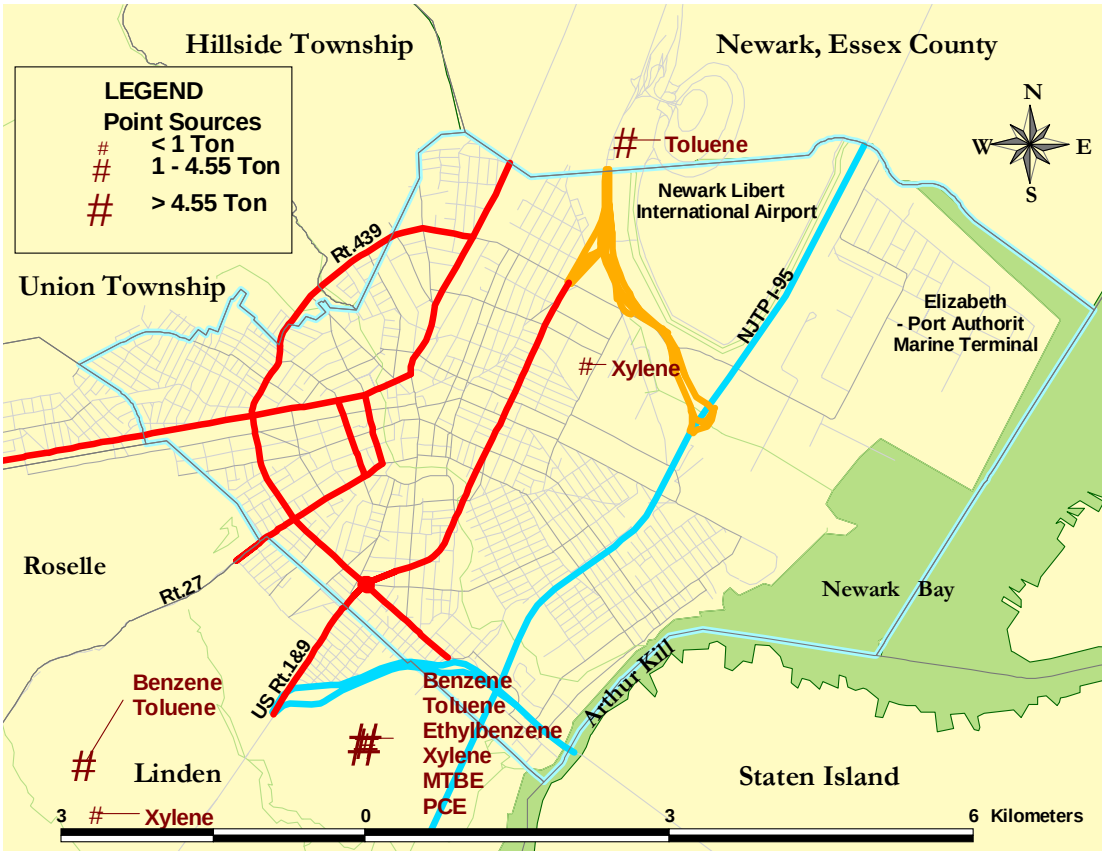

Figure 8. Identified and Selected Point Sources of VOCs Studied in Study Area, the City of Elizabeth, New Jersey (Source: 1999 NEI for HAPs, Version 3 Final)

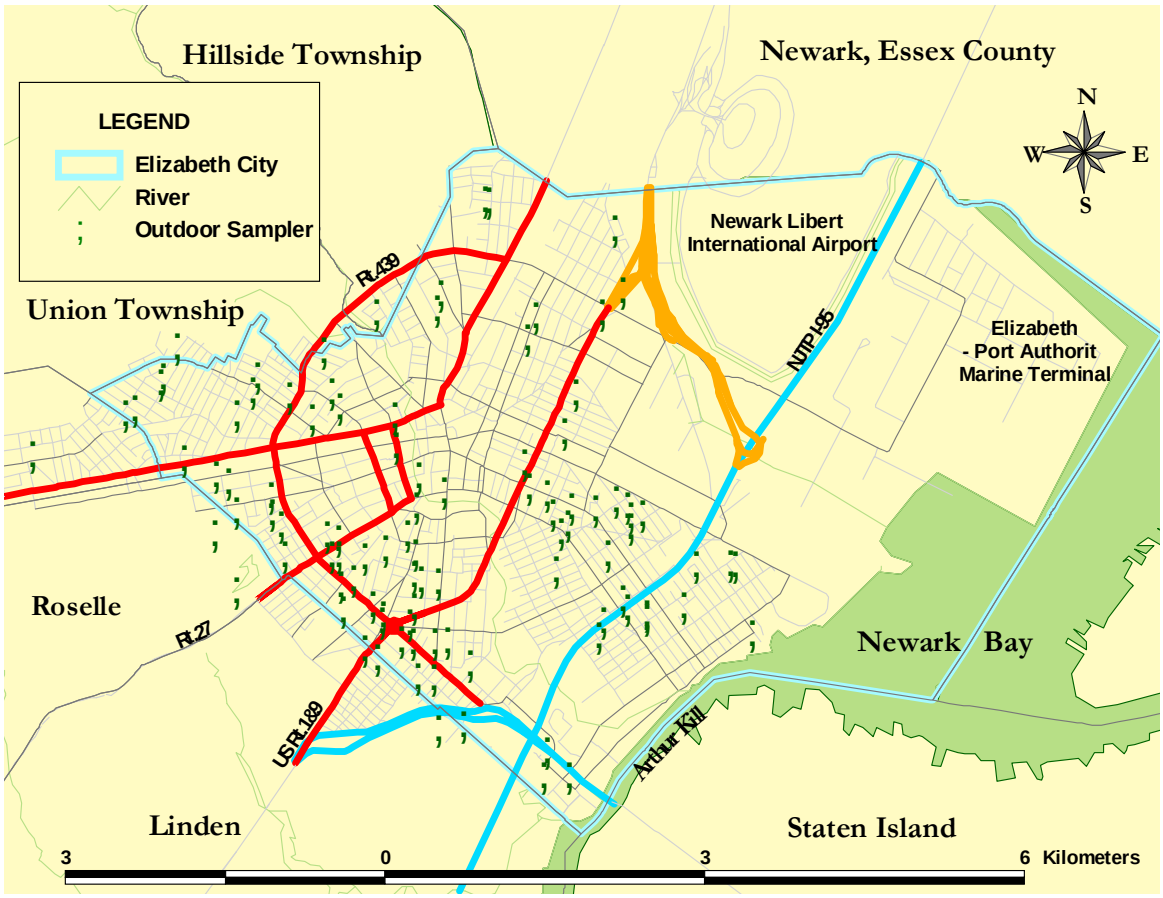

Figure 9. Approximate Locations of Outdoor Samplers in the City of Elizabeth, New Jersey (Locations randomly shifted by small amount for illustration purpose to preserve subjects' confidentiality, Source: RIOPA Questionnaire Database, 2003)

Statistical Analysis

Statistical Treatment of Data

The SAS system for Windows (version 8.02) and SPSS for Windows (version 12.0) were used for all statistical analyses in. The blank subtracted, temperature adjusted, and uncensored ambient air concentrations ($\mu g/m^3$) of the selected air toxics and $PM_{2.5}$ were evaluated.

The distributions of the residential ambient air concentrations were examined by the one-sample Kolmogorov-Smirnov (K-S) test to evaluate their normality. Natural log-transformation of the concentrations was performed because it provided distributions that were closer to a normal distribution with more constant variance than the un-transformed concentrations. Any zero values in the uncensored dataset were replaced with one half the minimum diction limit prior to the statistical analysis.

The sample means, standard deviations, median, percentiles, the minimum and maximum values for the variables were computed. The scatter plots of residential ambient air concentration and each independent variable were examined for obvious associations.

Bivariate Pearson correlation coefficients and the significance of the statistics were computed to examine the correlations between the response variables and the predictor variables for the purpose of preliminary selection of the more influential explanatory predictors among the groups of candidate variables. Correlations of un-transformed, ln-transformed, inversed, squared, and inverse squared of the concentration and predictor variables were examined.

Two samples were collected from most homes several months apart and on most

days one or two homes were visited, though occasional three or four homes were sampled on a single day. The Mixed Model Proc in SAS was run with home identification number and with date as the repeated measure to evaluate if whether multiple samples at the same location or date affected the results. No affect was observed.

Multiple Linear Regression Analyses

Multiple regression analysis was used to examine the association between the ambient air concentrations and the proximity and meteorological variables. A multiple linear regression equation that expresses the response variable as a linear combination of (p – 1) predictor variables, has the form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{P-1} X_{i,\,P-1} + \varepsilon_i$$

where:

Y_i is the response in the i th trial

$\beta_0, \beta_1 \cdots \beta_{P-1}$ are the parameters (regression coefficients)

$X_{i1}, X_{i2} \cdots X_{i,\,P-1}$ are the values of predictor variables

ε_i is the error term

This equation assumes that the relationship of independent variables with response variable is linear, and that the distribution of error terms is normal with equal variance. Two of the explanatory variable groups considered important for predicting residential ambient air concentrations of the selected VOCs were the proximity of a residence to the emission sources and the corresponding meteorological conditions. Distances from residences to mobile, area, and point emission sources identified from the emission inventories, wind speed, atmospheric stability, mixing height, temperature,

relative humidity, precipitation, and atmospheric pressure were used as the independent variables.

Selection of the predictors associated with elevated ambient air concentrations around residences were examined using several multiple linear regression analyses methods: forward selection, backward elimination, stepwise selection, r squares, and maximum r^2 improvement methods, to verify that consistent results were obtained independent of the type of regression model used. Final model were determined using stepwise selection. The default criteria of each method in the SAS program (version 8.02) were used for selecting variables to be included in the resulting model. The parameter selection criteria used for forward selection, backward elimination, and stepwise selection were $p<0.50$, $p<0.10$, and $p<0.15$, respectively. Due to the different levels in selection criteria, the number of predictors included in resulting models differed. The models selected by the different selection methods were compared and evaluated by the p values of parameter estimates of predictor variables and the composition of variables in the model. When the best-fitting model was selected for a VOC compound, the model and the corresponding statistics were also evaluated. The equality of error variances of the best-fitting model was visually examined on the appropriate diagnostic plots and statistics computed. See Appendix A for discussion of multicollinearity which was identified among several meteorological variables.

Identification and Tests of Outlying Observations

Details on how outliers were determined are given in Appendix A. Standardized residuals were examined with a criteria of *tinv* (0.95, *n-p*-1) of ± 1.654 based on a

minimum degree of freedom of 170 to determine if a value was a statistical outlier. Presence of outliers suggest other processes, not accounted for by the independent variables selected, was contributing to the concentration or there was analytical uncertainty in the measurement. The final model chosen excluded those values (which were <10% of the measurements) to determine the strength of the model for the data that could be predicted, as the focus of this analysis was to establish how proximity affected concentration. A separate analysis could be informative to indicate why outliers to the regression analysis exist. The actual degree of freedoms for each compounds were as follows; m,p-xylene (171); o-xylene (174); toluene (174); benzene (175); ethylbenzene (171); MTBE (169); and PCE (161).

To test if the outliers removed from the multiple regression model, biases the model outcome ANOVA tests were used to compare the means of independent variables between groups of outliers and non-outliers. To verify that removing the outlying observations did not eliminated specific conditions or situations, the analysis of variance (ANOVA) tests were performed on the means of the predictor variables between group of outliers and group of non-outliers. The regression model was run excluding the outliers to obtain the final, best fit equation for each compound.

Diagnostics of Unequal Error Variances and Multicollinearity

To test the assumption of equal error variance, the heteroscedasticity of the parameter estimates were tested To determine whether the error variance was constant over all cases (Neter et al., 1996). The null hypothesis for this test is that the errors are homoscedastic, independent of the predictors. Therefore, the equal error variance was

assumed in the best-fitting model when the probability (p) of the chi-square test was greater than 0.05.

The multicollinearity, which results from linear interactions between the predictor variables, was tested because codependency might be detrimental when interpreting the resulting regression model. First, the bivariate Pearson correlations between pairs of predictors included in the final models were examined to identify the highly correlated pairs of the predictors. Second, the magnitude of variance inflation factor (VIF) was examined to determine if it was greater than 10. Third, the condition index and eigenvalue were examined from the collinearity diagnostics. A condition index greater than 100 and an eigenvalue smaller than 0.01 was considered evidence of multicollinearity in the model since those values indicate the presence of highly correlated variables when the proportion of variation is greater than 0.5.

Use of Dummy Variables (for Seasonality)

The three indicator (dummy) variables were introduced to the finalized best-fitting models of selected VOCs. To avoid the not-fully ranked model problem, dummy variables for spring, summer, and fall were generated by assigning 1 for the season of the sampled date, and by assigning 0 for the other seasons. Therefore, the winter would be defined by all three indicator variables to be zero.

RESULTS

Dataset Extraction for Data Analysis

The RIOPA database was integrated with source emission inventory and meteorological information to provide datasets for statistical data analyses that contained accurate proximity information of emission sources of each sample with corresponding meteorological conditions for each 48-hour sampling period. The blank subtracted, temperature adjusted, uncensored residential ambient air concentrations of selected VOCs: *m,p*-xylene, *o*-xylene, toluene, benzene, ethylbenzene, and MTBE, carbon tetrachloride and PCE (as control compounds); Aldehydes: formaldehyde, acelydehyde, and acrolein; and Particulate Matter: PM2.5, elemental carbon, organic carbon and two PAHs were examined. The distances from residences to identified mobile, area, and point sources were determine as was the averages of meteorological variables for each time period a sample was collected.

Descriptive Statistics :

The sample means, standard deviations, median, percentiles, and the maximum values for the concentrations ($\mu g/m^3$) of selected the target compounds measured in residential ambient air are listed in Table 11. The sample means, standard deviations, median, percentiles, the minimum and maximum values of the closest distances from the location of the RIOPA sampler to the public roadways by its functional class, and by the roadway name are listed in Table 12 and 13, respectively. The sample means, standard deviations, median, percentiles, the minimum and maximum values of distances from sampler to the closest area and point sources are listed in Table 13. The sample means,

Table 11. Concentrations of Selected VOCs in Residential Ambient Air (µg/m^3, N=183)

Compounds	Mean	Standard Deviation	Percentiles				Maximum	Comparison NJ Urban Concentration
			25	50	75	90		
m,p-Xylene	3.25	4.29	1.51	2.37	51.21	3.97	6.44	2.6
o-Xylene	1.71	6.51	0.59	0.94	80.98	1.38	2.16	1.2
Toluene	6.82	5.83	2.59	4.83	32.88	9.36	14.67	5.7
Benzene	1.50	1.54	0.69	1.22	18.06	1.90	2.68	0.62
Ethylbenzene	1.34	2.74	0.46	0.99	36.24	1.74	2.51	0.92
MTBE	5.75	5.34	2.23	4.35	27.17	7.51	12.13	6.83
Tetrachloroethylene	1.10	3.09	0.50	0.74	41.82	1.11	1.50	0.40
Carbon Tetrachloride	0.84	2.28	0.48	0.69	39.1	0.81	0.94	0.09
Formaldehyde	6.35	2.81	2.71	7.09	10.7	8.29	9.33	2.3
Acetaldehyde	8.88	6.50	3.05	7.86	38.7	10.2	14.6	1.1
Acrolein	0.89	1.29	0.13	0.39	6.21	0.78	1.69	-
$PM_{2.5}$ Mass	20.4	10.7	13.8	18.2	71.7	25.5	30.9	15.8
Elemental Carbon	1.36	0.64	0.92	1.29	3.51	1.72	1.96	-
Organic Carbon	3.33	1.73	2.07	3.00	9.46	4.00	5.61	-

A NJDEP mean concentrations reported in Elizabeth, NJ, 2001 (www.state.nj.us/dep/airmon/toxics01.pdf)

Table 12. The Closest Distances from Sampler Location to the Public Roadways by Functional Classes (km, N=183)

Roads	Mean	Standard Deviation	Minimum	Percentiles			Maximum
				25	50	75	
FC11	1.53	1.05	0.04	0.68	1.33	2.28	3.70
FC12	2.53	1.16	0.02	1.47	2.87	3.44	5.58
FC14	0.50	0.54	0.01	0.11	0.33	0.65	2.49
FC16	0.19	0.17	0.01	0.07	0.13	0.32	0.78
FC17	0.29	0.22	0.02	0.11	0.25	0.39	0.97
FC19	0.03	0.02	0.00	0.02	0.03	0.04	0.13

13. The Closest Distances from Sampler Location to Individual Public Roadways (km, N=183)

Roads	Mean	Standard Deviation	Minimum	Percentiles			Maximum
				25	50	75	
I95 [a]	1.89	1.23	0.05	0.86	1.73	2.88	5.33
Rt1 [b]	1.10	0.83	0.03	0.42	0.93	1.72	3.62
Rt27 [b]	1.23	0.83	0.04	0.50	1.02	1.85	3.40
Rt28 [b]	1.54	0.87	0.10	0.88	1.51	2.08	3.59
Rt439 [b]	0.93	0.81	0.01	0.24	0.61	1.61	2.86

a: Interstate (FC11), b: Major Arterial (FC14)

Table 14. The Closest Distances from Sampler Location to Area Sources and Point Sources Likely Impact Elizabeth, NJ (km, N=183)

Emission Sources	Mean	Standard Deviation	Minimum	Percentiles			Maximum
				25	50	75	
Gas Station	0.36	0.21	0.03	0.22	0.36	0.49	1.01
Dry Cleaning Facilities	0.55	0.39	0.06	0.25	0.43	0.77	1.69
Refinery [a]	2.98	1.12	0.84	2.06	3.07	3.77	5.76
Tanker Terminal [b]	4.78	1.14	3.23	3.78	4.58	5.72	7.69
Industry [c]	2.51	1.00	0.62	1.75	2.60	3.26	5.63
Aviation Service [d]	5.92	1.15	2.80	5.03	6.18	6.91	8.63
Industry [e]	4.19	1.11	0.81	3.50	4.43	5.08	6.64
Industry [f]	3.27	1.14	0.99	2.46	3.24	3.98	6.04
Joint Meeting of Essex and Union [g]	2.77	1.20	0.40	1.91	2.36	3.80	5.81

a: Refinery = Xyl_PS1, Tol_PS1, Bzn_PS1, Ebz_PS1, MTBE_PS1, PCE_PS1; b: Tanker Terminal = Xyl_PS2, Tol_PS2, Bzn_PS2; c: Industry = Xyl_PS3, Ebz_PS2; d: Aviation Service = Xyl_PS4, Tol_PS5, Bzn_PS4; e: Industry = Tol_PS3; f: Industry = Tol_PS4; g: Joint Meeting of Essex and Union = Bzn_PS3

Table 15. The Meteorological Variables (N=183)

Variable, Unit	Mean	Standard Deviation	Minimum	Percentiles			Maximum
				25	50	75	
Temperature, K	284.2	8.0	265.5	279.4	284.6	289.9	303.3
Wind Speed, m/s	4.3	1.1	1.9	3.6	4.4	5.1	8.0
Relative Humidity, %	66.3	12.6	42.7	58.1	66.4	75.9	91.8
Atmospheric Pressure, mmHg	762.3	4.5	750.3	759.6	761.6	765.5	773.1
Precipitation, mm	0.0207	0.0249	0.000	0.000	0.010	0.040	0.130
Mixing Heights, km	1.027	0.362	0.414	0.767	0.948	1.214	2.099
Pasquill Stability Class	5.028	0.444	3.867	4.706	5.000	5.300	6.063

standard deviations, median, percentiles, the minimum and maximum values of the meteorological condition variables are listed in Table 15.

DISCUSSION

Prior to establishing the best-fit linear regression equations for each compound Bivariate Pearson Correlations were conducted to guide the inclusion of different variables and examine associations among the variables. The *ln* transformed concentration data were used since the concentration distribution was consistent with a log normal distribution and linear regression analyses assumes a normal distribution for the independent variable. The inverse distance was used since concentration declines inversely from line sources, such as roadways, or as the square of the inverse for point sources based on an idealized Guassian Dispersion. The square of inverse distance was also examine, but no differences in results were observed, so only the inverse distance was retained in the final mathematical models. As described in the method section andAppendix A outliers were identified for the regression model calculated from the entire data set and a second regression model was determined after eliminating the outliers. The variables selected in the model were examined for multicollinearity. Details for the model evaluated for each compound are given in Appendix A.

Test of Outlying Observations

To verify that the outlying observations were not eliminated based on specific conditions, the ANOVA tests were performed on the means of the predictor variables

between group of outliers and group of non-outliers. Duncan's multiple range test results indicated that means of predictor variables were not significantly different between groups of outliers and non-outliers for selected VOCs. The frequency of outliers removed is listed by season in Table 16.

Model Summaries

The relative contribution to residential ambient air concentrations due to proximity to ambient sources on the selected air toxics and $PM_{2.5}$ with corresponding meteorological conditions were determined by multiple linear regression analyses (Table 16). The *F* statistics were significant for overall models except carbon tetrachloride ($p<0.0001$). Probabilities for parameter estimates were more significant and the r^2 larger for the meteorological variables than the proximity variables. This implies that a greater percentage of the explanatory power of the regression equations for these compounds were due to changes in the meteorological conditions than the distance to a source (see below). There were some interactions between the predictor variables in the best-fitting model, especially between the meteorological variables. The model coefficients of determination for the compounds that included proximity predictors varied between 0.16 and 0.47 (Table 17). The samples and meteorological data were averaged over 48 hours, reducing the possibility of accounting for shorter term variability that could alter the air concentrations.

Among the variables associated with proximity to mobile source emissions, the inverse distance to major urban arterial roadways (FC14) was selected as significant predictor in best-fitting models of residential ambient air concentrations of all of the

aromatic compounds and the inverse distance to the NJ Turnpike (FC11) for $PM_{2.5}$, organic carbon and the two individual PAHs examined (coronene and Benzo[ghi]pyrene). The inverse distance to the closest gas station was included as a predictor in the models of residential ambient air concentrations of *m,p*-xylene, *o*-xylene, benzene, and MTBE. The inverse distance to areas in Elizabeth that had high truck traffic that included loading and unloading and therefore idling trucks was included in the models for $PM_{2.5}$ and elemental carbon while the inverse distance to the refinery in Linden, NJ was included in the regression equation for elemental carbon. The inverse distance to the closest dry cleaning facility was selected as a significant predictor variable in the model of residential ambient air concentration of PCE in Elizabeth, NJ. No variables associated with the inverse distance to sources were identified for the three aldehyde compounds. Nor were any of the proximity factors included in the control variable that did not have mobile source emissions, carbon tetrachloride, tetrachloroethylene, particulate sulfur and particular selenium.

Among the meteorological condition variables atmospheric stability, mixing height, temperature, wind speed, and relative humidity were significantly associated with one or more of the residential ambient air concentrations. The atmospheric stability and temperature were consistently included as statistically significant predictors in the best-fitting models of the aromatic compounds, MTBE and the particulate species, while mixing height was selected for acrolein. Atmospheric stability is calculated based on mixing height and temperature. Wind speed was included with a negative coefficient in most models.

A consistency in the parameter estimates of the proximity variables are observed among the aromatic compounds. The order of mobile emission strength in Elizabeth, NJ (Table 2) is toluene, xylenes (*m,p* xylenes is greater than *o* xylene), benzene and ethyl benzene, the same order as the magnitude of the coefficients in the regression equation, though the sum of the coefficients of o and *m/*p xylene exceeds that of toluene. The order of the coefficients for proximity to gasoline stations (GS^{-1}) is MTBE, *m,p* xylene, benzene and *o* xylene with GS^{-1} not included in the regression equation for toluene. MTBE is the compound with the highest concentration in gasoline and has the highest vapor pressure (0.309atm) of the VOCs studied. The next most prevalent compound of those that included GS^{-1} is *m,p* xylene. Lastly, while *o* xylene might be at a higher concentration than benzene in gasoline, benzene has a high vapor pressure (0.125atm) than the xylenes or ethyl benzene (0.0109-0.0125atm). Thus, the parameter coefficient order is consistent with the abundance of these compounds in gasoline as modified by the vapor pressure. It is unclear why proximity to gas stations was not included in toluene's regression equation since its concentration in gasoline is second only to MTBE and its vapor pressure is between that of benzene and *m,p* xylene, The lack of inclusion GS^{-1} in the regression equation for ethyl benzene might reflects its lower concentration in gasoline and the lower air concentration with more values being below detection.. The regression equation for MTBE did not include distance from arterial roadways, while the aromatic compounds did, but did include distance from the interstate highway. These differences may reflect the more efficient combustion and removal in the catalytic converter of MTBE compared to the aromatic compounds and lower tailpipe emissions along with a (Poulopoulos and Philippopoulos 2003).

Two polyaromatic hydrocarbons (PAHs) measured in the $PM_{2.5}$, coreonene and benzo[ghi]perylene, were evaluated for effects of proximity to mobile sources. Coronene has been used as an index PAH compound to differentiate between gasoline and diesel vehicles because coroene is found in emissions from gasoline powered vehicles, but has not been detected in diesel emissions (Rogge et al, 1993). Benzo[ghi]perylene is present in both diesel vehicle and gasoline vehicle exhaust, so should be an individual compound representative of mobile source emissions (Harrison et al. 1996). It should be noted that these compounds are also emitted from other combustion sources, so may not be solely from mobile sources. Both compounds included the inverse distance to FC11 as well as atmospheric stability, temperature and wind speed in the regression equation (Table 17). The elemental carbon was associated with FC14. FC14 has three (Rt 1/9, Rt 27, Rt 439) roadways that are major truck thoroughfare. A number of the homes in the study on very close to FC14 and FC14 was the mobile source area associated with the aromatic hydrocarbons. No clear difference in what sources contributed to $PM_{2.5}$ mass and the two individual PAHs that may be markers of mobile sources, was identified. The weakest associations were observe for organic carbon which is expected to have more sources besides combustion and diesel emissions than the other components of PM. All PM components were influenced by a variety of meteorological factors, proximity to the NJ Turnpike, major arterial roads and/or truck loading/unloading areas.

As a check on the possibility that there was an inherent bias in the sampling or analyses that caused the associations between the proximity to mobile sources in the regression equations to volatile compounds without mobile sources, carbon tetrachloride and tetrachloroethene were evaluated. Carbon tetrachloride has little industrial or

commercial uses and therefore minimal sources in Elizabeth, NJ, while tetrachloroethene is the primary solvent used in the dry cleaning industry. No parameters, neither proximity nor meteorological variables, were include in the regression equations for carbon tetrachloride at a $p<0.5$ criteria indicating that no local sources nor distance to roadways influenced the variability in its measured concentration. This is consistent with the lack of local sources. Meteorological variables were included in the regression equation for tetrachloroethene, but not proximity to roadways or gasoline stations. Since tetrachloroethene is used in dry cleaning, the distance between the sampling locations and dry cleaning facilities were determine and evaluated in the regression equation. The final regression equation for tetrachloroethene included the inverse distance to dry cleaning facilities, atmospheric stability, temperature, wind speed and relative humidity with similar partial r^2 and coefficients for the meteorological identified for the regression equations for the compounds derived from mobile sources (Table 17).

To evaluate whether all particulate components might show associations with mobile sources, selenium an element measured in PM2.5 that is not expected to be associated with mobile sources was also examined. Its regression equation did not include proximity or meteorological to mobile sources.

Effects of Source Proximity

The common interpretation of a regression coefficient is that it estimates the change in the response variable per unit increase in the predictor variable. This estimation has limitations when the predictor variables are seriously intercorrelated. When highly correlated predictor variables vary together, the magnitude of the outcome variable

change with a single predictor variable is altered. Since the multicollinearity in the models was not serious for the immediate remedial measures (Appenix A), based on the previous diagnostics, the effect of individual parameter estimates on the concentrations were evaluated by holding all the variables constant except for the variable being evaluated. This approach allows for the model to be evaluated for the effect of a single variable across its range of values when considering all other variable to be constant. It is a type of sensitivity analysis. The assigned constant values used were the median value for the meteorological variables and the maximum value for the distance. The maximum value of the distance was used since the smallest changes in concentrations with distance would be expected at the furthest distance from the source. A plot of the predicted air concentration with distance for each aromatic compound, MTBE and the $PM_{2.5}$ components derived from the best fit regression equation are given in figures 11-19.

The shape of the decline with distance follows an exponential form since the regression equations included distance as an inverse term and the concentration was expressed as a log normal concentration. For the roadways (both FC11 and FC14) and gasoline stations, the decline in predicted concentration is rapid during the first 200m with little change due to roadways after that distance. The magnitude of this change between 20 meters, the distance of the closest samples, to 200 meters was a factor of approximately two for the PM2.5 constituents to four for the aromatic compounds and MTBE. The scatter plots of concentration with distance for the actual data (figure 20 to 45) are consistent with the rapid falloff in concentration with distance predicted by the regression equations, though the falloff may be at a slightly further distance, though the changes in concentrations appear to be small. The predicted effect for the PM area

sources of truck loading and unloading for $PM_{2.5}$ and elemental carbon is over a longer distance than the roadways. This is probably a statistical artifact the multiple area sources associated with truck loading and unloading that are all to the east/north east of Elizabeth and the difficulty in assigning the appropriate distance to the site since it covers a large area (figures 46 and 47). The roadways FC19 which are small local roads show a maximum effect on $PM_{2.5}$ mass at 10 to 20 meters. This is suspect as a true mobile source of $PM_{2.5}$ from streets in this category is minimal since the traffic is very light with little if any truck traffic. The scatter plot suggests that only a few data points are responsible for the observation so it may be a statistical associate with other causes. All homes will be close to roadway FC19, even if they are near major roadways as well, as is evident by the maximum distance to a roadway classified as FC19 for any home was less than 100 meters.

One meteorological parameter that we could not adequately incorporate in the models was wind direction. Several different approaches to examine wind direction were examined including categorizing wind direction based on the amount of variability in wind direction during the sampling period as well as evaluation of the dominant direction. However, the micrometeorology around the sample sites could not be definitively represented by the meteorological station at Newark Airport since directional changes are expected around buildings and roadways. Thus, the effect of wind direction could not be adequately represented in the model and therefore final models did not include that term.

The meteorological variables contributed more to the explanatory power of the regression equations than the proximity variables. One possible reason this is that there were more homes sampled at distances greater than 200 meters, than closer than 200 meters, the distance with the maximum predicted effect of the roadway. If only homes within 200 m of a major roadway were included in the study it is possible that the effect of proximity would be stronger. The regression equations suggest that the effect of distance due to mobile sources is minimal for homes further than 200 meters from major roadways and that concentration changes nears homes more than 200 meters from roadways or gasoline stations would be dependent upon meteorology which controls the urban background levels for constant emission sources within an urban center and transport of pollutants from outside Elizabeth, NJ. Exploratory analyses of only homes within 200 meters and homes within 500 meters of FC11, the NJ Turnpike, suggest that inverse distance to the NJ Turnpike was a potential predictor variable, but the term FC11 did not reach statistical significance at $p<.15$ for the aromatic compounds probably due to the small n in that sub-sample of homes.

None of the regression models for the three aldehyde compounds studied, formaldehyde, acetaldehyde and acrolein, included the inverse distance to any of the mobile source proximity terms, even though they are exhaust emission products. The positive association with the distance to FC11 roadways for formaldehyde implies that roadways are not a source of formaldehyde. It is more likely that it is a result of an association among FC11 distance, formaldehyde concentration and third variable not evaluated. It is possible though that photochemical production of formaldehyde increases

with distance from roadway in a manner similar to ozone, which is higher away from roadways than directly adjacent from roadways as there is time component to its maximum concentration. To attempt to evaluate whether the affect of proximity to roadways could be observed in the absence of photochemistry, regression equations were also determined for data when the mean temperature during sampling was <10ºC, days when photochemistry is expected to be minimal. Again, only meteorological variables were only included in the regression equation for formaldehyde and acetaldehyde with a $p<0.15$ (Table 17). These analyses had a smaller so had less statistical power to identify an association.

Summary

Mobile sources (cars, trucks and gasoline stations) are a main source for aromatic hydrocarbons, methyl *tert* butyl ether, and $PM_{2.5}$, elemental carbon and selected PAHs in Elizabeth, NJ. Meteorological factors, in particular atmospheric stability, wind speed and temperature were statistical predictors of the overall concentration of these pollutants in the ambient air surrounding homes in the area. The air concentrations at homes that were very close to roadways and gasoline stations within 200 to 500 meters, were inversely related to the distance to those sources. Increases in the concentrations for the closest residences are predicted to be factors of two to four above what might be considered the background levels for the area. Area sources that were associated with truck activity or possibly other mobile source (airport or shipping terminal) also appears to increase the PM levels associated with diesel emissions. These increases in ambient air for homes near ambient sources could potentially result in corresponding increases in personal exposure for individuals living in homes without smokers since the ambient air surrounding homes penetrates into the home and a strong association has been found between ambient air concentrations outside a portion of the homes studied during the RIOPA study with both indoor and personal air for these compounds (examples in Figures 48 to 50 – from Weisel et al. 2004b).

Table 16. Frequency of Non-detects, Outliers, and Non-outliers by Season

***m,p*-Xylene**	Non Detects		Outliers		Non-Outliers	
Season	N	%	N	%	N	%
Fall	1	25	4	30.77	51	30.72
Spring			1	7.69	36	21.69
Summer	1	25	6	46.15	46	27.71
Winter	2	50	2	15.38	33	19.88
***o*-Xylene**	Non Detects		Outliers		Non-Outliers	
Season	N	%	N	%	N	%
Fall			4	26.67	52	31.14
Spring			1	6.67	36	21.56
Summer	1	100	4	26.67	48	28.74
Winter			6	40	31	18.56
Toluene	Non Detects		Outliers		Non-Outliers	
Season	N	%	N	%	N	%
Fall			1	6.67	55	33.33
Spring			7	46.67	30	18.18
Summer	3	100	4	26.67	46	27.88
Winter			3	20	34	20.61
Benzene	Non Detects		Outliers		Non-Outliers	
Season	N	%	N	%	N	%
Fall			5	27.78	51	31.1
Spring			1	5.56	36	21.95
Summer			5	27.78	48	29.27
Winter	1	100	7	38.89	29	17.68
Ethylbenzene	Non Detects		Outliers		Non-Outliers	
Season	N	%	N	%	N	%
Fall	3	42.86	5	29.41	48	30.19
Spring			1	5.88	36	22.64
Summer	3	42.86	7	41.18	43	27.04
Winter	1	14.29	4	23.53	32	20.13
MTBE	Non Detects		Outliers		Non-Outliers	
Season	N	%	N	%	N	%
Fall			6	21.43	50	34.01
Spring	6	75	7	25	24	16.33
Summer	1	12.5	10	35.71	42	28.57
Winter	1	12.5	5	17.86	31	21.09
PCE	Non Detects		Outliers		Non-Outliers	
Season	N	%	N	%	N	%
Fall	5	33.33	3	23.08	48	30.97
Spring	1	6.67	5	38.46	31	20
Summer			3	23.08	50	32.26
Winter	9	60	2	15.38	26	16.77

Table 17. Summary of Finalized Best-fitting Models of Selected VOCs (-p<0.15 used as criteria for inclusion)

Pollutant Total r^2	**Row Heading**	**Intercept**	**Mobile/Area/Point Source**		**Meteorological Variables**			
m,p-Xylene	**X_i**	**β_0**	**FC14^{-1}**	**GS^{-1}**	**Stab**	**K**		
0.33	**β_i(SE)**	4.9(1.7)[b]	7.9(4.4)[d]	17.4(6.3)[b]	0.54(0.11)[a]	-0.02(0.005)[a]		
	P-r^2		0.01	0.04	0.18	0.09		
o-Xylene	**X_i**	**β_0**	**FC14^{-1}**	**GS^{-1}**	**Stab**	**K**	**U**	
0.42	**β_i(SE)**	4.5(1.4)[b]	7.4(4.5)[d]	9.5(5.5)[c]	0.52(0.09)[a]	-0.02(0.004)[a]	-0.12(0.04)[b]	
	P-r^2		0.01	0.02	0.27	0.09	0.04	
Toluene	**X_i**	**β_0**	**FC14^{-1}**		**Stab**	**K**		**RH**
0.31	**β_i(SE)**	3.1(1.8)[d]	14.7(4.4)[b]		0.71(0.12)[a]	-0.02(0.006)[b]		0.01(0.005)[c]
	P-r^2		0.04		0.22	0.03		0.02
Benzene	**X_i**	**β_0**	**FC14^{-1}**	**GS^{-1}**	**Stab**	**K**	**U**	
0.41	**β_i(SE)**	10.(1.3)[a]	10.1(1.5)[a]	5.5(3.3)[d]	16.1(5.6)[b]	0.30(0.10)[a]	-0.04(0.004)[a]	
	P-r^2			0.01	0.03	0.10	0.25	
Ethylbenzene	**X_i**	**β_0**	**FC14^{-2}**		**Stab**	**K**	**U**	
0.16	**β_i(SE)**	6.0(2.4)[c]	9.7(5.6)[c]		0.44(0.16)[b]	-0.03(0.007)[b]	-0.11(0.07)[c]	
	P-r^2		0.02		0.08	0.06	0.015	
MTBE	**X_i**	**β_0**	**FC14^{-2}**	**GS^{-1}**	**Stab**	**K**	**U**	
0.25	**β_i(SE)**	-2.7(2.3)[**]	22.3(14.3)[c]	33.6(8.3)[a]	0.24(0.15)[c]	0.01(0.007)[c]	-0.19(0.06)[a]	
	P-r^2		0.01	0.09	0.01	0.01	0.12	
PERC	**X_i**	**β_0**		**DCF^{-1}**	**Stab**	**K**	**U**	**RH**
0.31	**β_i(SE)**	2.5(1.3)[c]		32.7(12.4)[b]	0.14(0.09)[*]	-0.01(0.004)[b]	-0.14(0.04)[a]	0.01(0.003)[b]
	P-r^2			0.04	0.01	0.05	0.18	0.03
Formaldehyde	**X_i**	**β_0**				**K**	**U**	
0.15	**β_i(SE)**	9.2(1.1)				-0.11(.03)	-0.31(.22)	
	P-r^2					.094	.014	
Acetaldehyde	**X_i**	**β_0**				**K**	**U**	
0.13	**β_i(SE)**	2.2(0.3)				.023(.007)	-.13(.06)	
	P-r^2					.093	.036	
Acrolein	**X_i**	**β_0**			**MH**			
0.046	**β_i(SE)**	-.68(.33)			0.61(.32)			
	P-r^2				.046			

Pollutant Total r^2	Row Heading	Intercept	Mobile/Area/Point Source			Meteorological Variables		
$PM_{2.5}$	**X_i**	**β_0**	**$FC11^{-1}$**	**$FC19^{-1}$**	**$TRUCK^{-1}$**	**Stab**	**U**	
0.47	**β_i(SE)**	1.0(0.6)	20(11)	4.2(1.7)	51(30)	0.43(.09)	-0.13(.04)	
	P-r^2		.016	.052	.016	.32	.066	
EC	**X_i**	**β_0**	**REF^{-1}**		**$TRUCK^{-1}$**	**Stab**	**RH**	
0.40	**β_i(SE)**	-2.5(0.6)	630(26)		78(36)	0.32(0.13)	.011(.004)	
	P-r^2		.033		.050	.078	.24	
OC	**X_i**	**β_0**	**$FC11^{-1}$**			**Stab**	**Precip**	
0.33	**β_i(SE)**	-2.2(.7)	39(21)			0.66(0.14)	.013(.006)	
	P-r^2		.043			.25	.035	
Coronene	**X_i**	**β_0**	**$FC11^{-1}$**			**Stab**	**K**	**U**
0.67	**β_i(SE)**	24(4)	133(42)			0.81(0.28)	-0.10(.01)	-0.41(.10)
	P-r^2		0.091			.06	.28	.24
Benzo[ghi]-pyrene	**X_i**	**β_0**	**$FC11^{-1}$**			**Stab**	**K**	**U**
0.66	**β_i(SE)**	22(4)	123(38)			0.71(.26)	-0.087(.01)	-0.38(.10)
	P-r^2		.094			.060	.26	.25
Sulfur	**Xi**	**β_0**	**O_3**			**Stab**	**K**	**U**
0.52	**β i(SE)**	4.3(2.2)	29(5)			0.44(.11)	-0.023(.008)	-0.11(.05)
	P-r2		.090			.14	.23	.039
Selenium	**Xi**	**β_0**	**O_3**			**Stab**	**K**	**U**
0.41	**β i(SE)**	1.2(3.5)	27(9)			0.93(.19)	-0.15(.08)	-0.019(.007)
	P-r2		.059			.29	.029	.034
Analysis of aldehyde data for days when the temperature was <10°C, to evaluate role of photochemistry								
Formaldehyde	**X_i**	**β_0**				**K**	**U**	**MH**
0.13	**β_i(SE)**	1.5(0.5)				-0.03(.02)	-0.13(.07)	0.43(.23)
	P-r^2					.063	.007	.03
Acetaldehyde	**X_i**	**β_0**				**K**		
0.13	**β_i(SE)**	2.2(0.3)				-0.03(0.02)		
	P-r^2					0.068		
Acrolein	**X_i**							
0.046	**β_i(SE)**							
	P-r^2							

X_i, *i* th predictor variable; β_0, intercept of model; β_i, parameter estimate of *i* th predictor; SE, standard error of parameter estimates; r^2, coefficient of determination; P-r^2, Partial r square of the variable.
$^{-1}$, indicates inverse values; $^{-2}$, indicates inverse square values.
$p<0.15$ used as selection criteria for inclusion of a variable in the model
FC14^{-1} is the inverse distance (m) to the nearest major arterial roadways
GS^{-1} is the inverse distance (m) to the nearest gasoline station
PS^{-1} is the inverse distance (m) to a point source (Linden Refinery)
DCF^{-1} is the inverse distance (m) to the nearest dry cleaning facility
TRUCK^{-1} is the inverse distance (m) to the major truck loading areas
Airport $^{-1}$ is the inverse distance (m) to Newark International Airport
Stab is atmospheric stability
K is temperature (ºK)
U is the wind speed (m
MH is the mixing height (km)
Precip is precipitation (total mm)

Regression Model Predictions Figures 10 – 20

There figures show the change in concentration as predicted by the regression models while varying the variable indicated from the minimum to maximum value observed during the study and holding all other variables in the model constant (median value for the meteorological variables or maximum value for the distance variables). The side bar is a box and whisker plot of the measured concentrations during the study (mean, median, 5^{th}, 25^{th}, 75^{th}, 95^{th} percentiles) for comparison.

Scatter Plots of Distance to Concentration Figures 21-45

These figures are the scatter plots of the concentration measured with the determined nearest distance between each home and the nearest roadway in each class or gasoline station for all values in the study. The figures provide a visualization of the association concentration with distance to mobile sources without consideration for meteorology, a major factor that influences concentration.

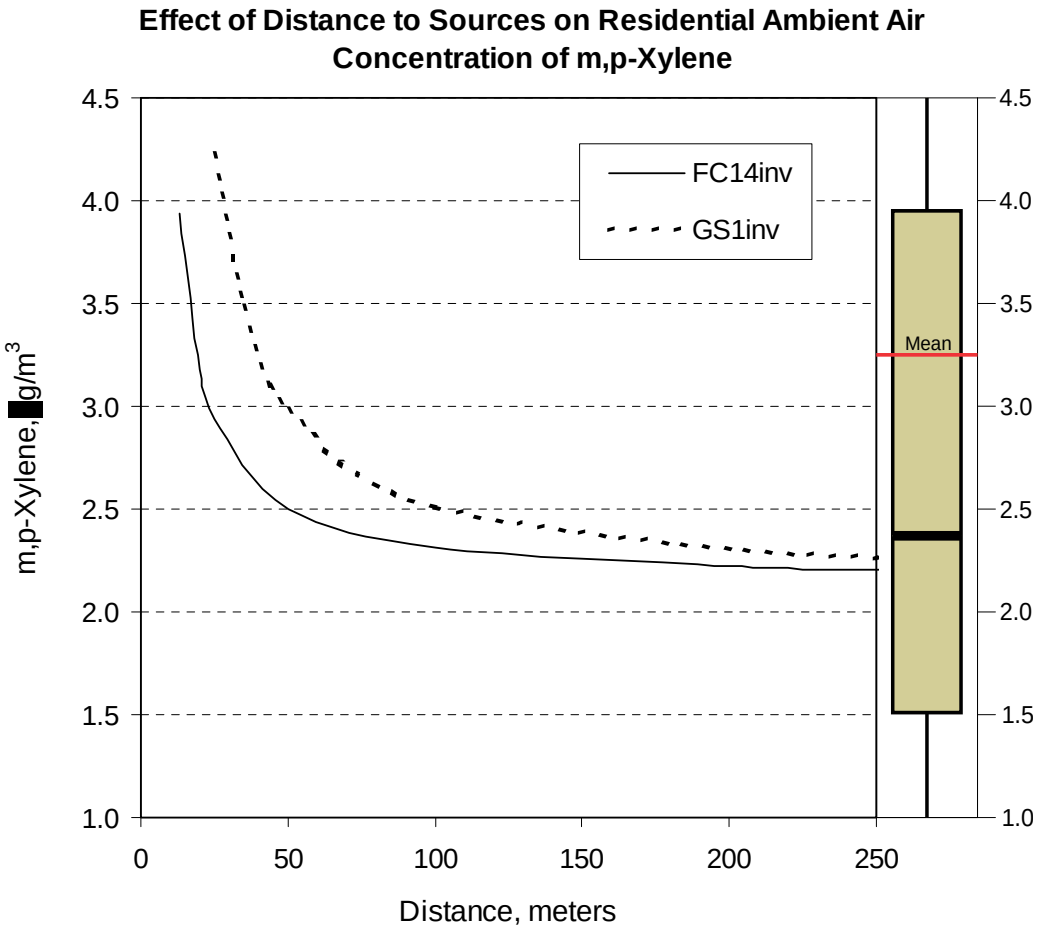

Figure 10: Effect of the Distance to the Emission Sources on the Residential Ambient Air Concentration of *m,p*-Xylene Estimated by the Best-fitting Model (Box Plot Shows Mean and Quartiles of Distribution of *m,p*-Xylene Concentrations)

.

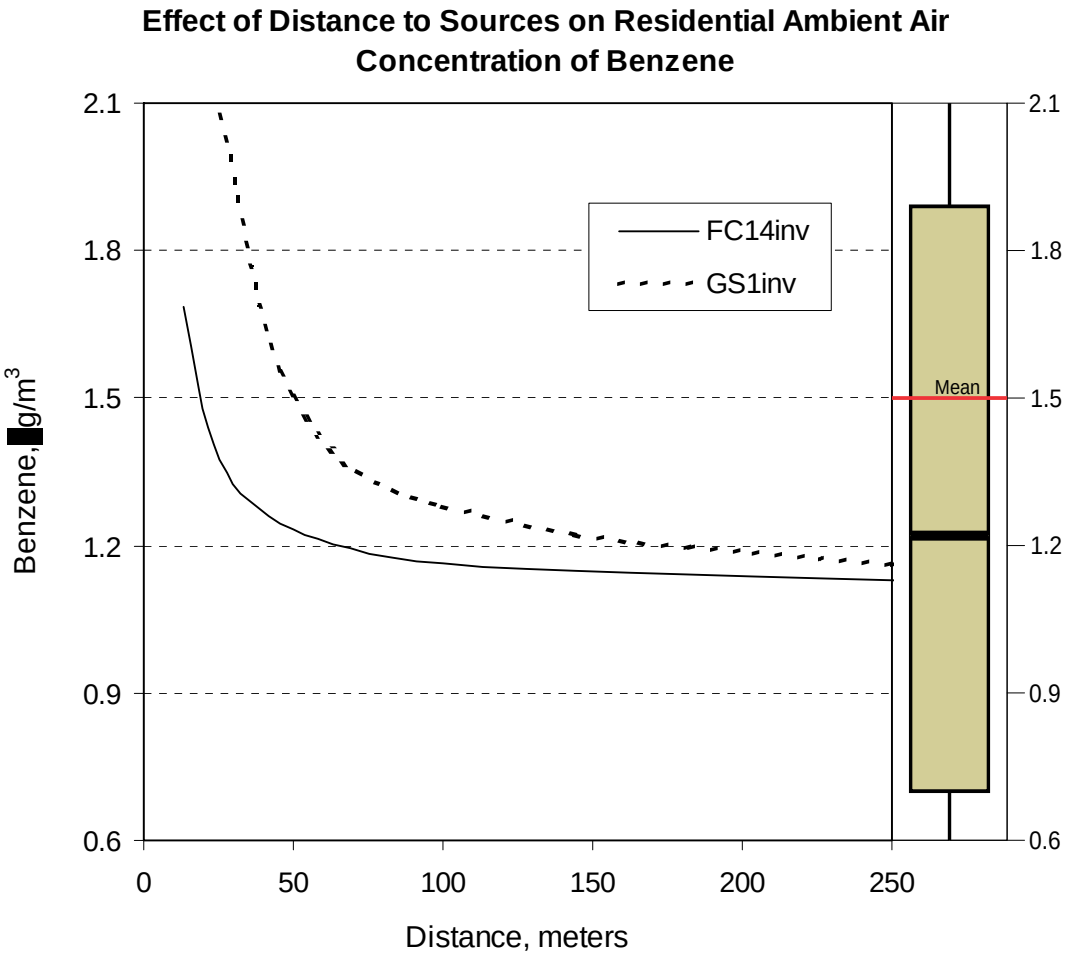

Figure 11: Effect of the Distance to the Emission Sources on the Residential Ambient Air Concentration of Benzene Estimated by the Best-fitting Model (Box Plot Shows Mean and Quartiles of Distribution of Benzene Concentrations).

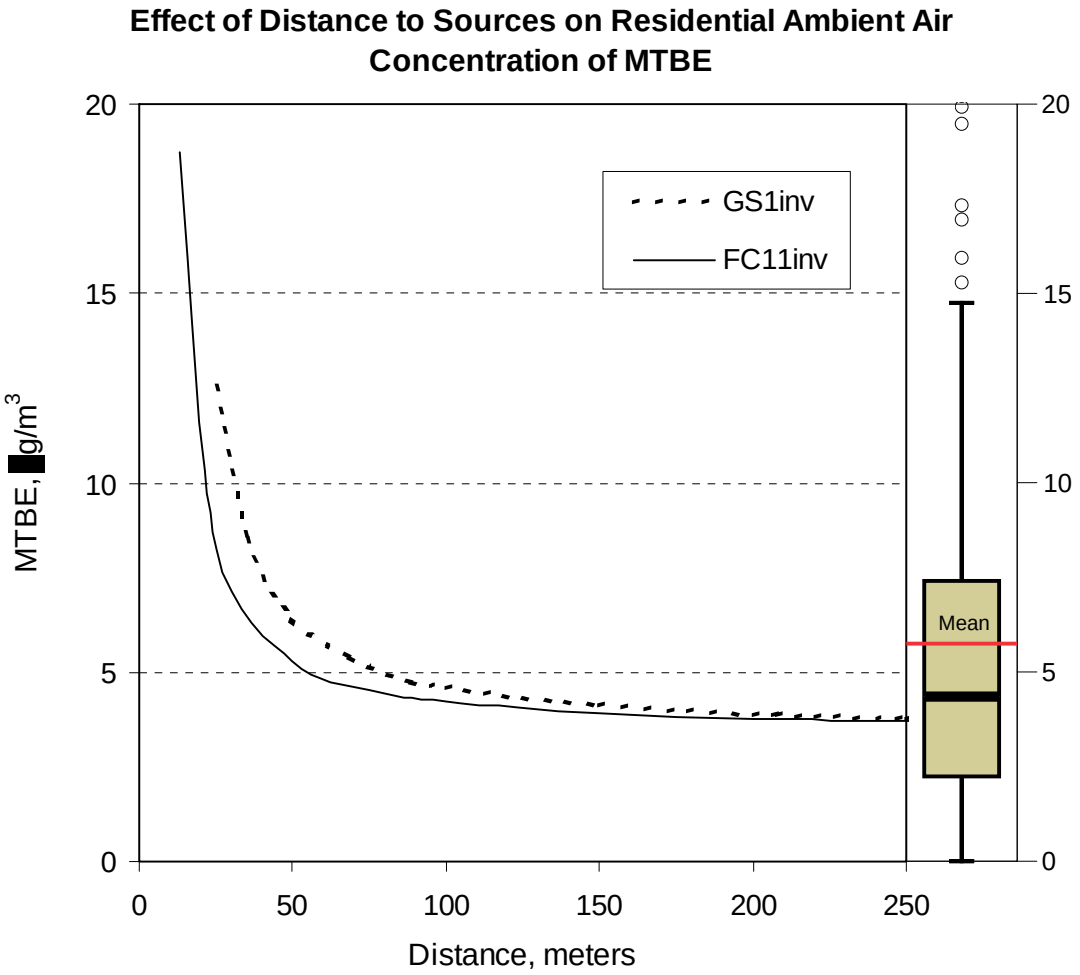

Figure 12: Effect of the Distance to the Emission Sources on the Residential Ambient Air Concentration of MTBE Estimated by the Best-fitting Model (Box Plot Shows Mean and Quartiles of Distribution of MTBE Concentrations).

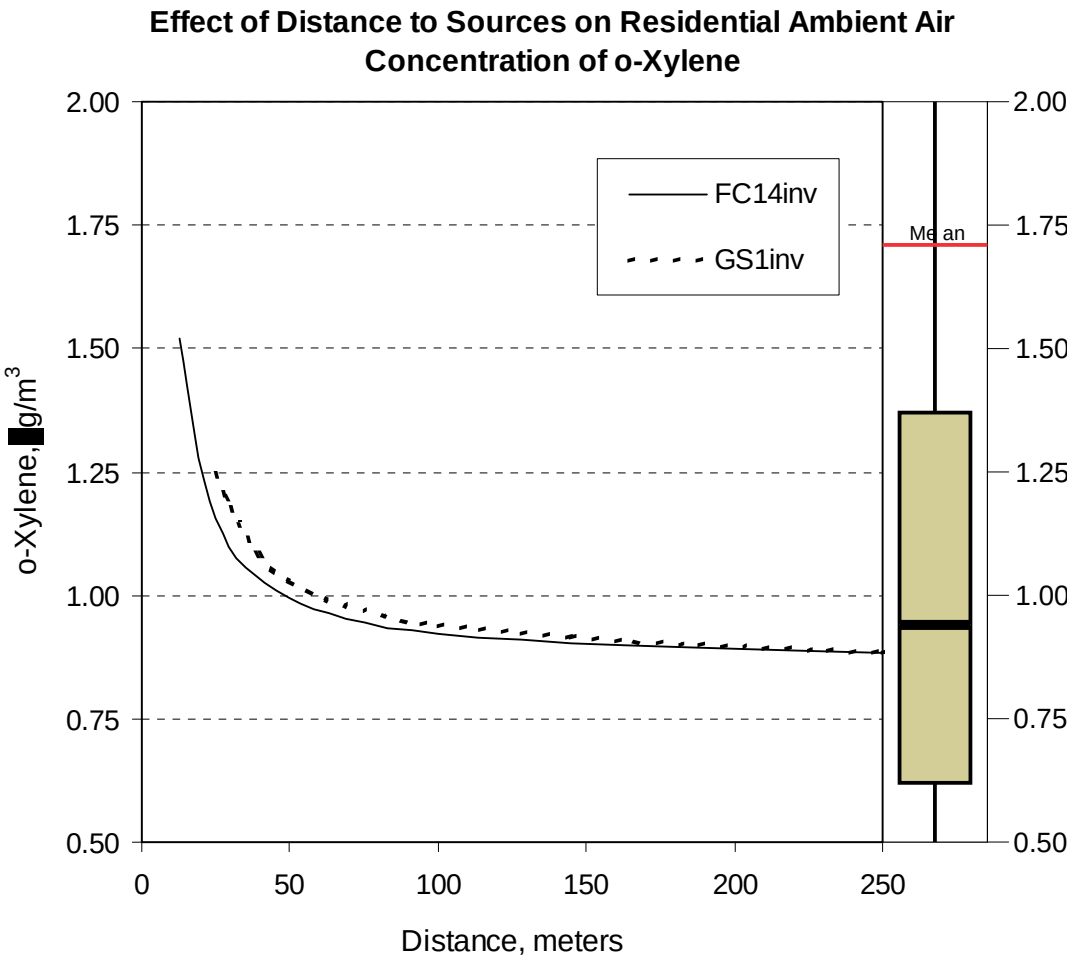

Figure 13: Effect of the Distance to the Emission Sources on the Residential Ambient Air Concentration of *o*-Xylene Estimated by the Best-fitting Model (Box Plot Shows Mean and Quartiles of Distribution of *o*-Xylene Concentrations).

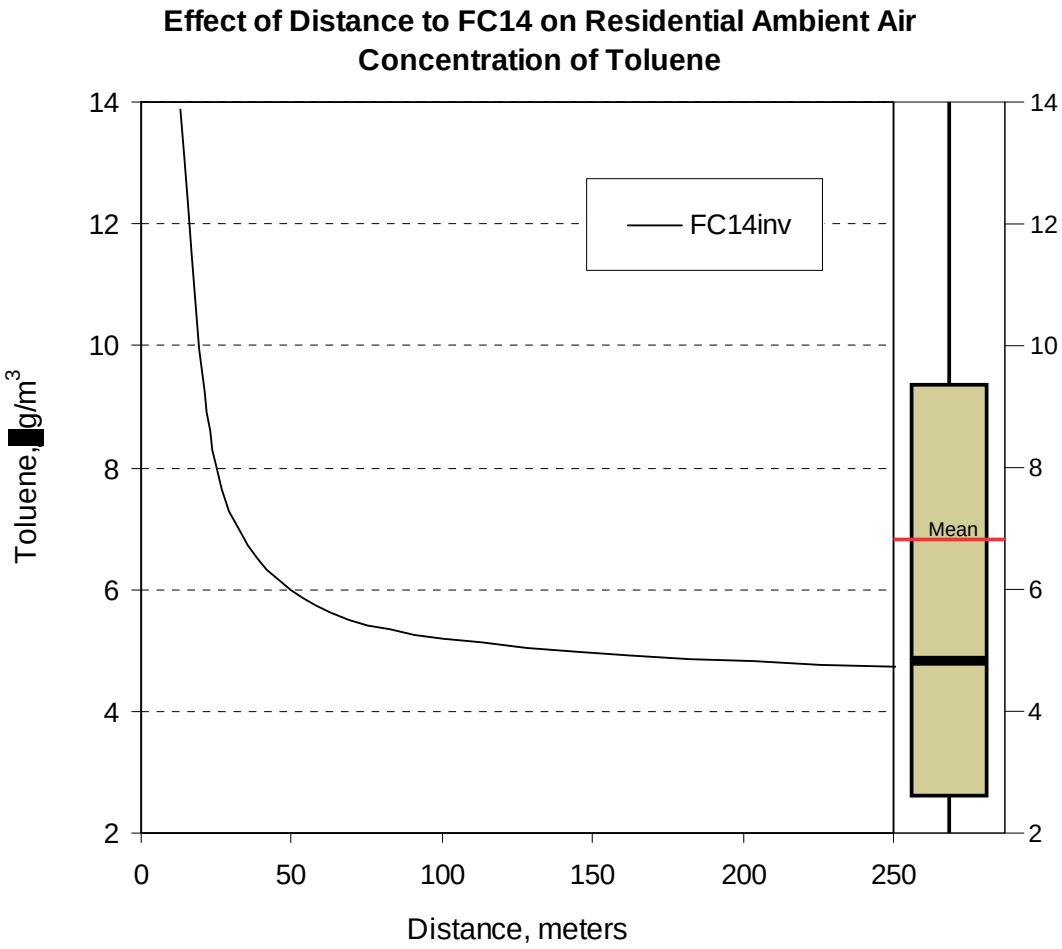

Figure 14: Effect of the Distance to the Emission Sources on the Residential Ambient Air Concentration of Toluene Estimated by the Best-fitting Model (Box Plot Shows Mean and Quartiles of Distribution of Toluene Concentrations)

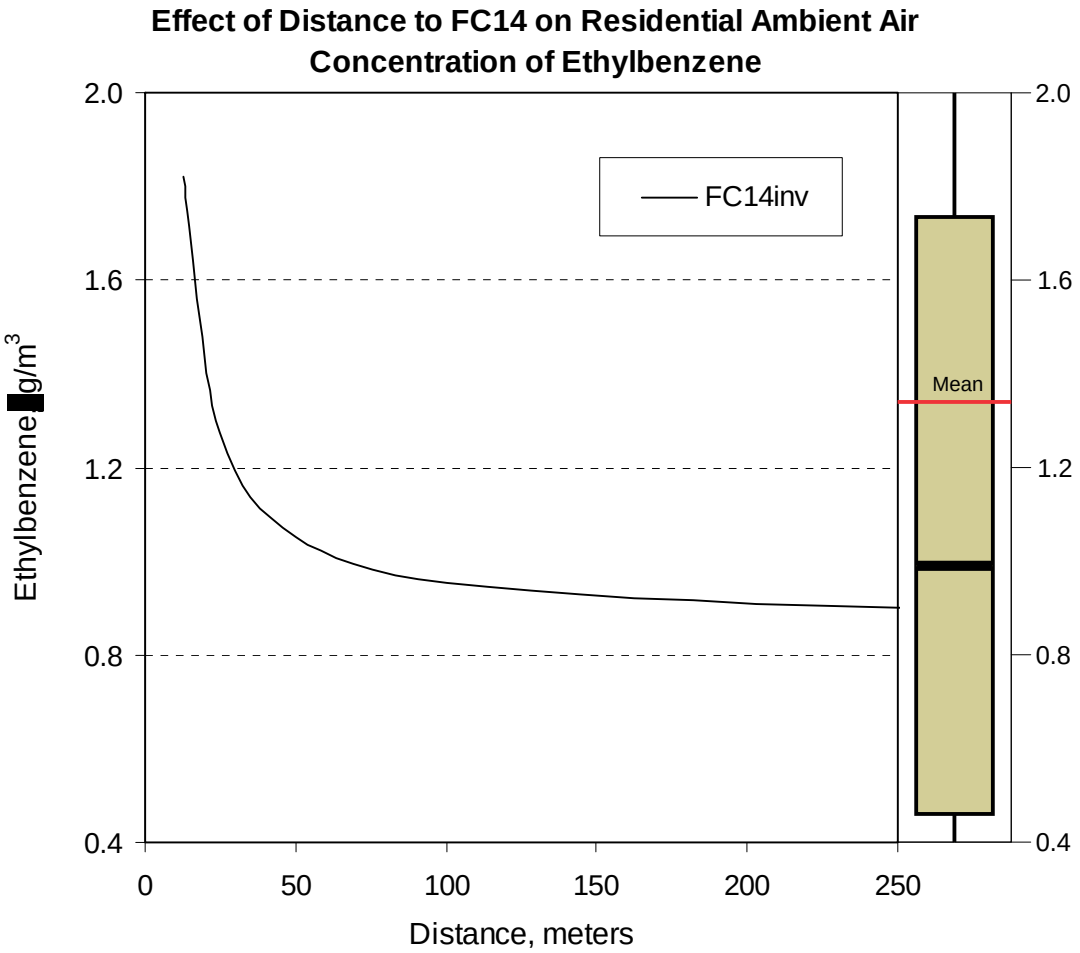

Figure 15: Effect of the Distance to the Mobile Source Emission on the Residential Ambient Air Concentration of Ethylbenzene Estimated by the Best-fitting Model (Box Plot Shows Mean and Quartiles of Distribution of Ethylbenzene Concentrations)

.

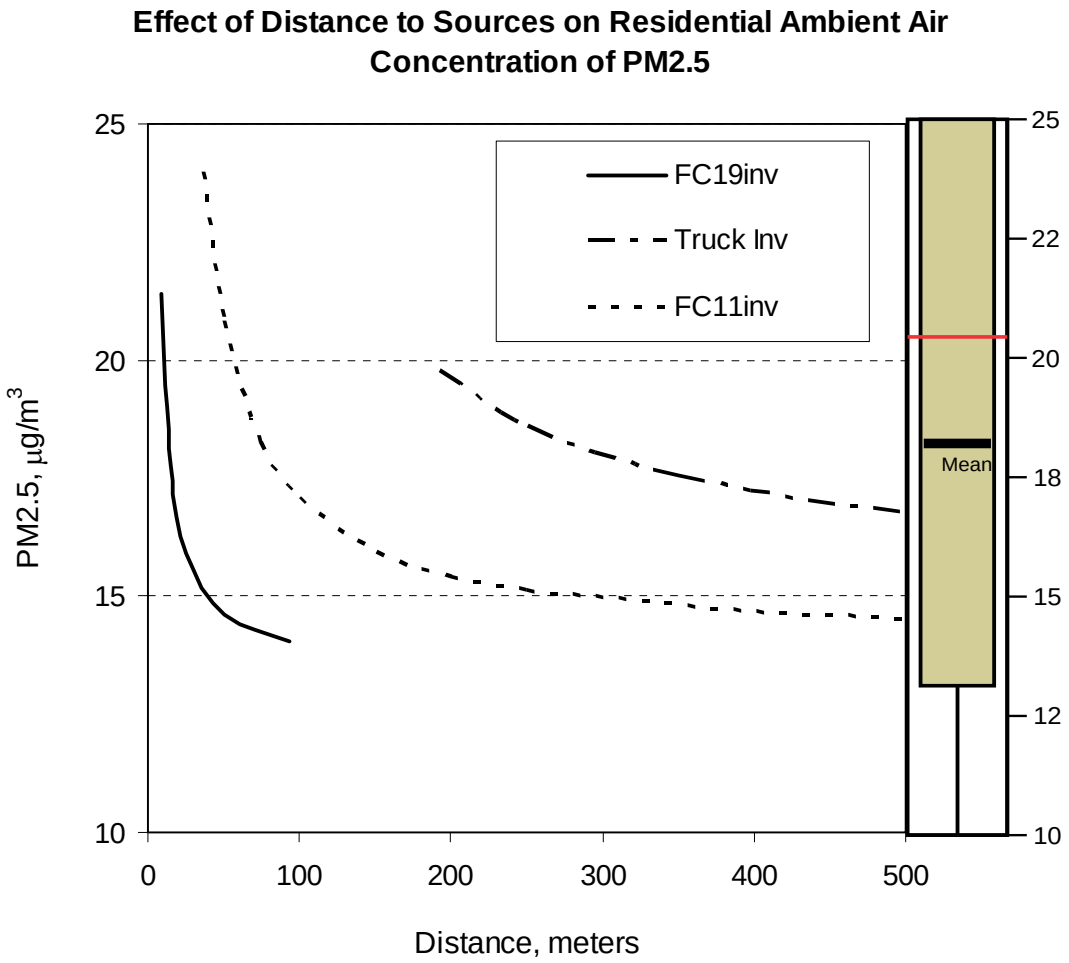

Figure 16: Model Prediction Of $PM_{2.5}$ Concentration With Distance To F11, F19 And Truck Loading And Unloading Region Estimated By The Best-Fitting Model (Box Plot Shows Mean And Quartiles Of Distribution Of $PM_{2.5}$ Concentrations)

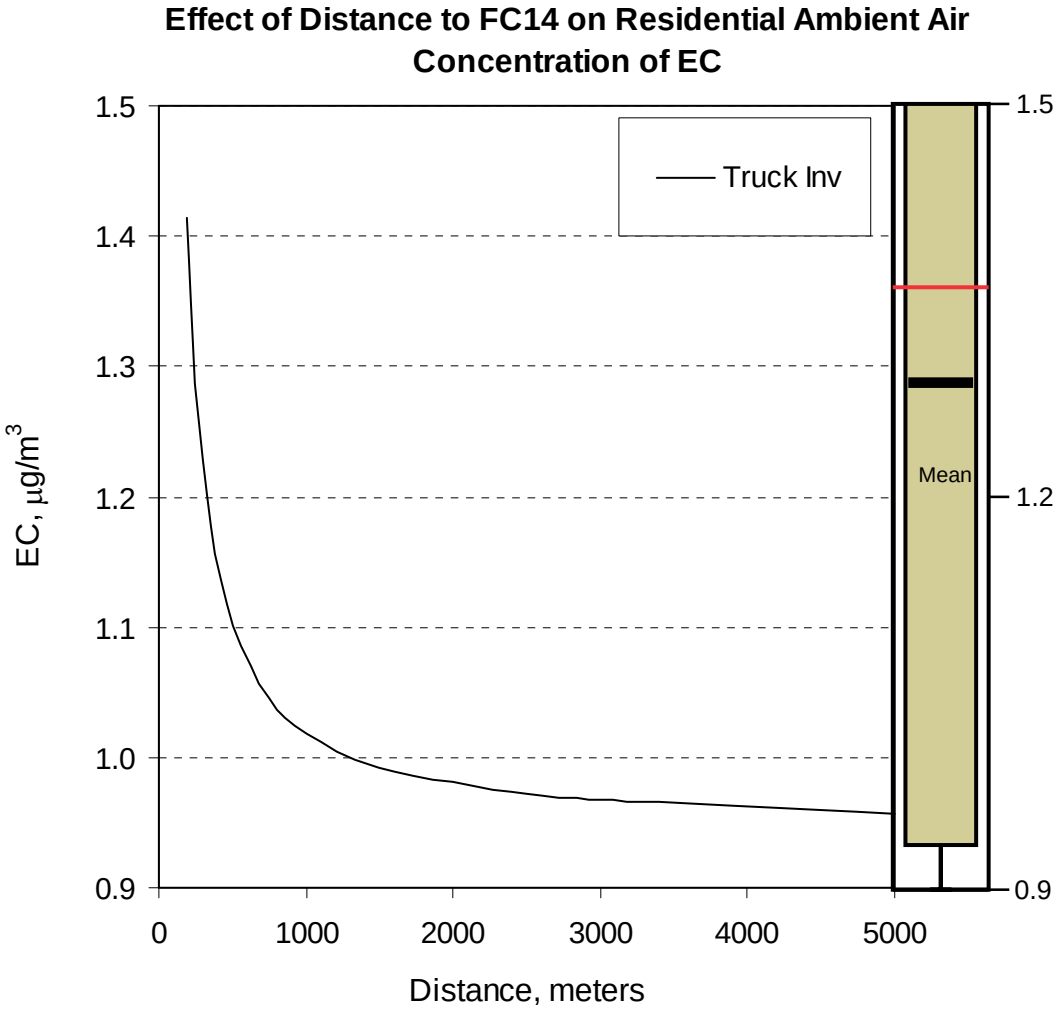

Figure 17: Model prediction of elemental carbon concentration with distance to Truck loading/dock area Estimated By The Best-Fitting Model (Box Plot Shows Mean And Quartiles Of Distribution Of Elemental Carbon Concentrations)

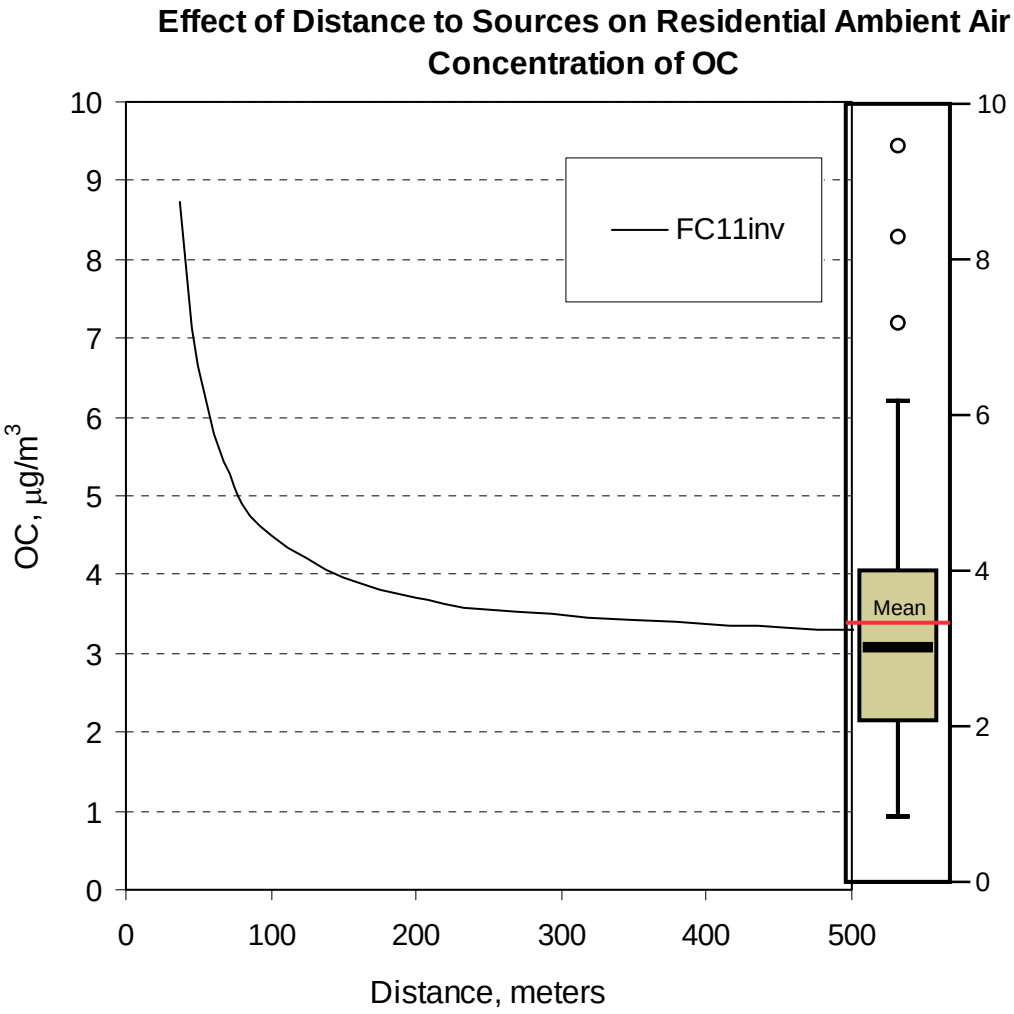

Figure 18: Model Prediction Of Organic Carbon Concentration With Distance To F11 Roadways Estimated By The Best-Fitting Model (Box Plot Shows Mean And Quartiles Of Distribution Of Organic Carbon Concentrations)

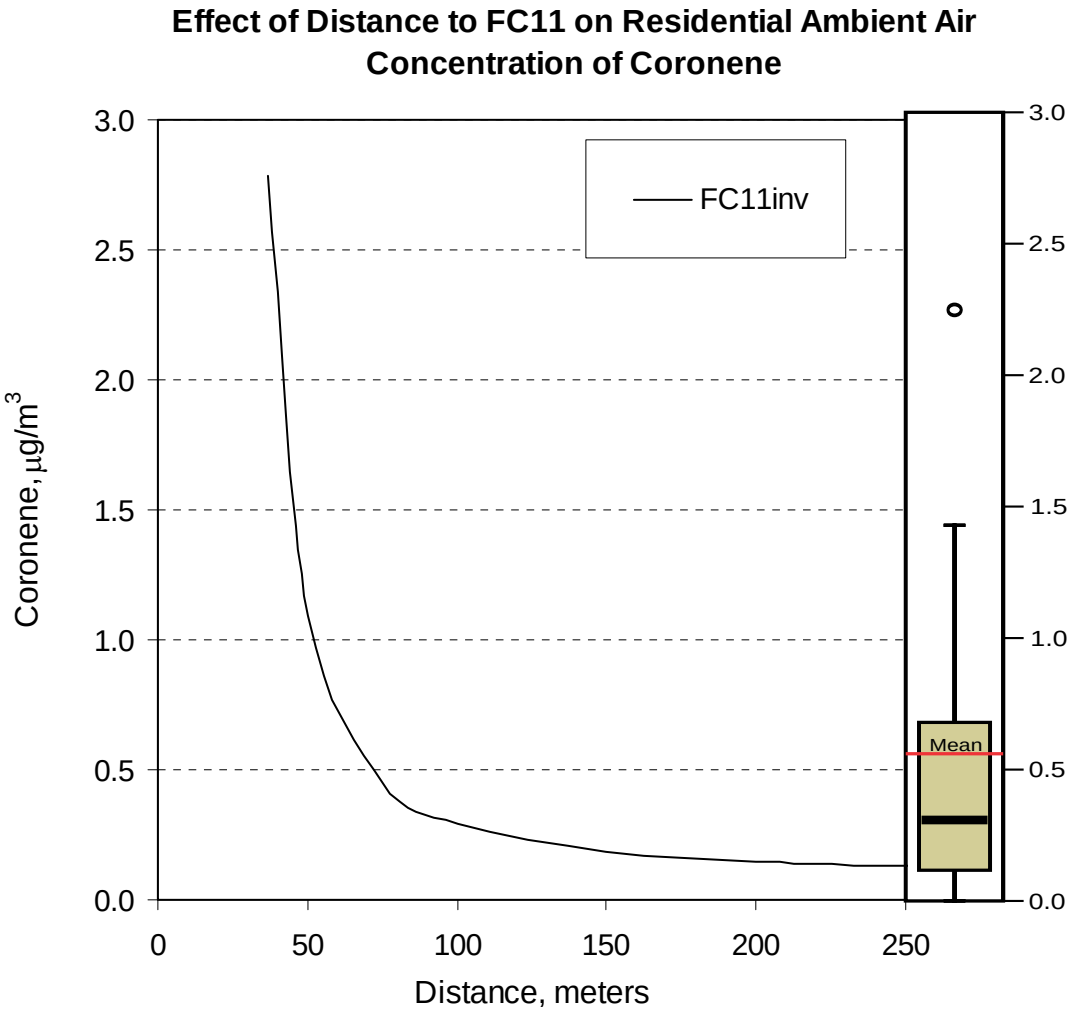

Figure 19: Model Prediction Of Coronene Concentration With Distance To F11. Estimated By The Best-Fitting Model (Box Plot Shows Mean And Quartiles Of Distribution Of Coronene Concentrations)

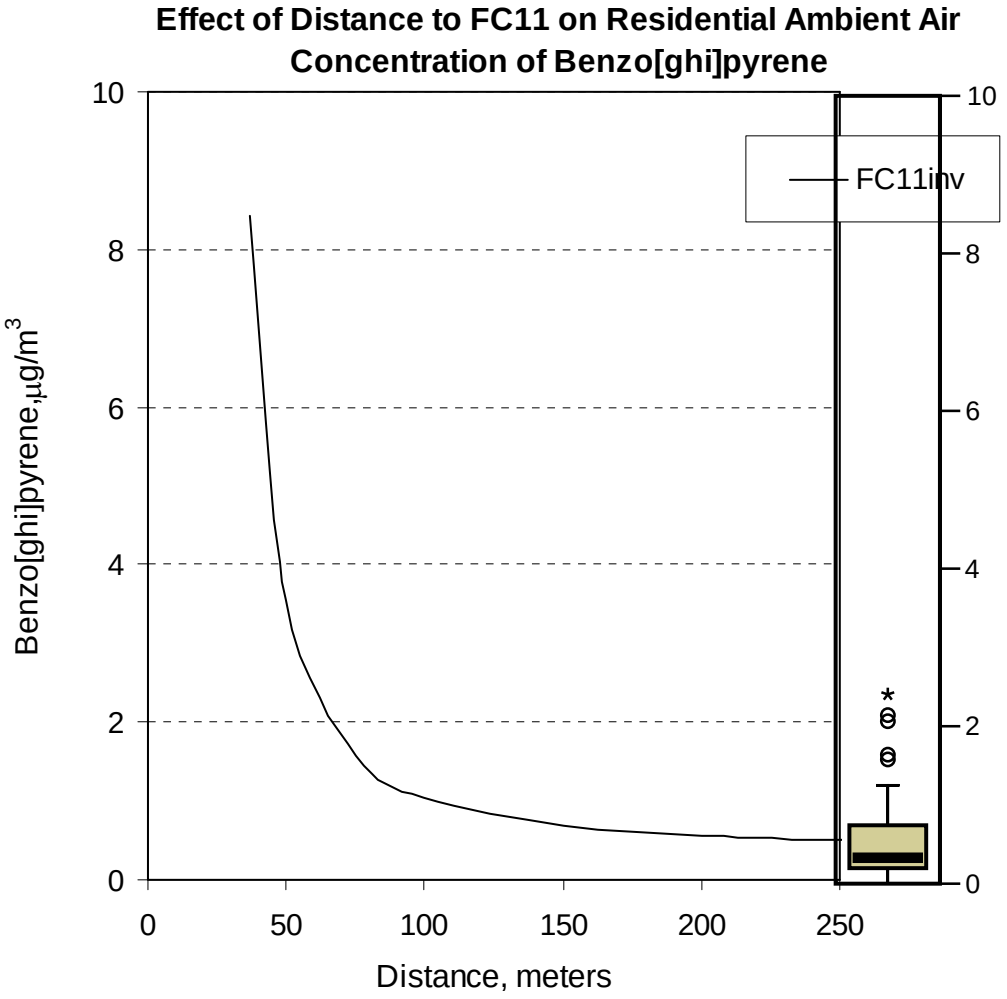

Figure 20: Model Prediction Of Benzo[ghi]pyrene Concentration With Distance To F11. Estimated By The Best-Fitting Model (Box Plot Shows Mean And Quartiles Of Distribution Of Benzo[ghi]pyrene Concentrations)

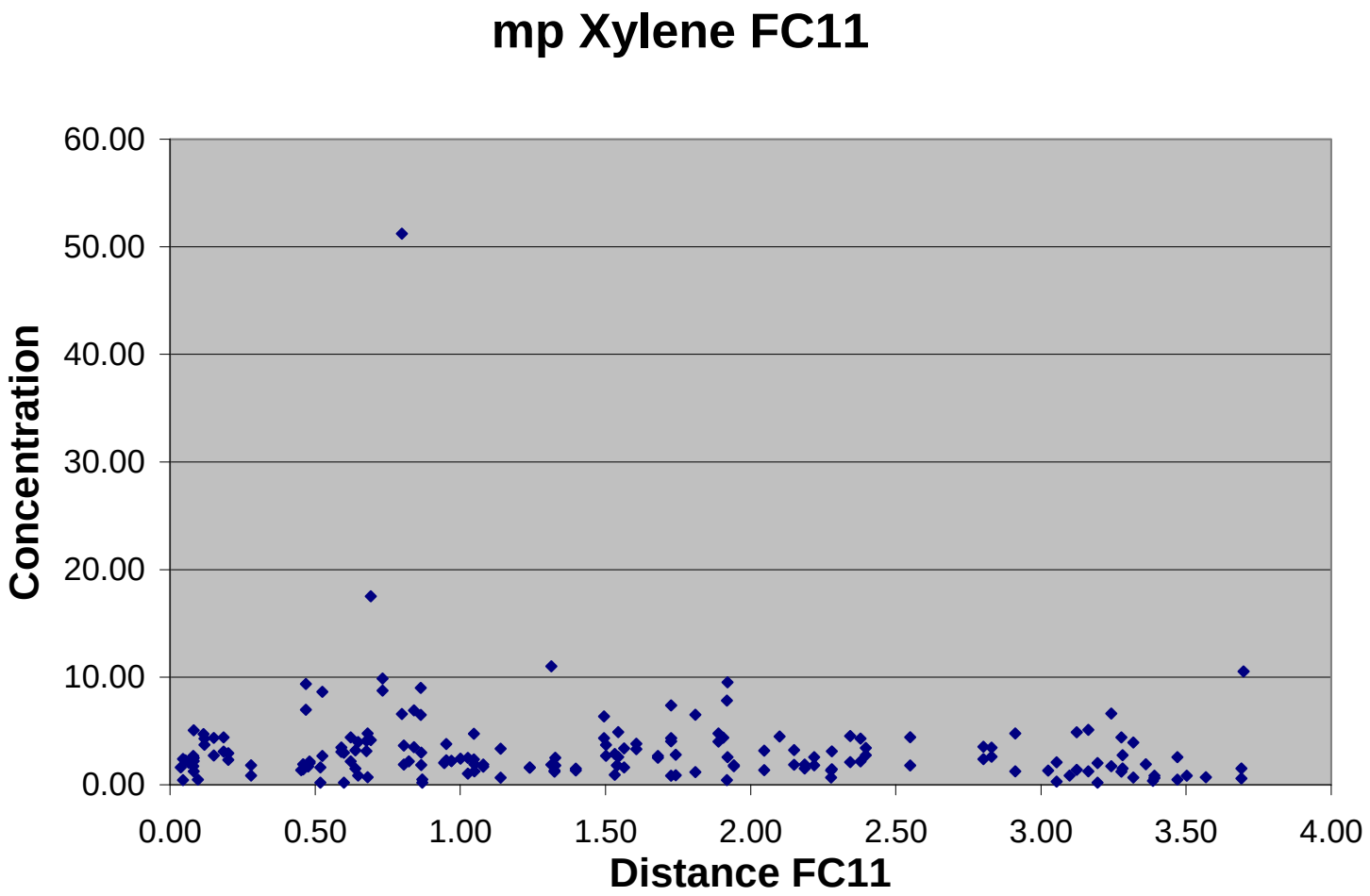

Figure 21. Scatter plot of *m/p* xylene with distance from FC11 Roadways, major urban arterial.

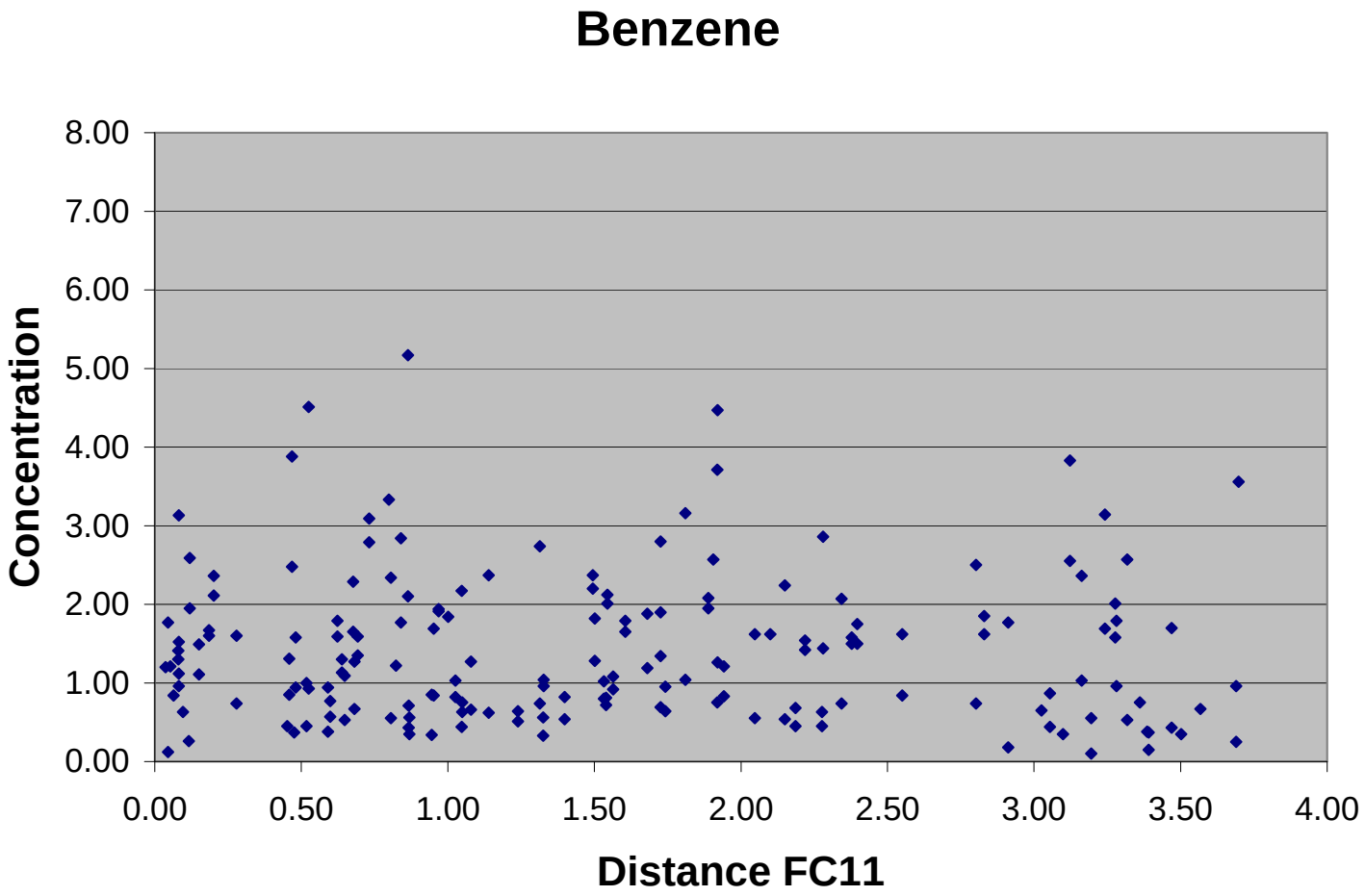

Figure 22. Scatter plot of benzene with distance from FC11 Roadways, major urban arterial.

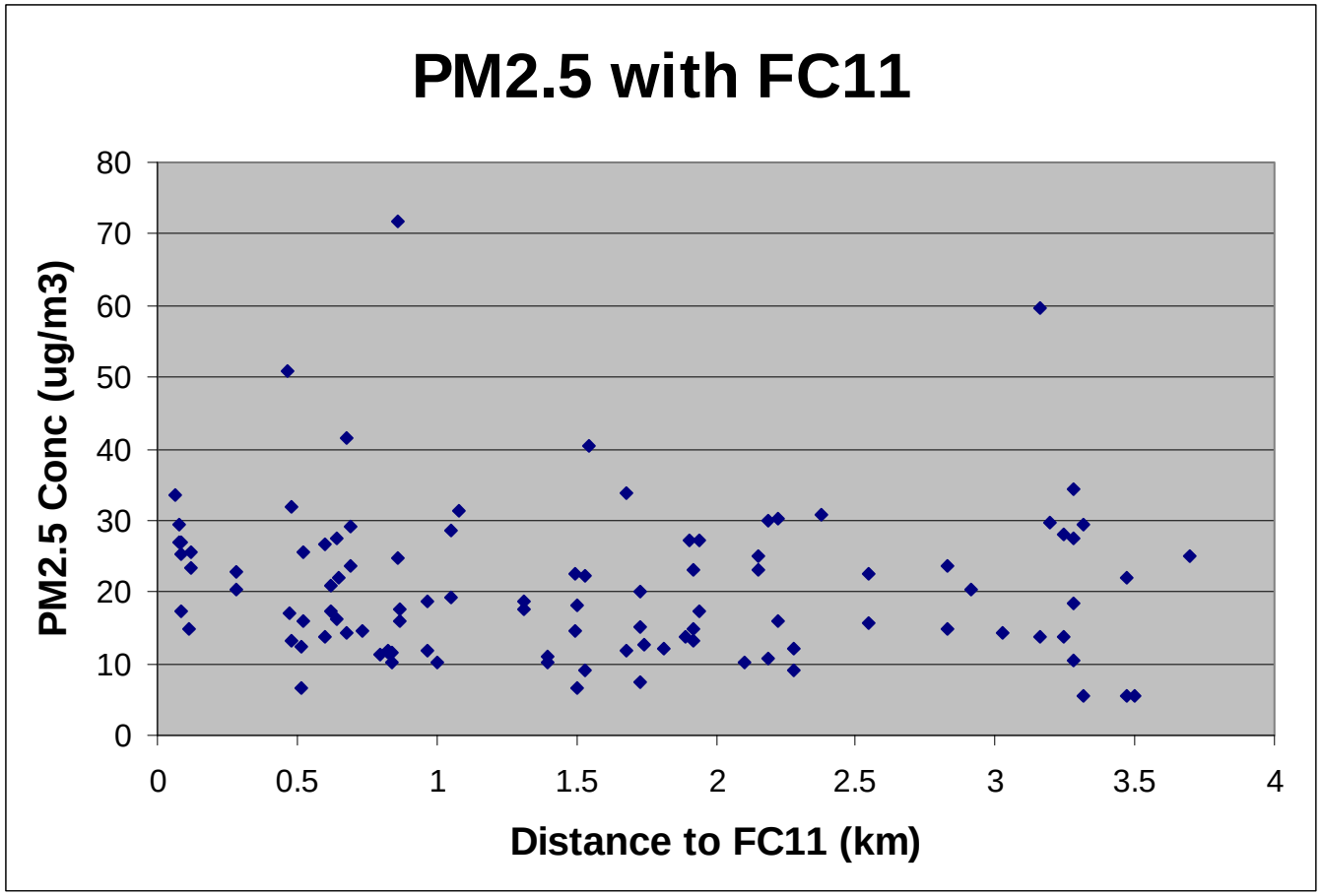

Figure 23. Scatter plot of $PM_{2.5}$ with distance from FC11 Roadways, major urban arterial.

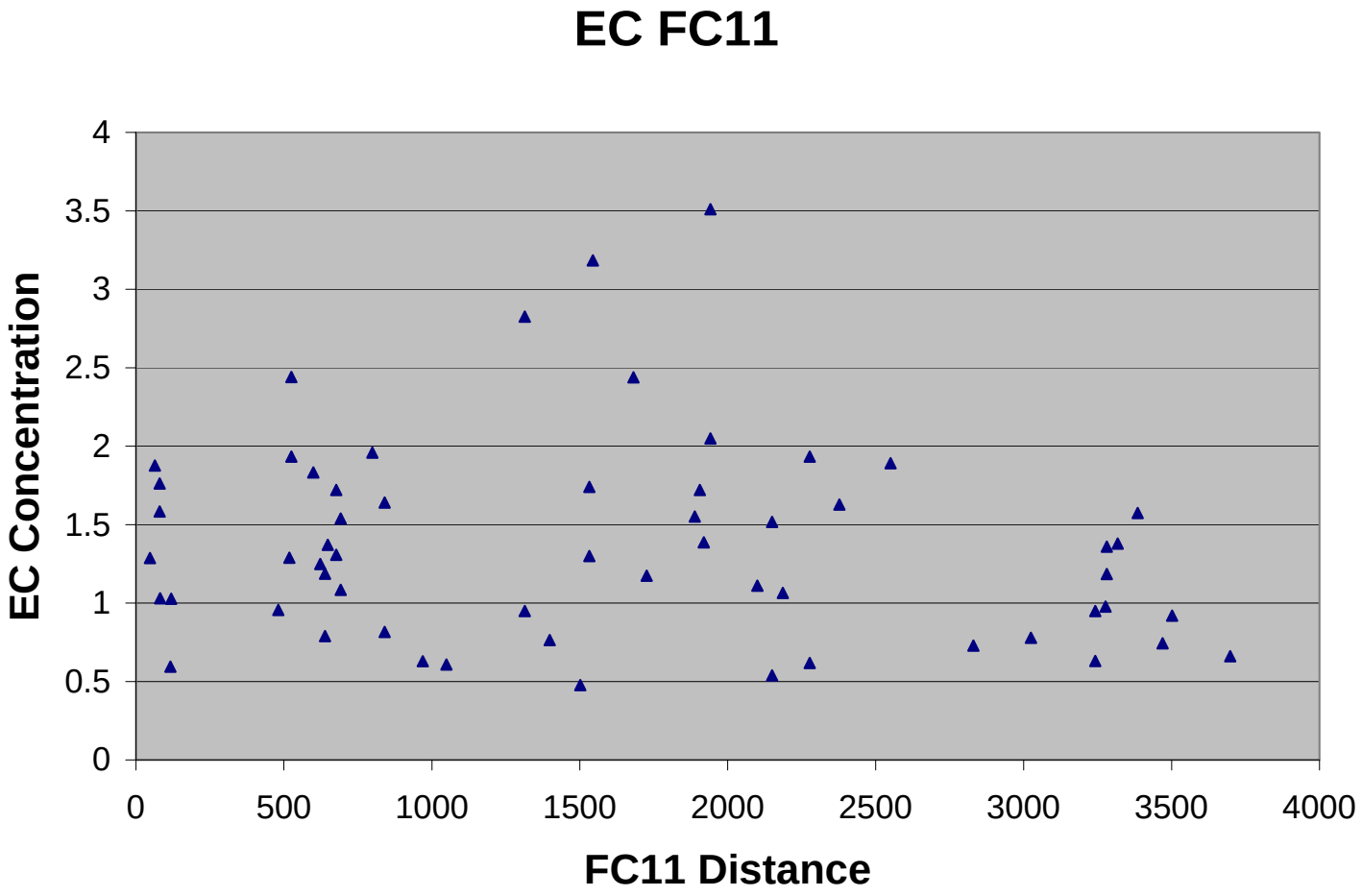

Figure 24. Scatter plot of elemental carbon with distance from FC11 Roadways, major urban arterial.

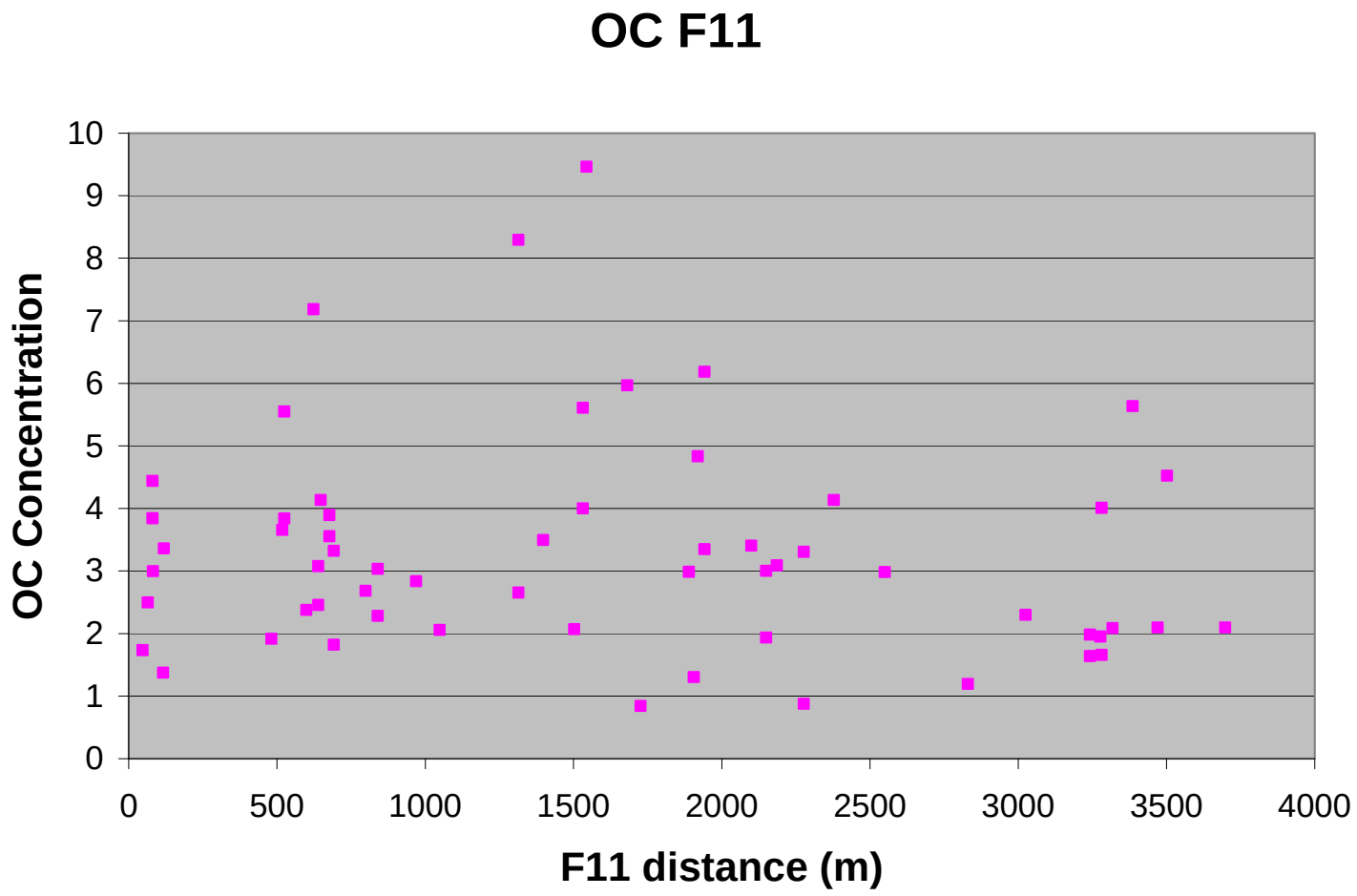

Figure 25. Scatter plot of organic carbon with distance from FC11 Roadways, major urban arterial.

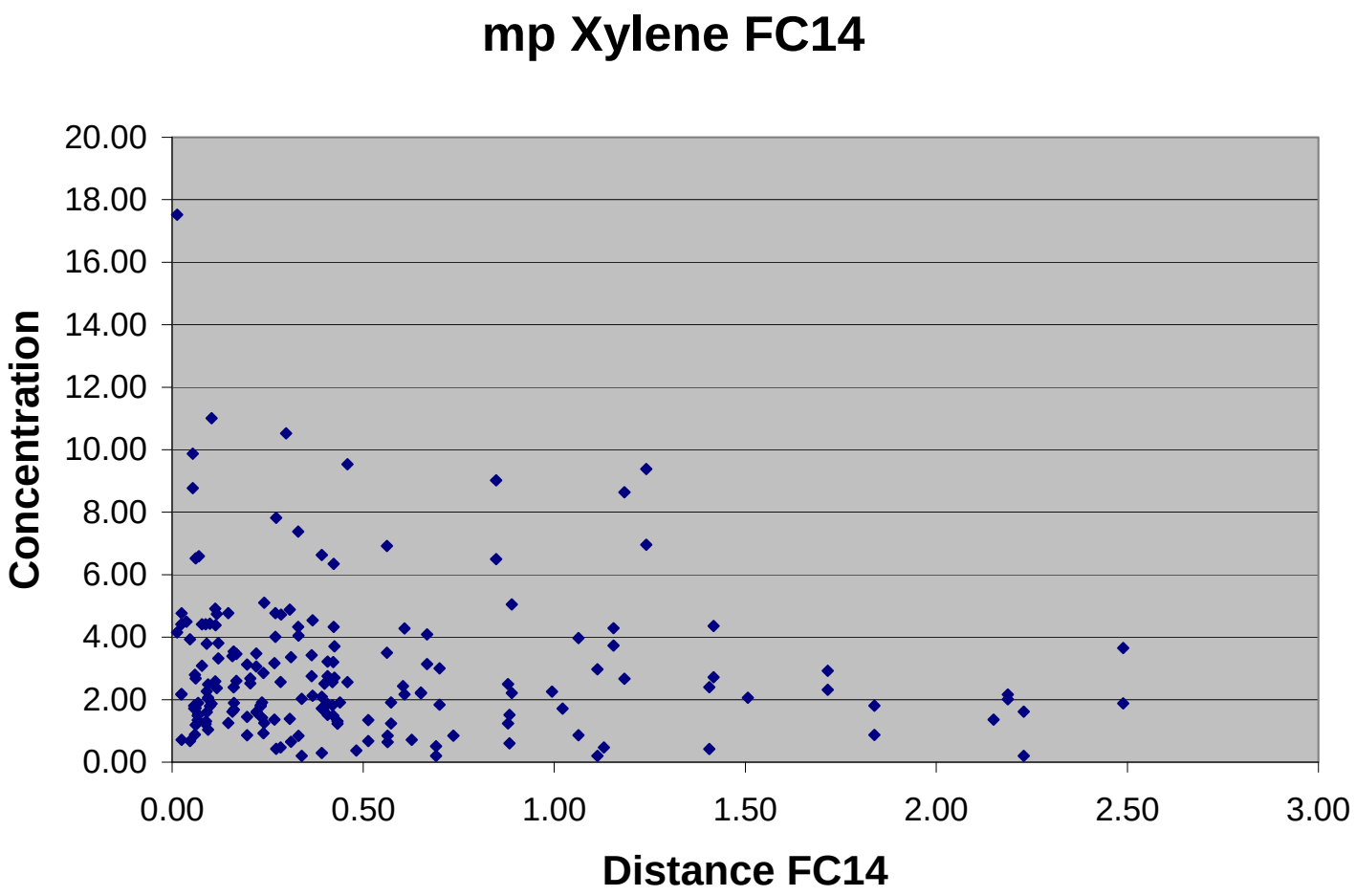

Figure 25. Scatter plot of *m/p* xylene with distance from FC14 Roadways, interstate.

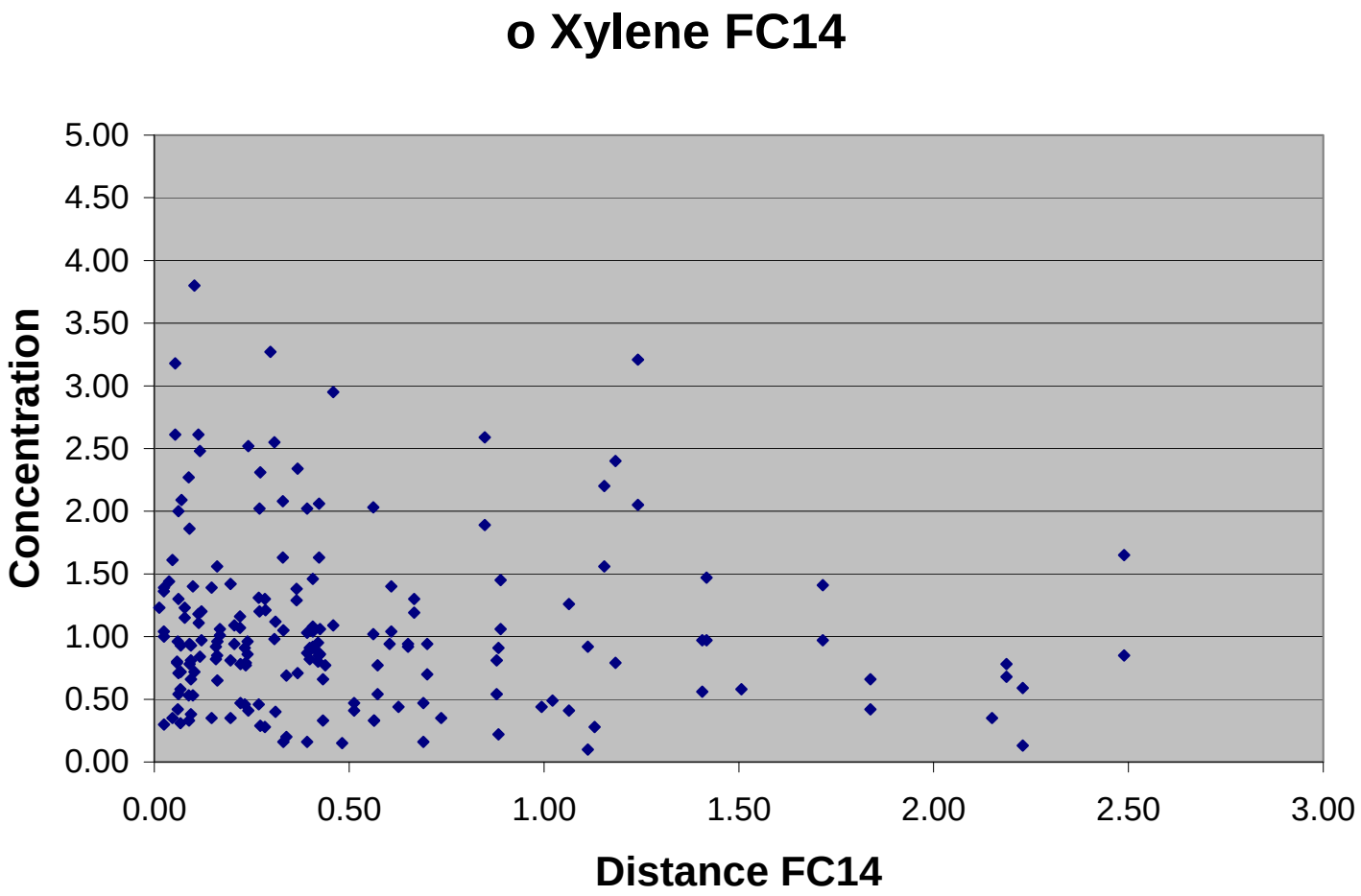

Figure 26. Scatter plot of *o* xylene with distance from FC14 Roadways, interstate.

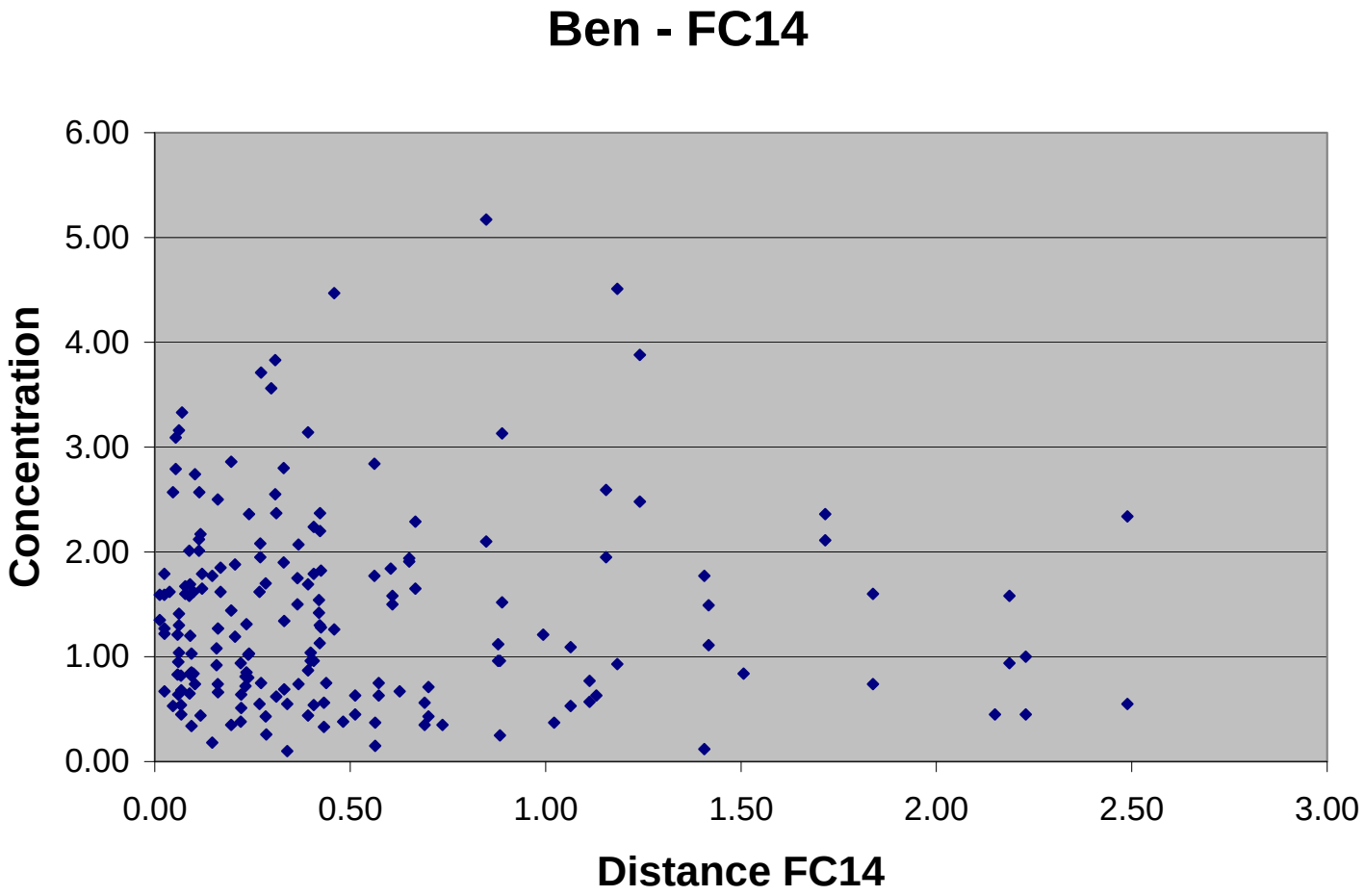

Figure 27. Scatter plot of benzene with distance from FC14 Roadways, interstate.

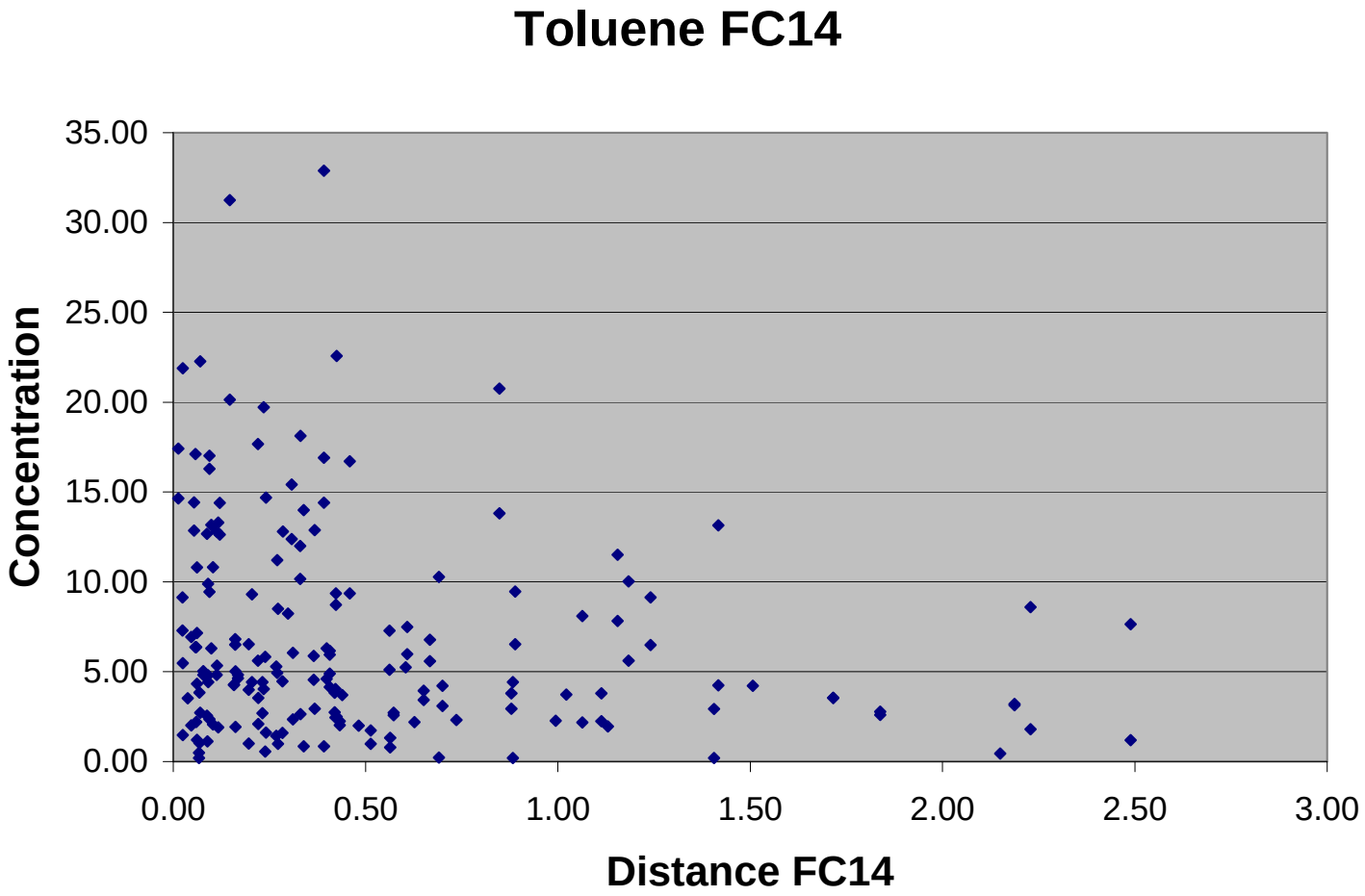

Figure 29 Scatter plot of toluene with distance from FC14 Roadways, interstate.

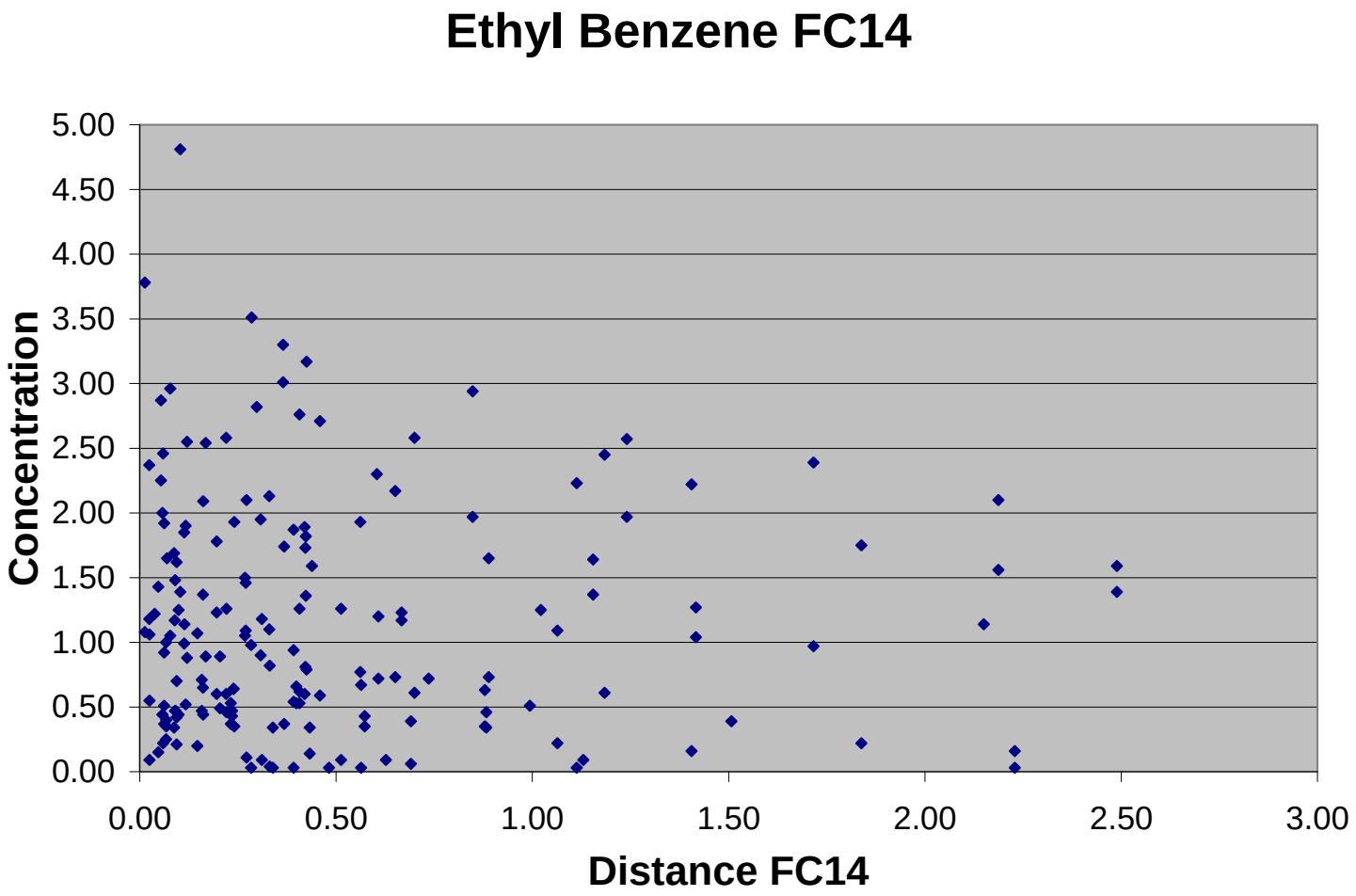

Figure 30. Scatter plot of ethyl benzene with distance from FC14 Roadways, interstate.

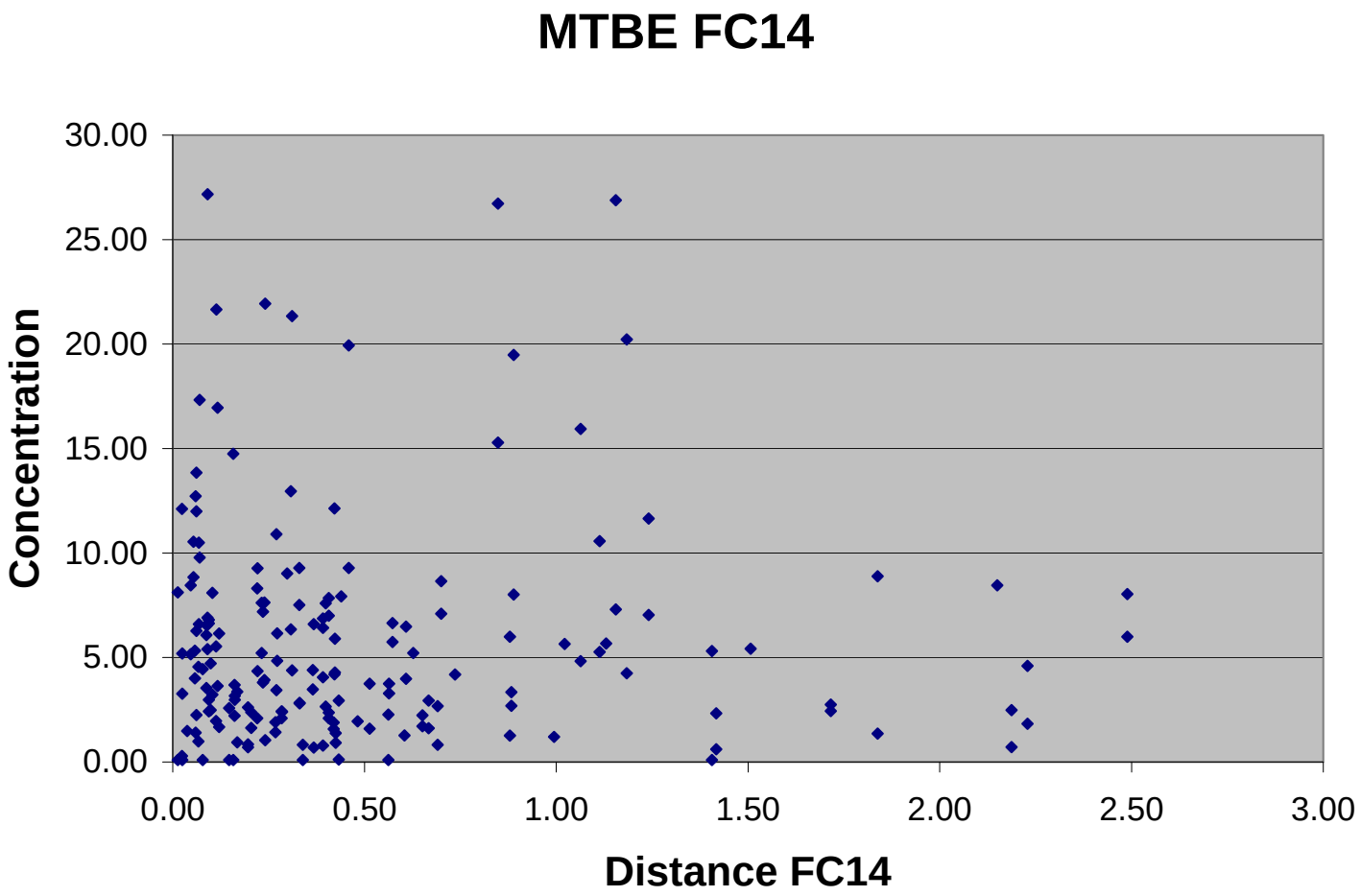

Figure 31. Scatter plot of methyl *tert* butyl ether (MTBE) with distance from FC14 Roadways, interstate.

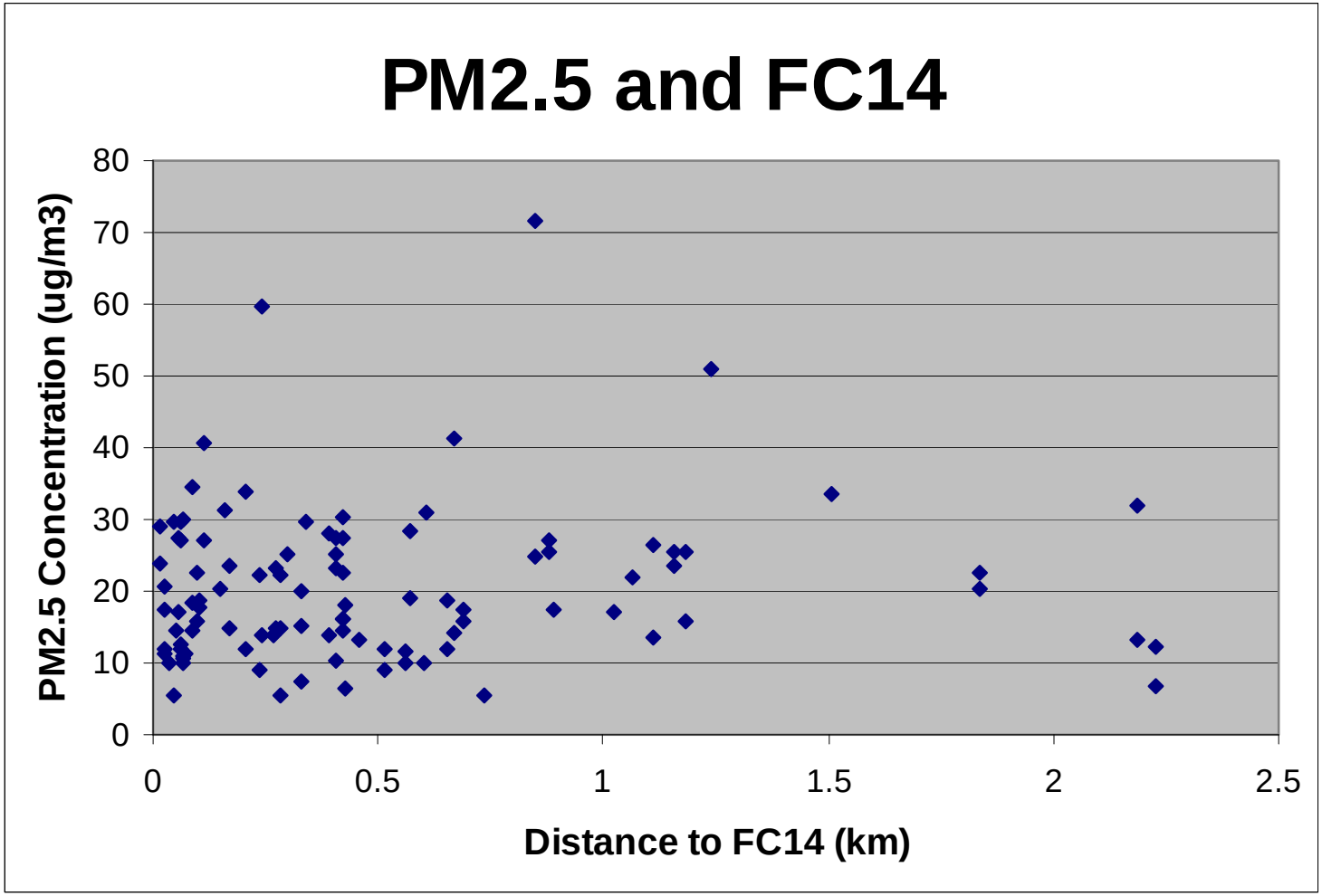

Figure 32. Scatter plot of $MP_{2.5}$ mass with distance from FC14 Roadways, interstate.

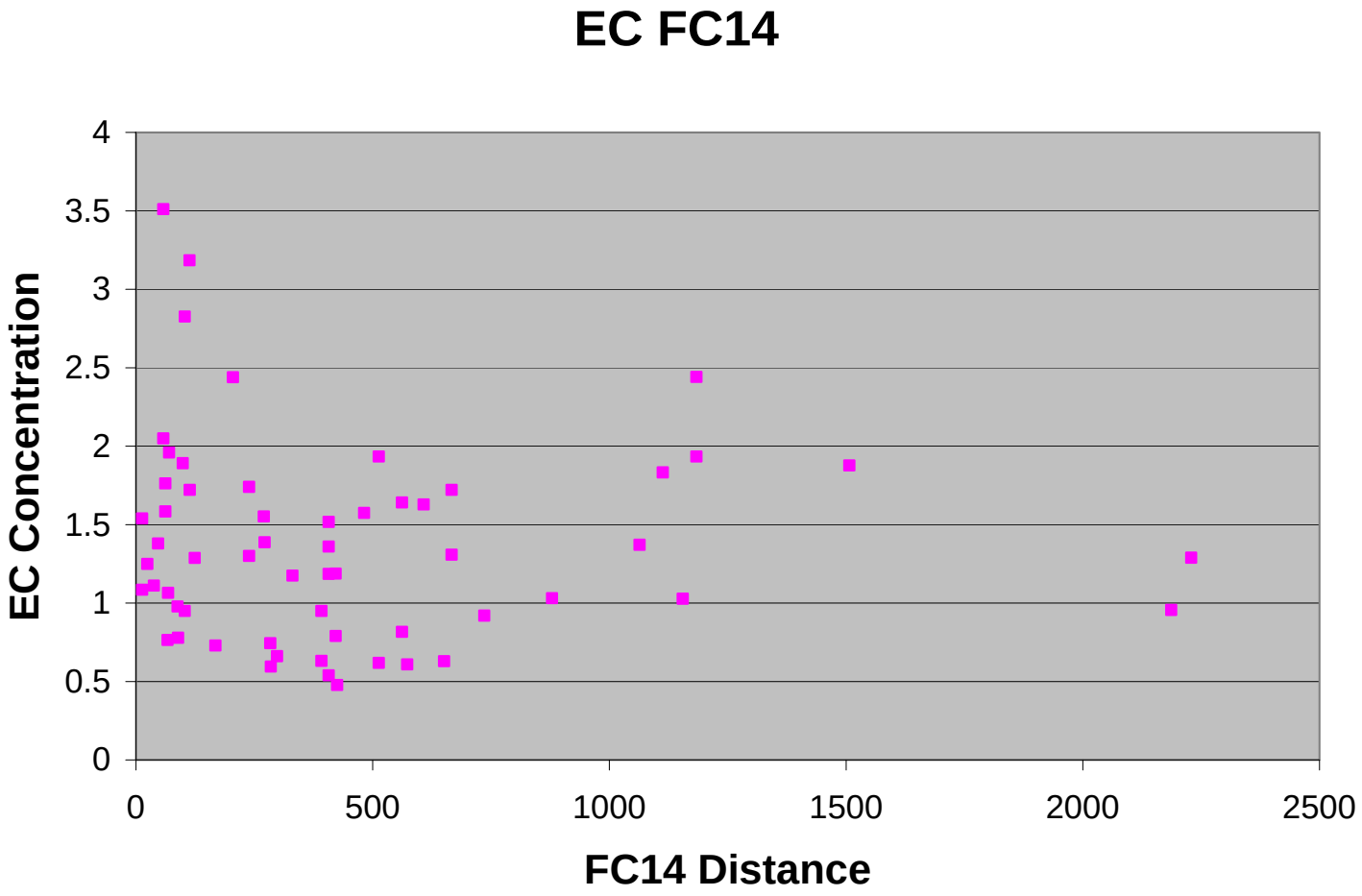

Figure 33. Scatter plot of elemental carbon with distance from FC14 Roadways, interstate.

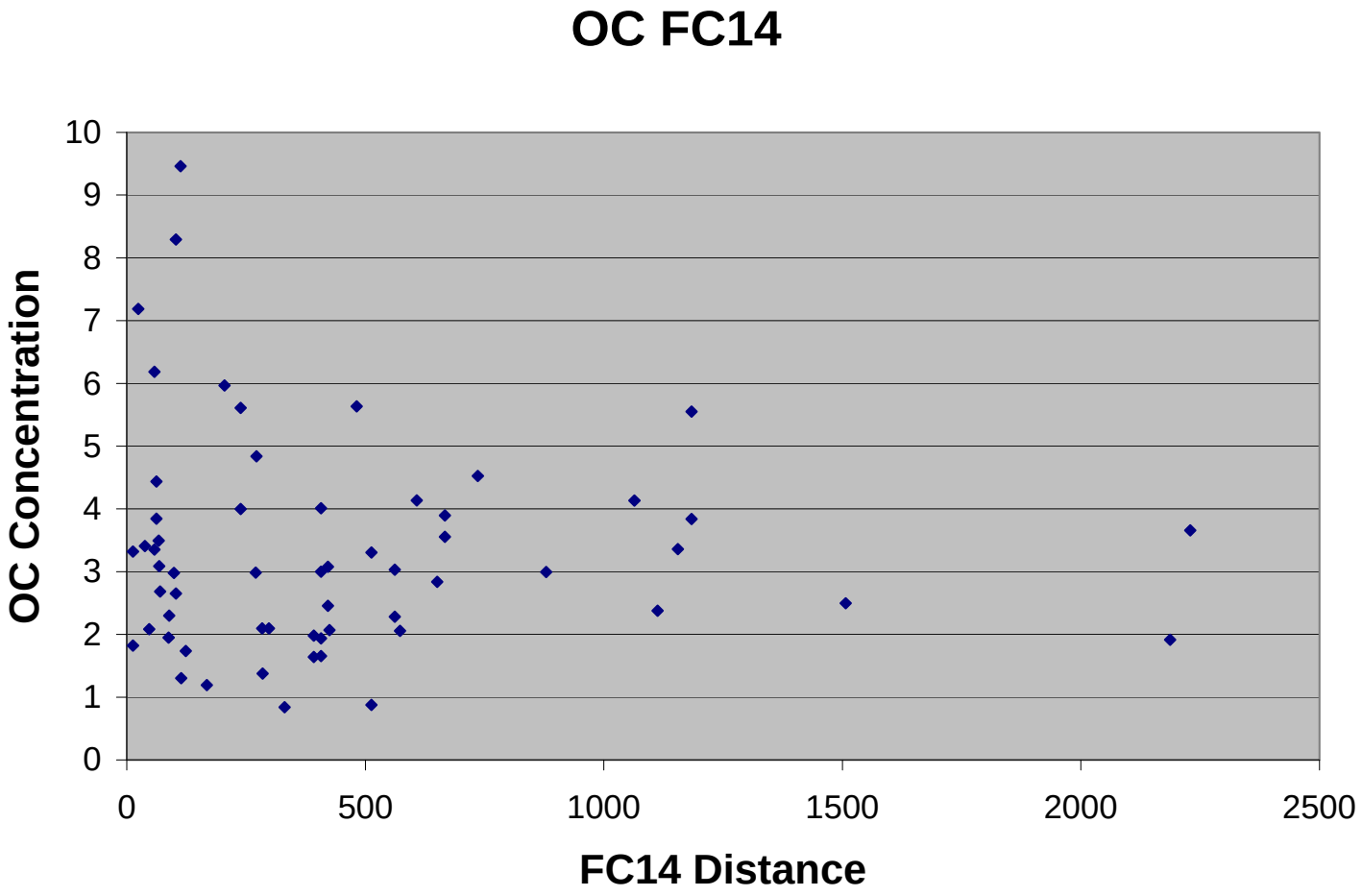

Figure 34. Scatter plot of organic carbon with distance from FC14 Roadways, interstate.

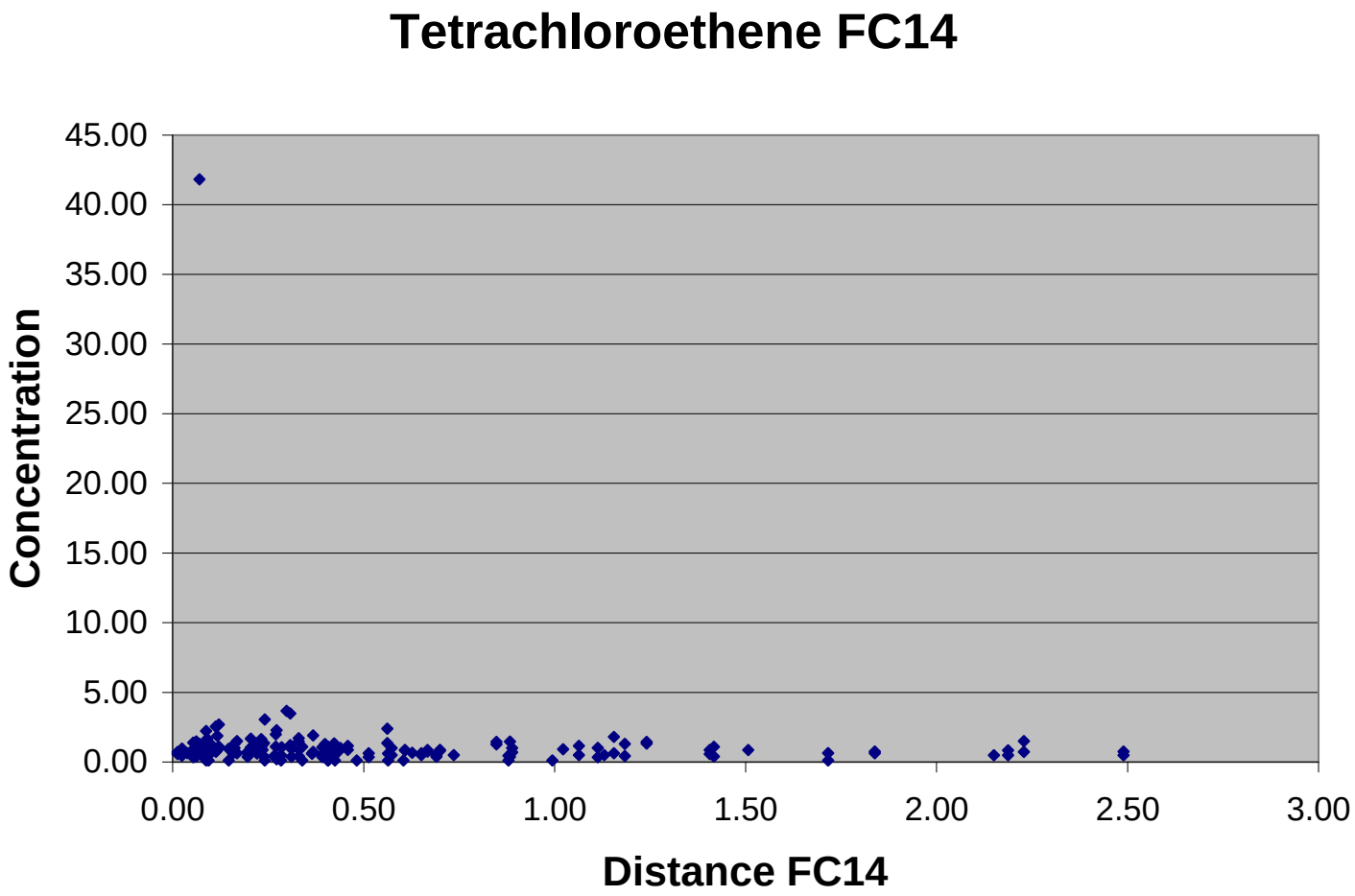

Figure 35. Scatter plot of tetrachloroethylene with distance from FC14 Roadways, interstate.

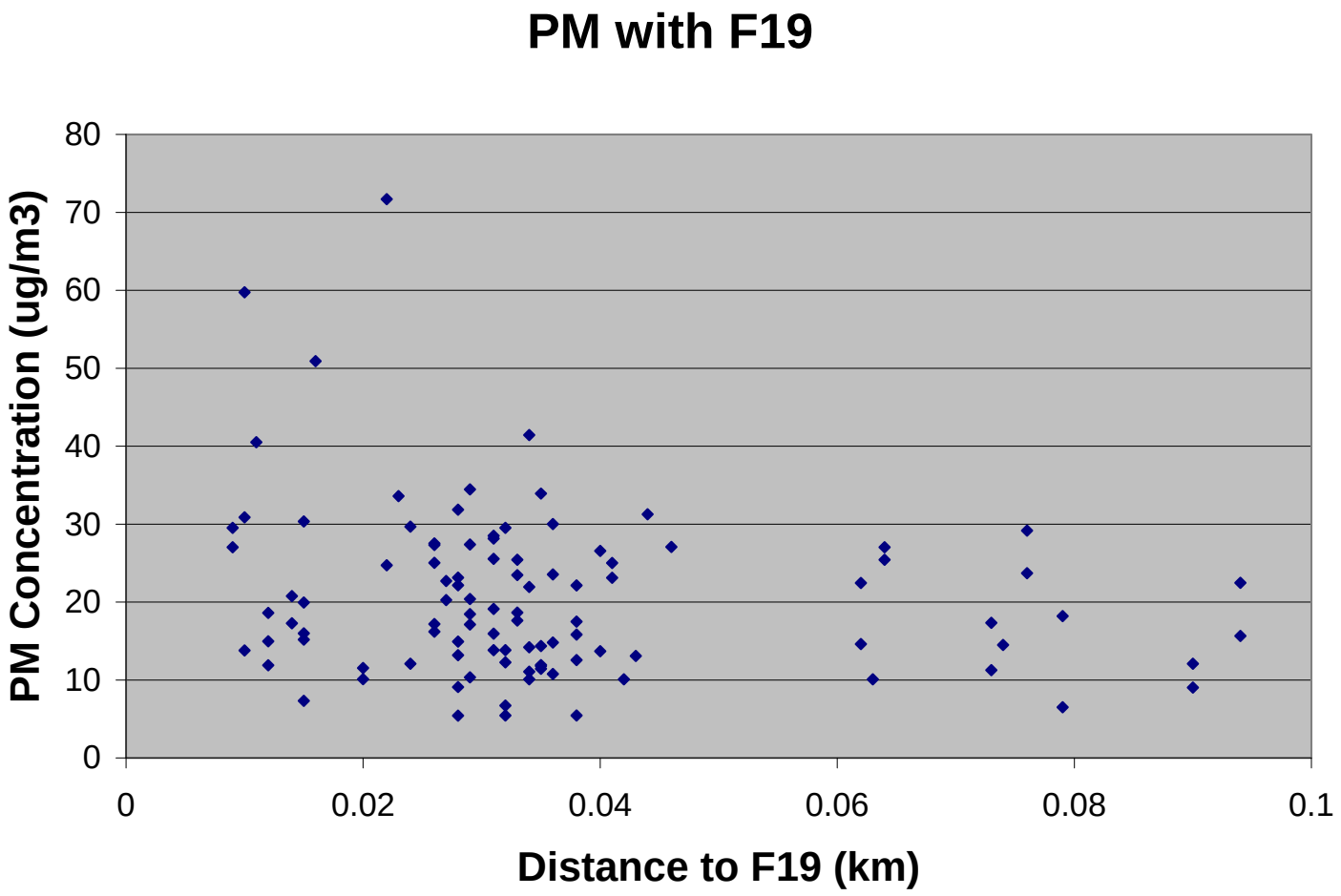

Figure 36. Scatter plot of $PM_{2.5}$ Mass with distance from FC19 Roadways, small local roads.

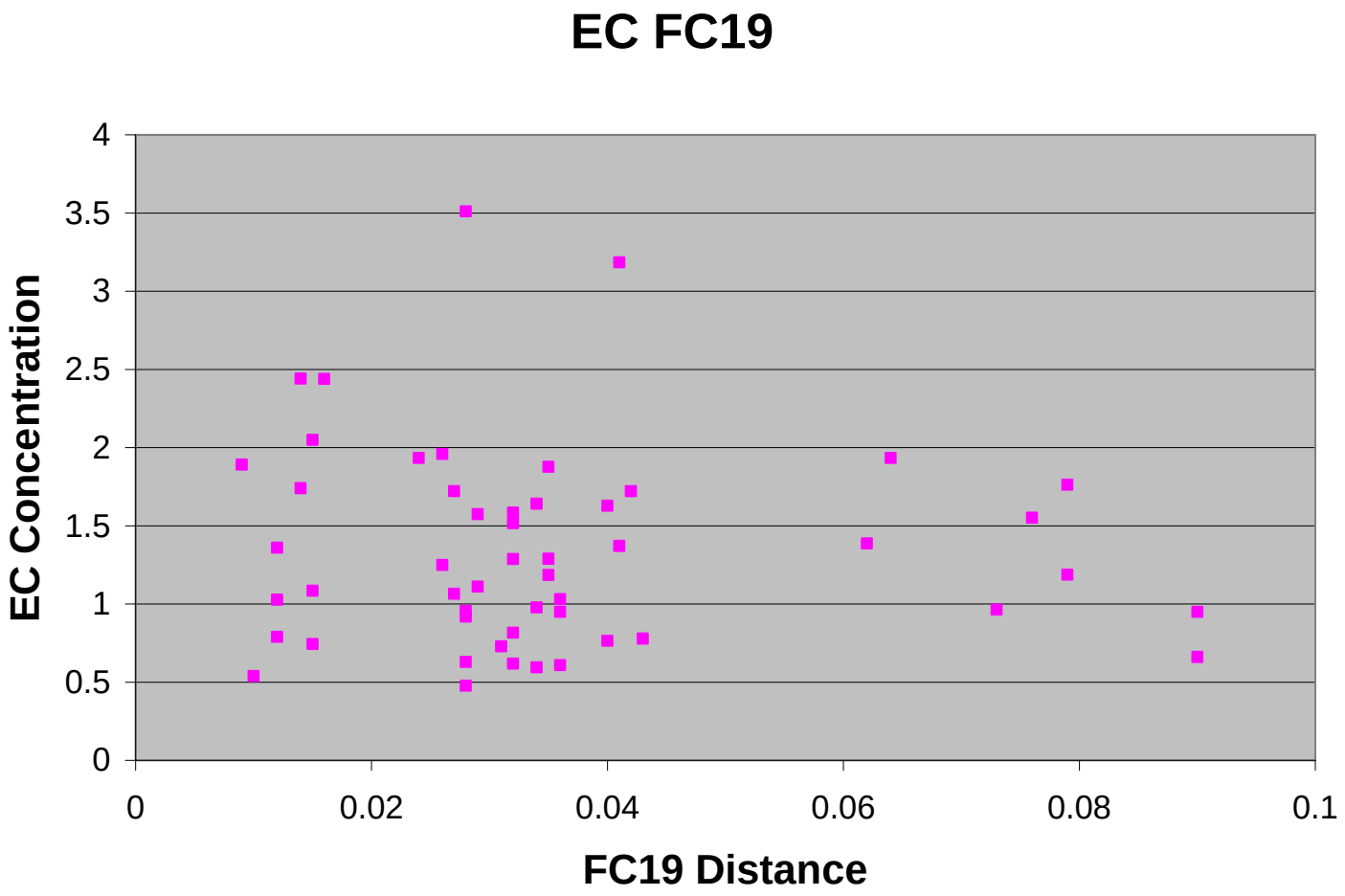

Figure 37. Scatter plot of elemental carbon with distance from FC19 Roadways, small local roads

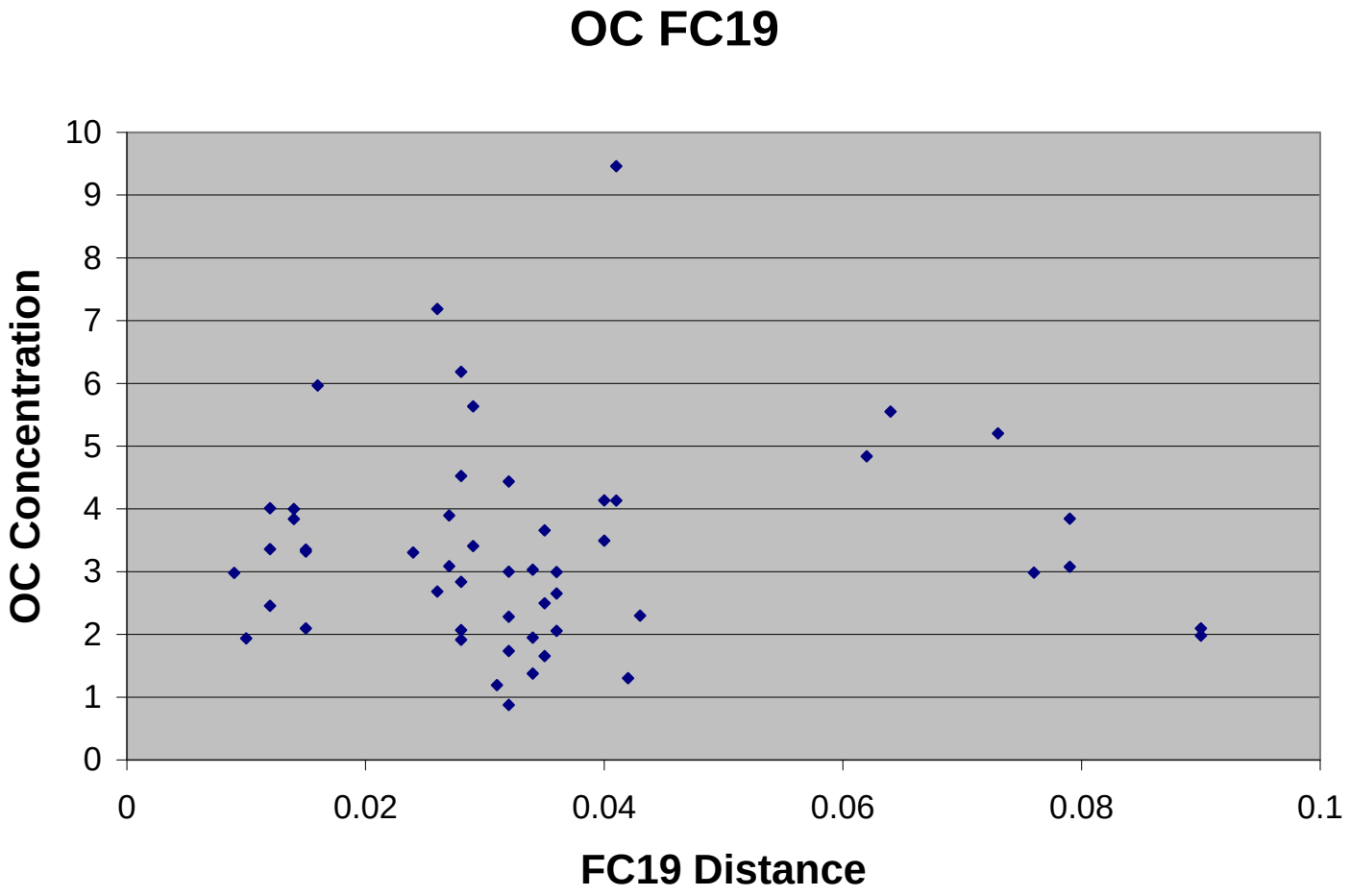

Figure 38. Scatter plot of organic carbon with distance from FC19 Roadways, small local roads

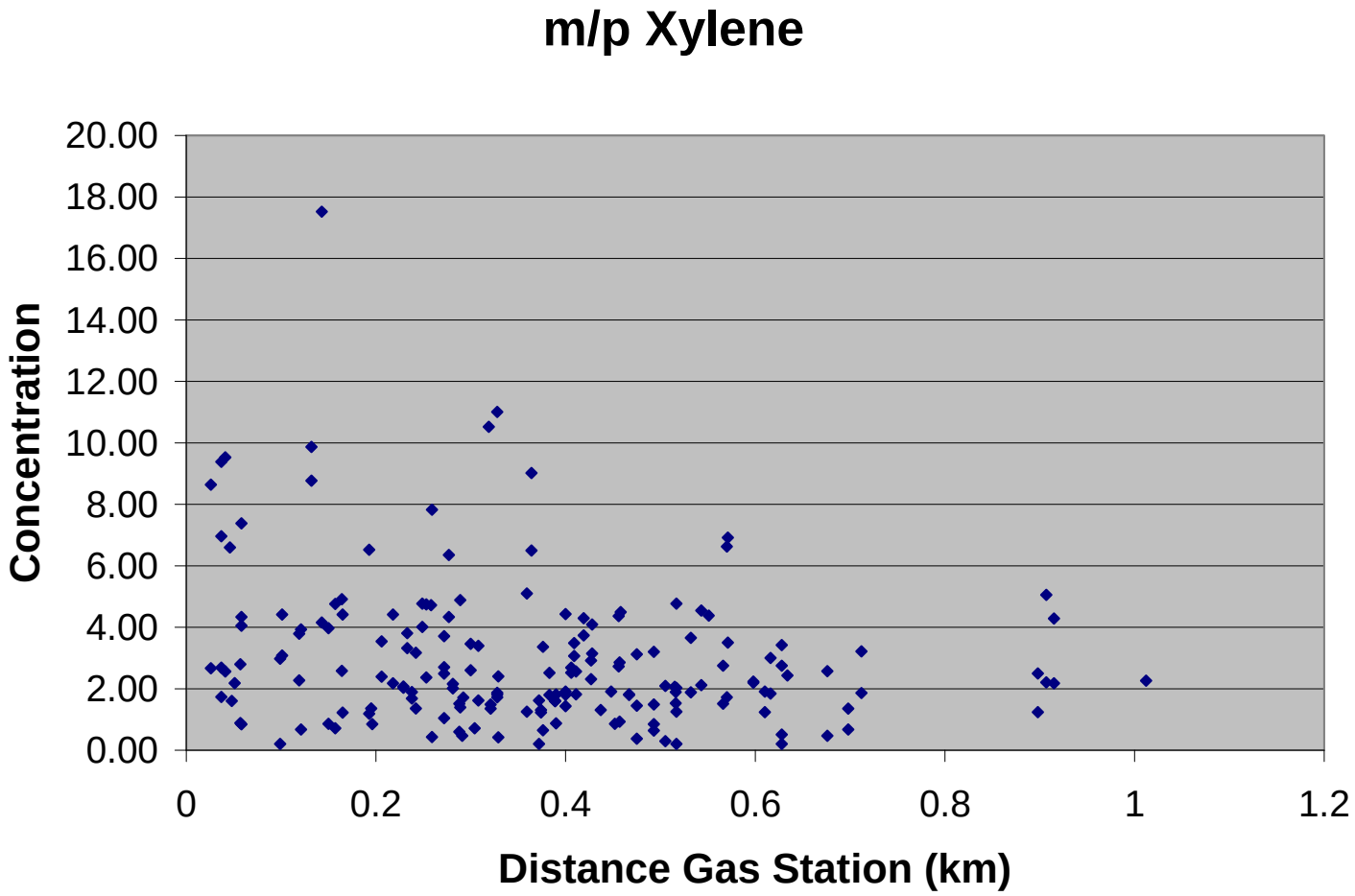

Figure 39. Scatter plot of *m/p* xylene with distance from closest gasoline station.

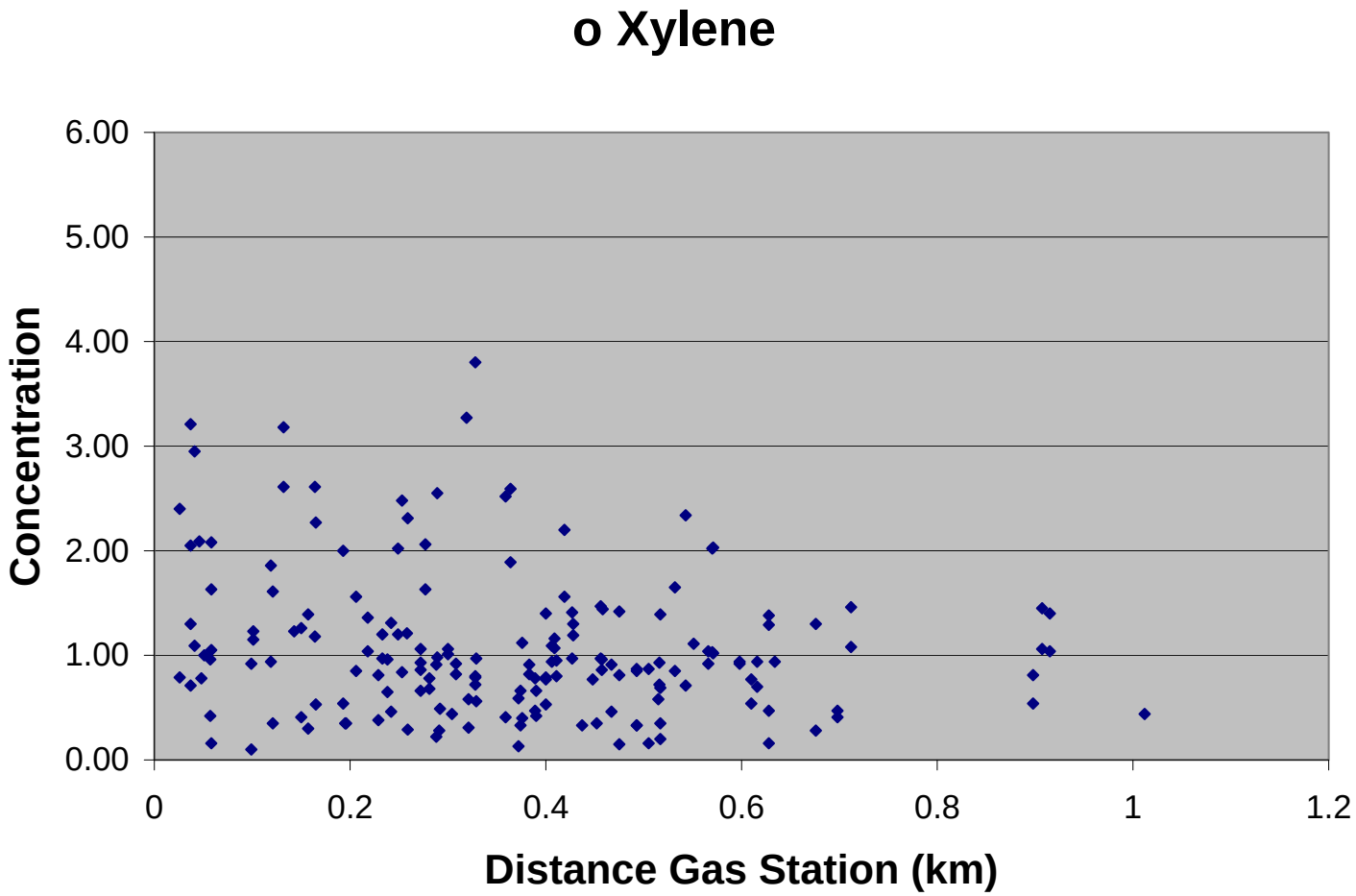

Figure 40. Scatter plot of *o* xylene with distance from closest gasoline station.

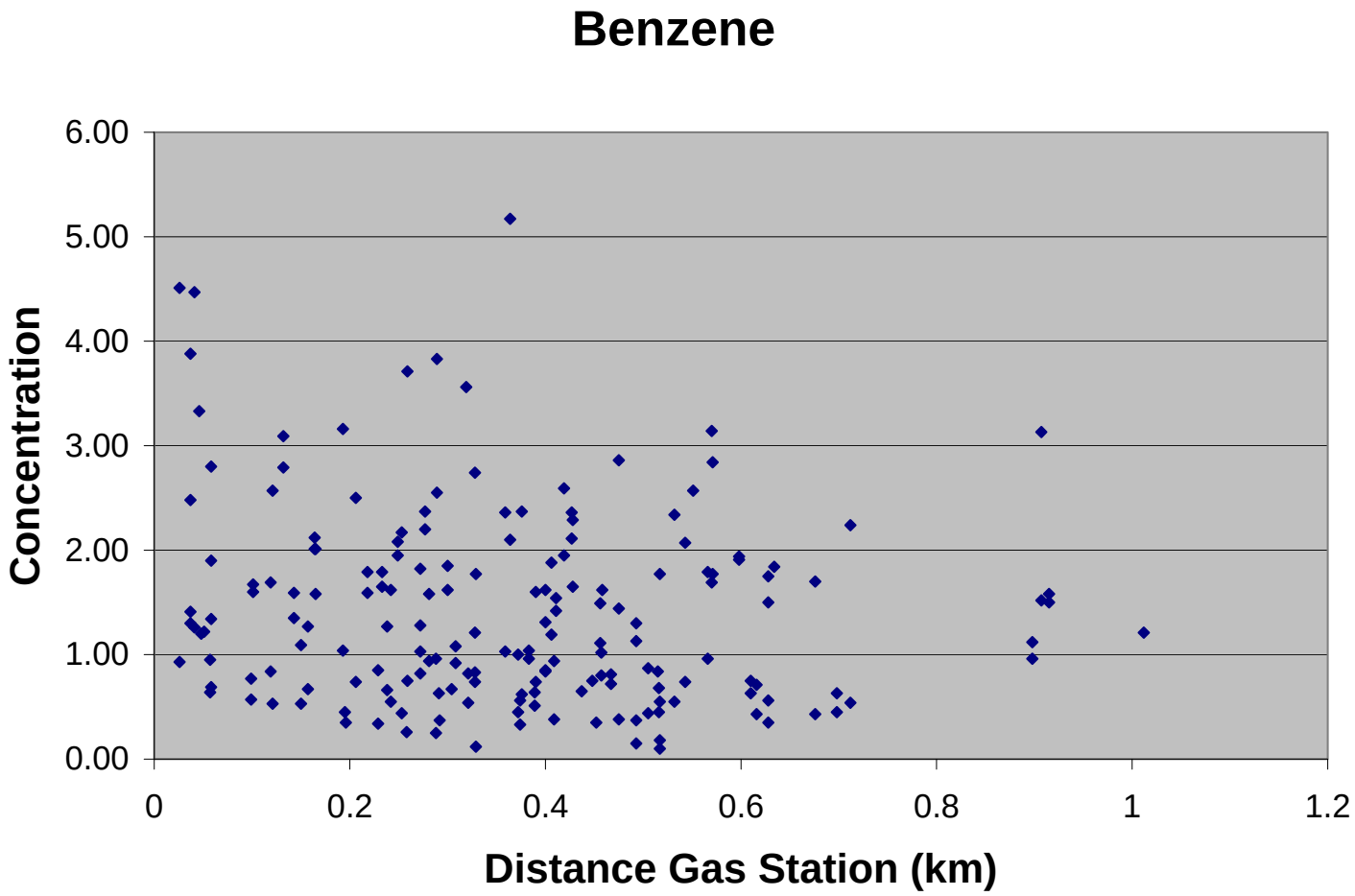

Figure 41. Scatter plot of benzene with distance from closest gasoline station.

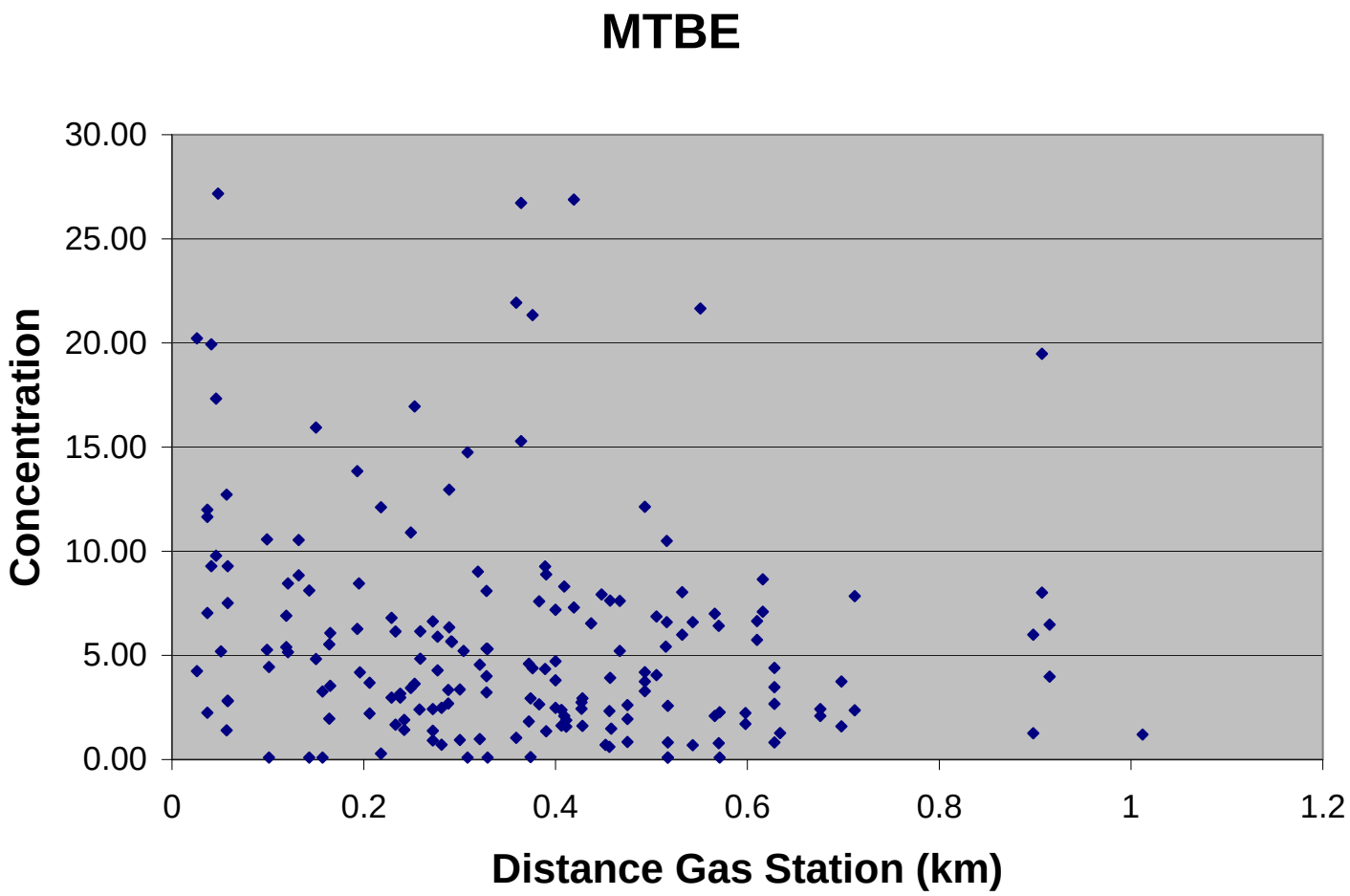

Figure 42. Scatter plot of methy *tert* butyl ether with distance from closest gasoline station.

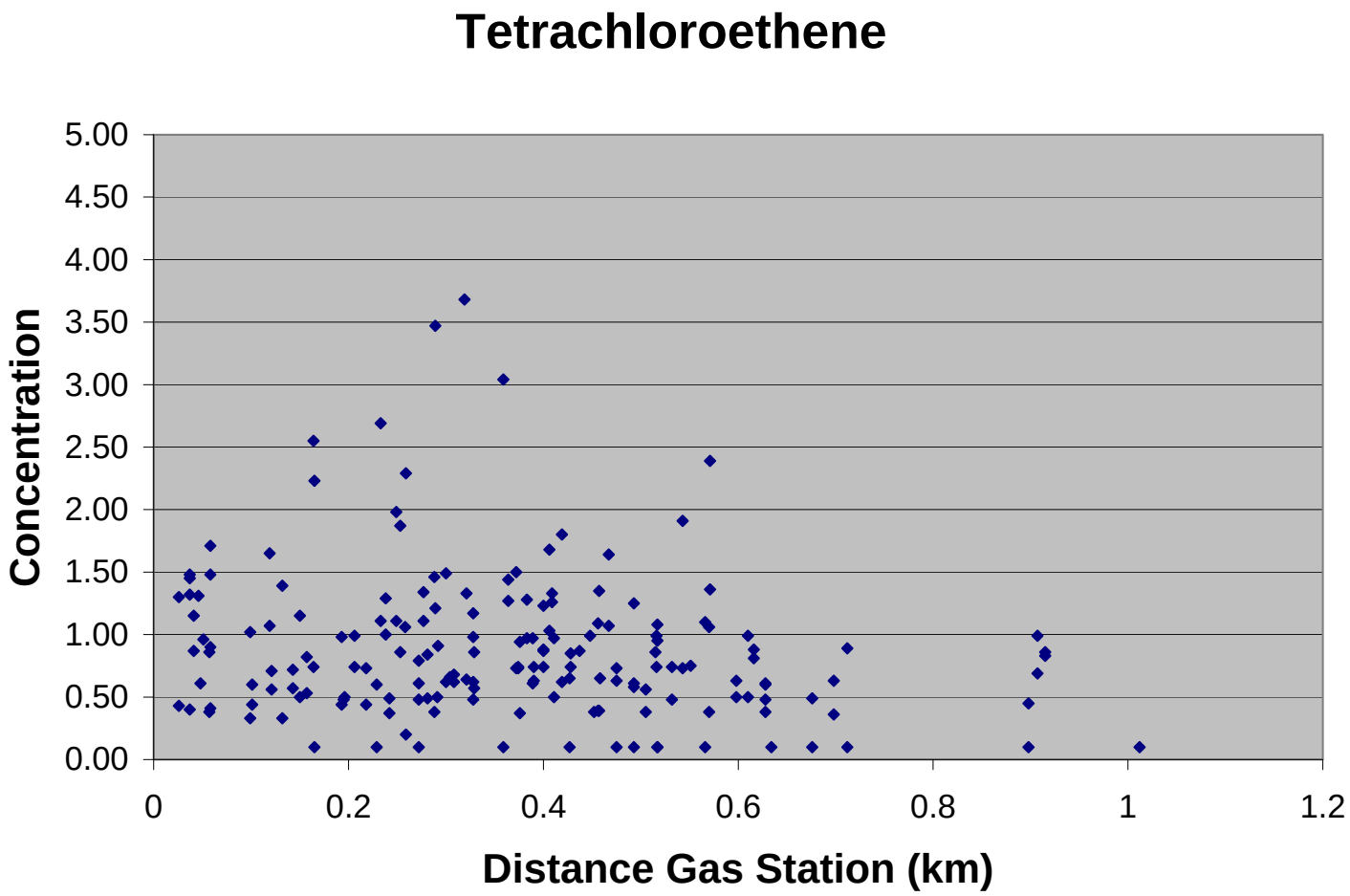

Figure 43. Scatter plot of tetrachloroethylene with distance from closest gasoline station.

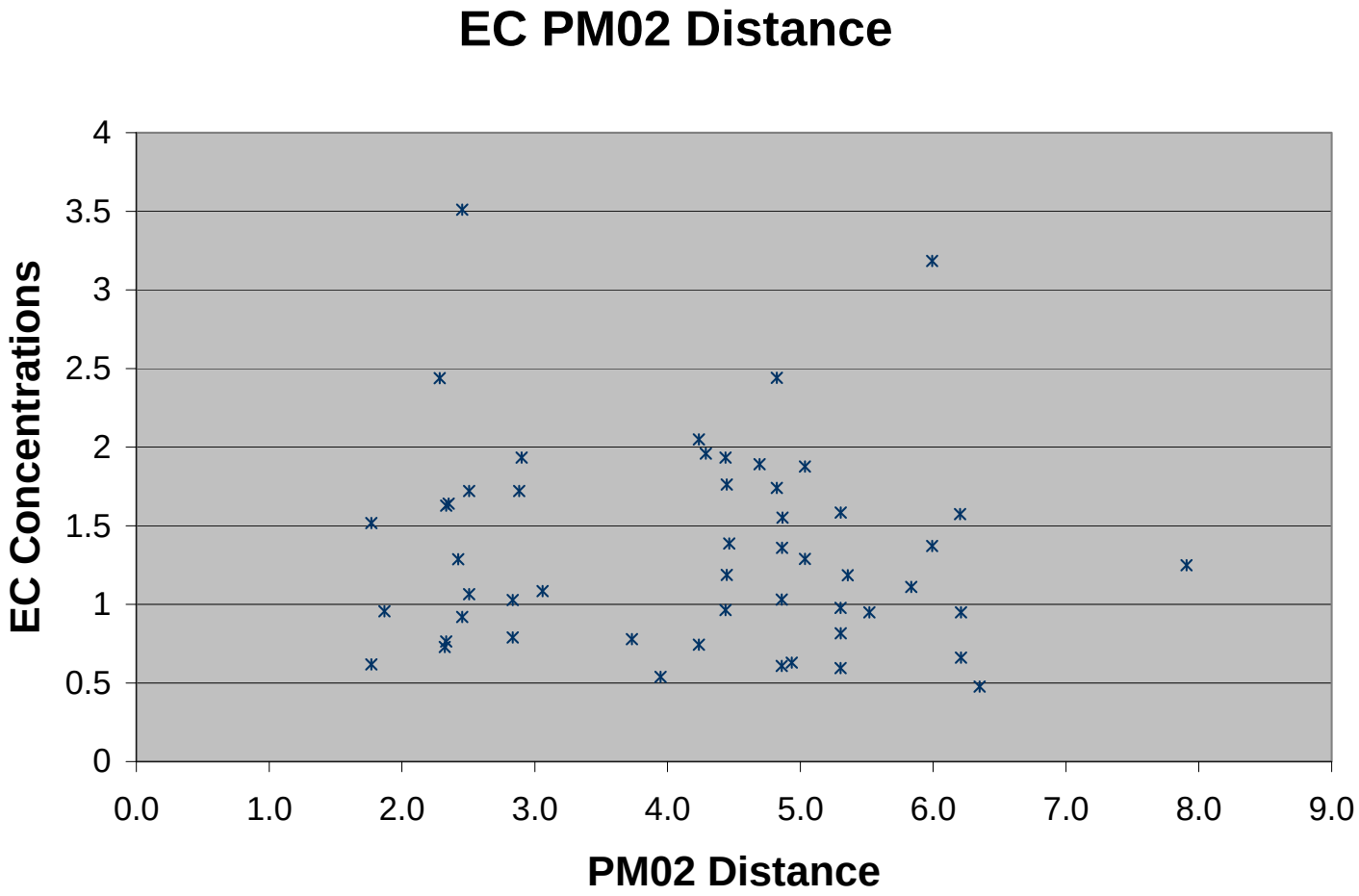

Figure 44. Scatter plot of $PM_{2.5}$ Mass with distance from PM02, truck loading area.

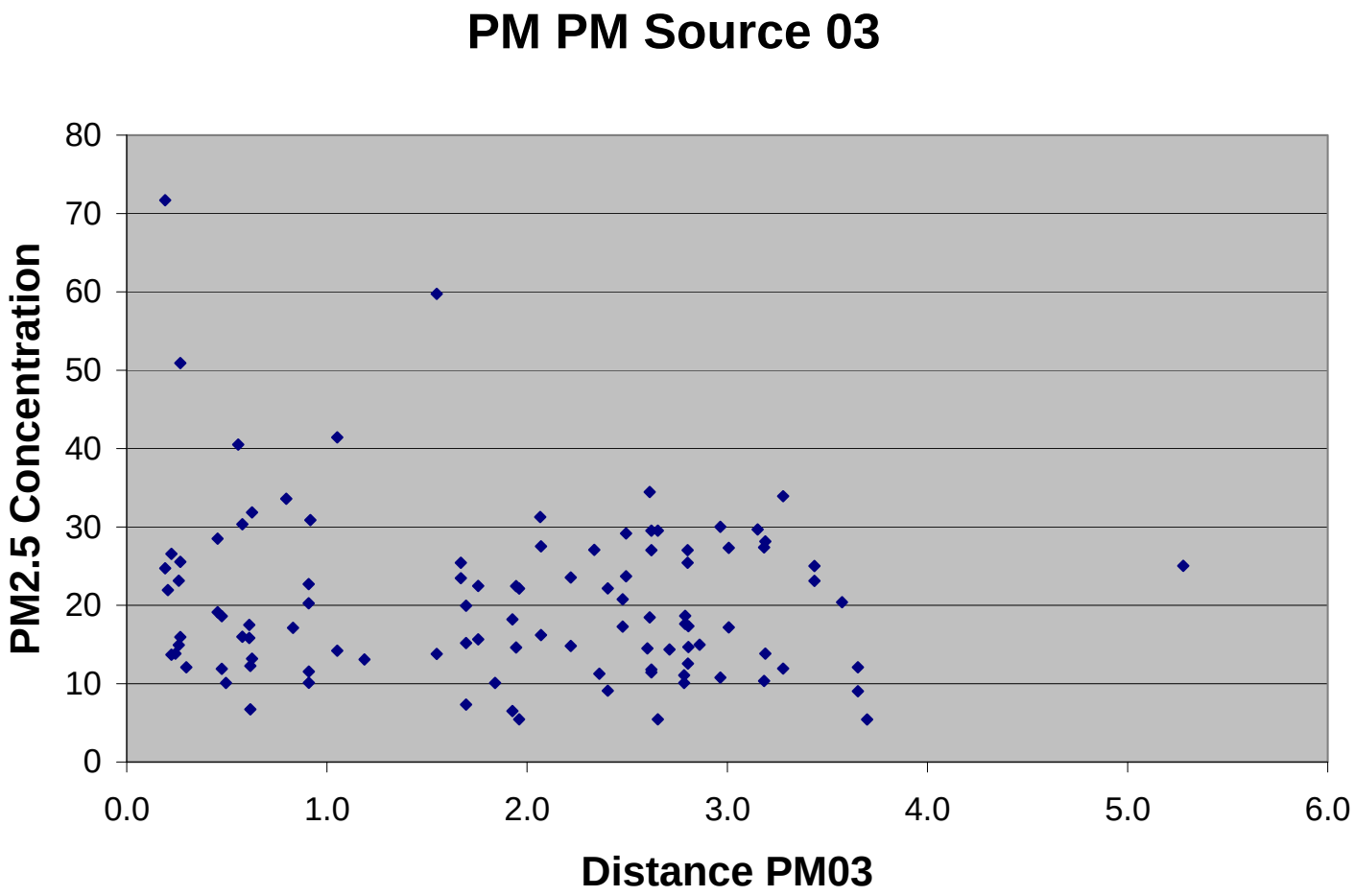

Figure 44. Scatter plot of $PM_{2.5}$ Mass with distance from PM 03 truck loading area and dock.

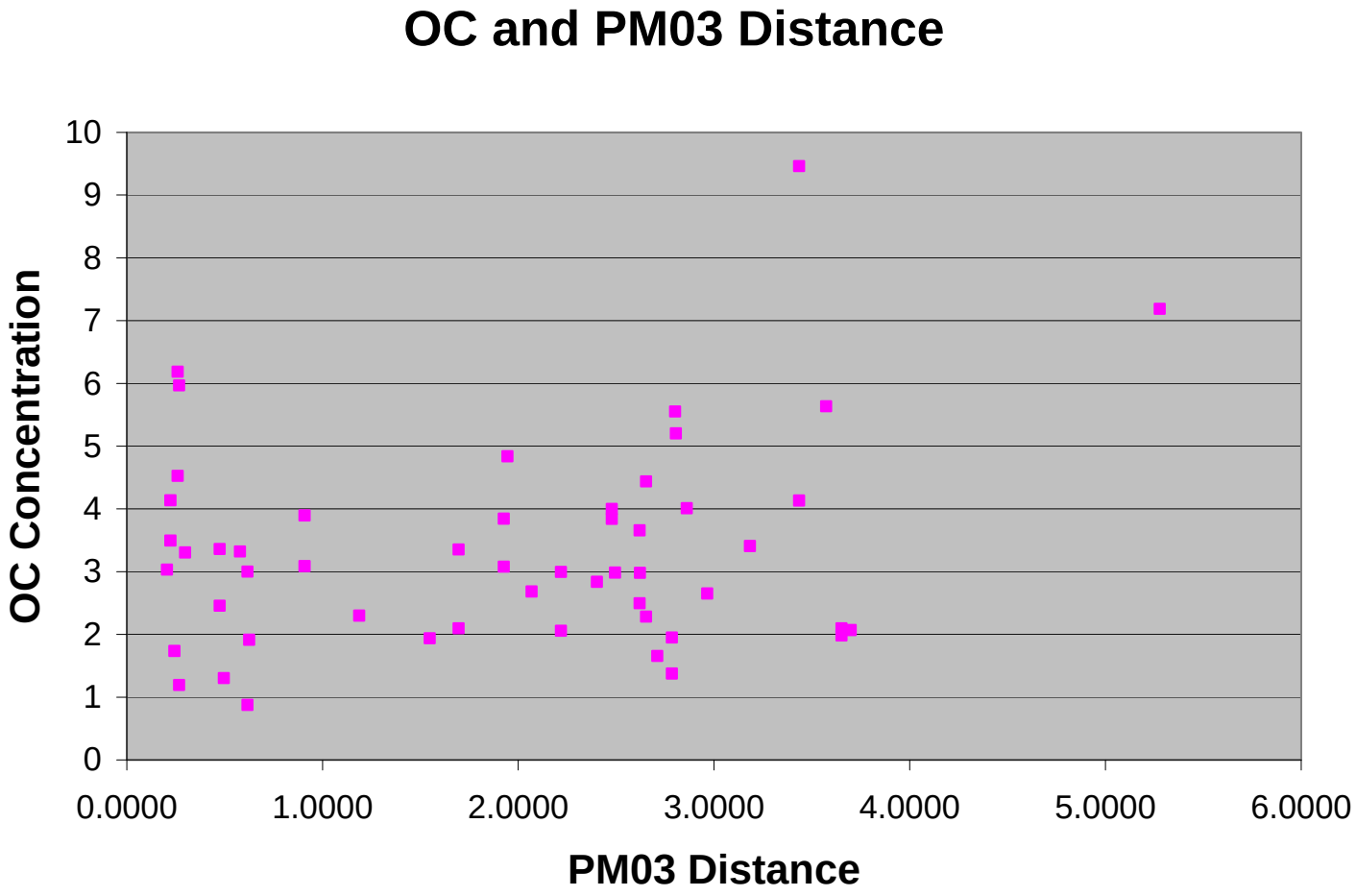

Figure 45. Scatter plot of organic carbon with distance from PM 03 truck loading area and dock..

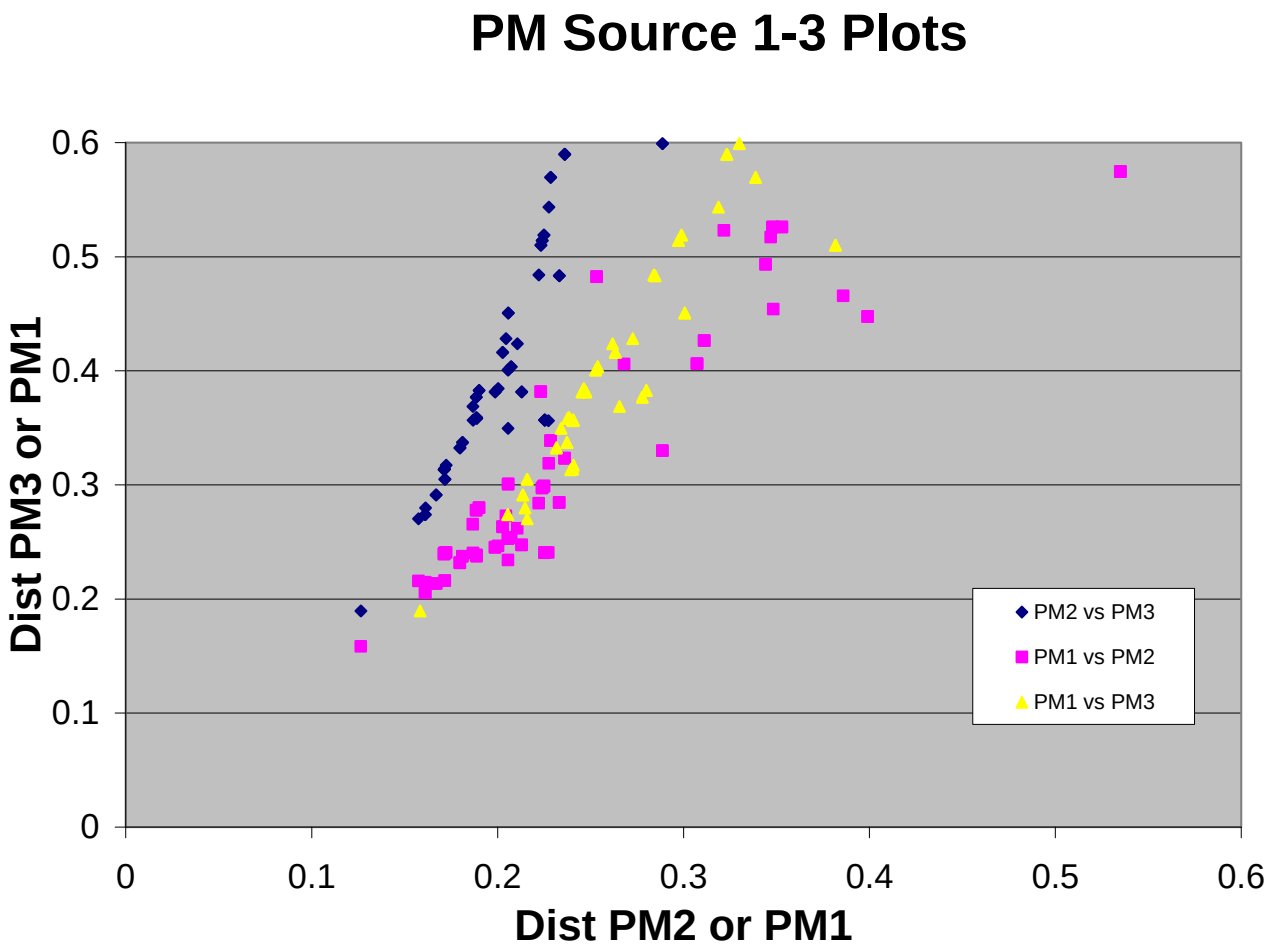

Figure 46. Scatter plot of distance from homes to PM 1, 2 and 3 source regions. Patterns same indicating a high correlation for homes close to these sources (a few hundred yards).

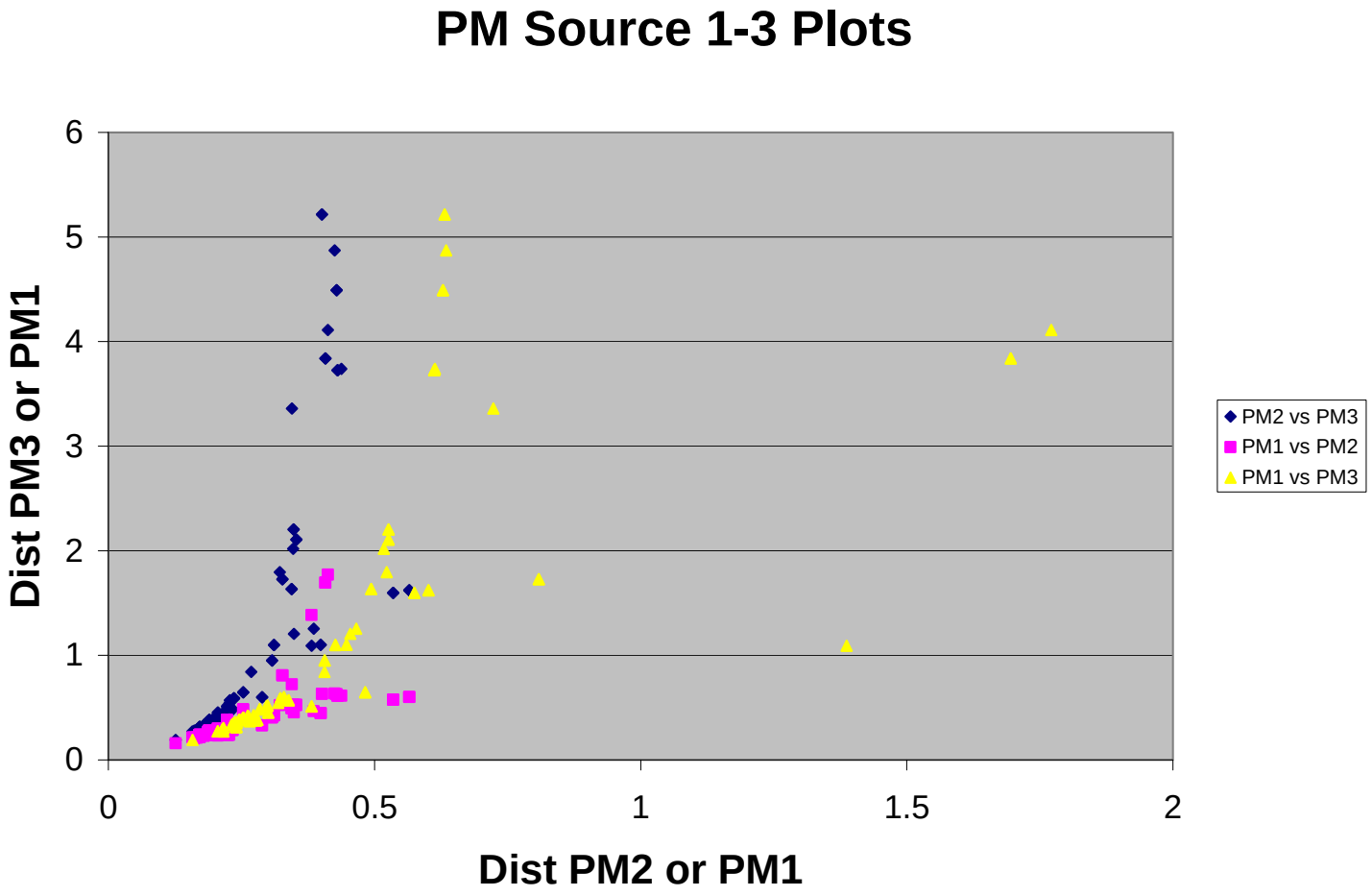

Figure 47. Scatter plot of distance from homes to PM 1, 2 and 3 source regions to 2 km distance.

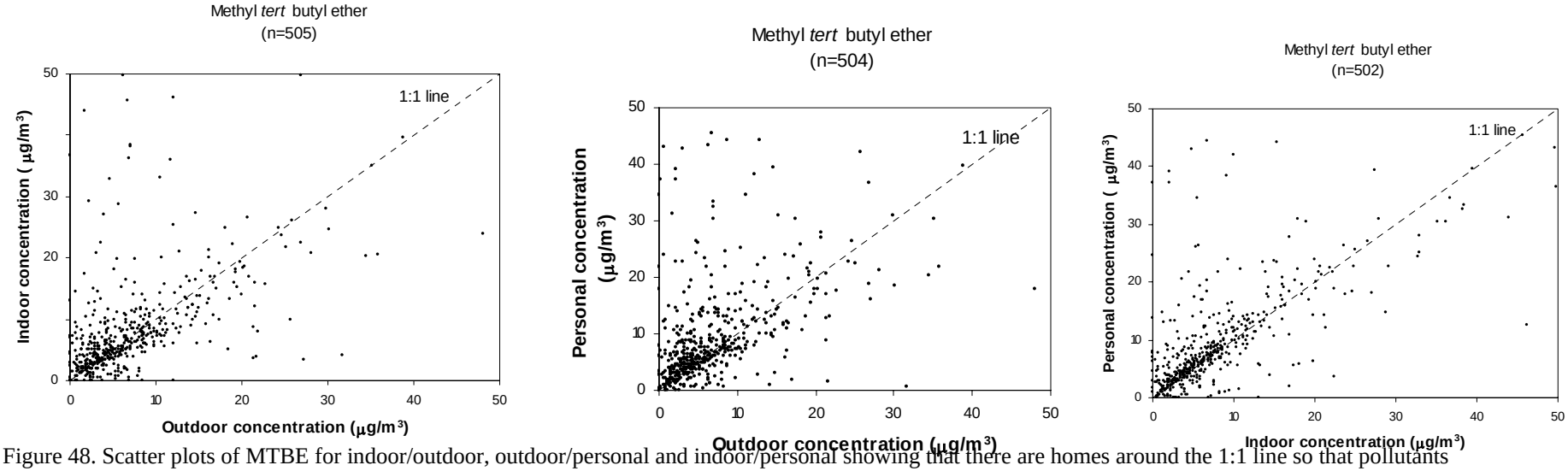

Figure 48. Scatter plots of MTBE for indoor/outdoor, outdoor/personal and indoor/personal showing that there are homes around the 1:1 line so that pollutants arising from outdoor will affect personal exposure.

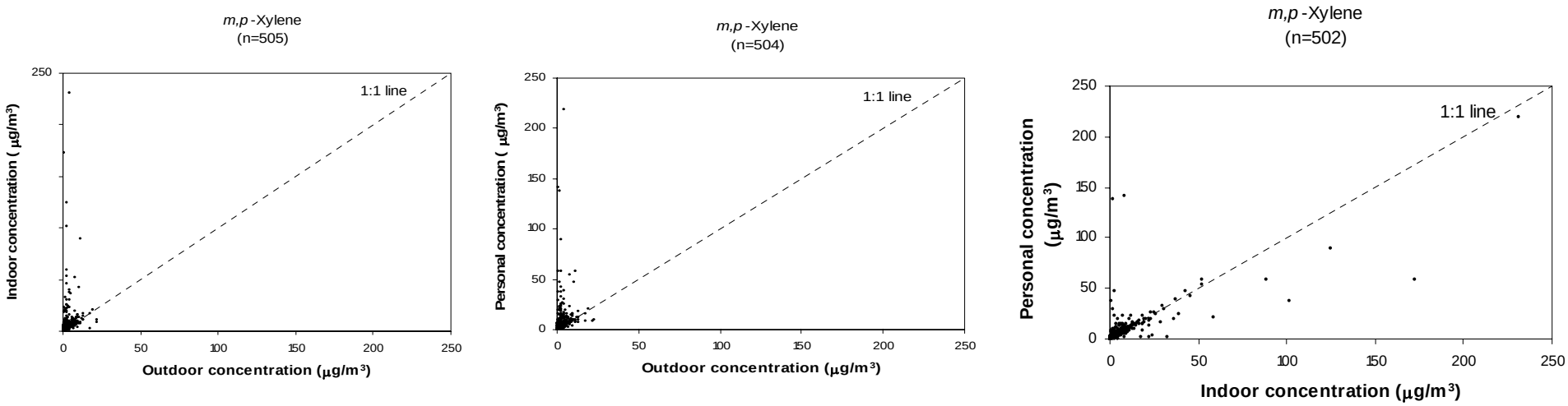

Figure 49. Scatter plots of MTBE for indoor/outdoor, outdoor/personal and indoor/personal showing that a subset of homes around the 1:1 line so that pollutants arising from outdoor will affect personal exposure.

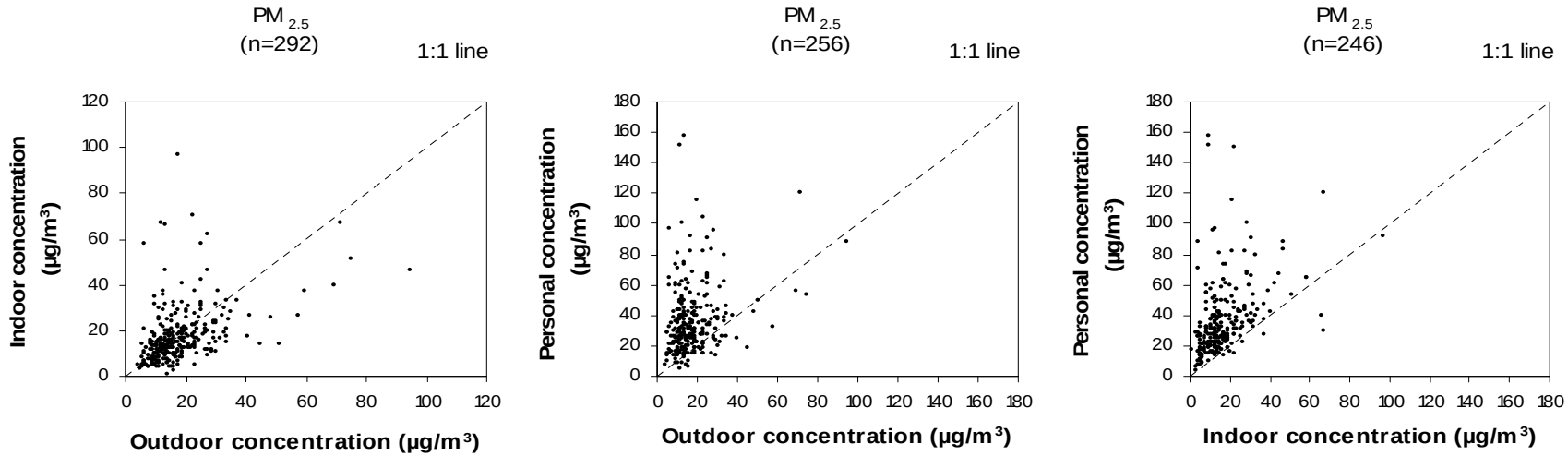

Figure 50. Scatter plots of $PM_{2.5}$ Mass for indoor/outdoor, outdoor/personal and indoor/personal showing that some homes are parallel to the 1:1 line so that pollutants arising from outdoor will affect personal exposure.

References:

Harrison, RM Smith, DJR and Luhana , L Source apportionment of atmospheric polycyclic aromatic hydrocarbons collected from an urban location in Birmingham, UK EST 30, 825-832, 1996.

Netter, J. Kutner, MH, Nachsheim, CJ and Wasserman, W Applied Statistical Models. 4th edition, R.D. Irwin, Inc, Homewood, IL 1996.

Poulopoulos, SG and Philippopoulos, CJ, "The Effect of Adding Oxygenated Compounds to Gasoline on Automotive Exhaust Emissions" Transactions of the ASME, 125, 344-350, 2003

Rogge, WF Hildemann, LM Mazurek, MA Cass, GR and Simoneit, BRT, Sources of fine organic aerosol. 5. Natural gas home appliances. Environmental Science and Technology, 27, 636-651, 1993.

Weisel, CP, Zhang, JJ, Turpin, BJ, Morandi, MT Colome, S, Stock, TH Spektor, DM, Korn, L, Winer, A, Alimokhtari , S, Kwon, J, Mohan, K Harrington, R, Giovanetti, R Cui, W, Afshar, M, Maberti, S, Shendell, D "The Relationships of Indoor, Outdoor and Personal Air (RIOPA) Study: Study Design, Methods and Quality Assurance/Control Results, Journal of Exposure Analysis and Environmental Epidemiology, In Press 2004.

Weisel, Zhang, Turpin, Morandi, Colome, Stock and Spektor "Relationships of Indoor, Outdoor and Personal Air (RIOPA)", HEI Final Report, In Press 2004.

Appendix A

m,p-Xylene

Bivariate Pearson Correlation

The correlation coefficient between the ln-transformed *m,p*-xylene concentrations and the distance to urban interstate (FC11) roadways was -0.20 (p=0.007). The correlation coefficients between the ln-transformed *m,p*-xylene concentrations and the inverse distance to urban major arterial (FC14) roadways and the inverse distance to urban collector (FC17) roadways were 0.19 (p=0.01) and 0.22 (p=0.0034), respectively. The correlation between the ambient air concentration of *m,p*-xylene and distances to individual roadways were examined for the major roadway classes (I-95 for FC11; Rt.1, Rt.27, Rt.28, Rt.439 for FC14). The distance to the I-95 was statistically significantly correlated to the ln-transformed concentration of *m,p*-xylene in the residential ambient air (-0.194, p=0.0093). The distance to the US Highway Route 1 also was statistically significantly correlated to the ln-transformed concentration of *m,p*-xylene in the residential ambient air (-0.272, p=0.0002).

The correlation coefficient between ln-transformed ambient air concentration of *m,p*-xylene and the inverse distance to the closest gas station was 0.28 (p=0.0002). For *m,p*-xylene, only the point sources that were closer than 3 km from any of the sampled homes and emissions larger than 0.9 tons of annual total generation, were considered in the data analysis. Two point sources met the above criteria; one refinery in Linden, and an industrial emission in Elizabeth. Only the distance between the refinery and the residences had a statistically significant correlation with the ln-transformed *m,p*-xylene concentrations (-0.17, p=0.022).

The Pearson correlation coefficients of the meteorological variables and the ln-transformed *m,p*-xylene concentrations that were statistically significantly correlated at α =0.05, were atmospheric stability, 0.348 ($p<0.0001$); mixing height, -0.254 ($p=0.0009$); wind speed, -0.235 ($p=0.0014$); and temperature, -0.19 ($p=0.0101$). The correlation coefficients of precipitation and relative humidity were 0.125 ($p=0.091$) and 0.129 ($p=0.082$), respectively. Atmospheric pressure was not correlated to the ambient *m,p*-xylene concentration ($p=0.47$).

Preliminary Selection of Predictors

The preliminary regression analysis was performed on the ln-transformed *m,p*-xylene concentration to determine the relative importance of variables within the same types (proximity and meteorological) of independent variables. The distances to the roadways, either original or transformed, were grouped by its FC to examine the importance of proximity of the mobile sources to the *m,p*-xylene air concentration. When the distances to the functional classes were analyzed, the distances to the urban interstates (FC11) and the urban principal arterials (FC14) were included in the resulting linear regression model ($p<0.15$) with $r^2 = 0.0819$. For the gas stations and the point sources, the inverse form of the closest distance was always selected as the largest explanatory predictor variable in the model regardless of the selection methods. The proximity to the refinery was also selected as the larger explanatory variable along with an industry site in the model. Among the meteorological variables, the 48-hour averaged mixing heights, temperature, and wind speed were selected from the preliminary regression analysis. When the 48-hour averaged atmospheric stability was introduced to the initial group of other meteorological variables,

the model was improved (increased r^2), but the mixing height was eliminated from the resulting model.

Selection of the Best-fitting Model

The variables selected by the different regressions methods were relatively consistent. Atmospheric stability, temperature and wind speed were included as predictors in the model as were the inverse distances to the major roadways (FC11) and to gasoline stations (Table A-1). The association of the distance to the refinery was not significant. The parameters and analysis of variance of the regression equations for the *m,p*-xylene ambient air concentration for the best-fitting model with 6 variables selected are given in Table A-1. The C(p), which is Mallows' C_p statistic, associated with this particular subset of variables was determined to be 7.0. The resulting model was appropriate in number of parameters, because the number of parameters (*p*) including the intercept in the best-fitting model exactly matched to the same value of the C(p). The diagnostic plots, the residual plot against the predicted values, the normal probability-probability (PP) plot and normal quantile-quantile (QQ) plot of the residuals were generated and visually examined (Figures A-1,2 and Appendix B). The residuals were randomly distributed without showing any obvious trend or any particular pattern (Figure A-1.) indicating close to a normal distribution and the constant variances. The PP plot was nearly linear so it could be considered the error term of the model follows a normal distribution. Based on the visual diagnosis, there was no significant evidence of lack of fit or of significant unequal error variance for the best 7-parameter regression model.

Possible Outliers were found (Figure A-2) using the test statistics (± *tinv,* .95, *n-p*-1 = 175) of ± 1.6545. The regression equation was recalculated after removal of the seventeen Outliers. The parameters for the best fit equation are given in Table A-2. An increase in the

r^2 was obtained after removal of the outliers. The residuals plotted against the predicted (Figure A-3.) seemed more randomly distributed compared to those in Figure A-1.

Diagnostics of Equal Variances and Multicollinearity Diagnostics

To test the assumption of equal variance, the heteroscedasticity of the parameter estimates were tested as well as multicollinearity (Appendix C). The chi-square was 18 with a probability of 0.19, a value greater than 0.05. Therefore, the variances of the parameter estimates could be concluded as not being significantly different. As a consequence, the equal error variances in parameter estimates were assumed in the best-fitting 6-parameter regression model. The multicollinearity of predictor variables in the best-fitting model was tested. The bivariate Pearson correlations between pairs of predictors included in the model were examined for any significant correlation between the predictors.

The variance inflations for all predictors were close to 1, a value smaller than 10, suggesting that there was no significant collinearity between the predictors in the model. However, the collinearity diagnostics suggest that there were possible co-dependences, which might overspecify the model outcome. In particular the meteorological conditions were somewhat correlated and commercial enterprises, such as gasoline stations, are preferentially located on or near major roadways so correlations in the proximity variables could exist.

In order to attempt to reduce the multicollinearity diagnosed, the temperature, which had larger proportion of variation than 0.5, was removed from the best-fitting model and the multicollinearity of the resulting model diagnosed. When the temperature was removed from the model, the coefficient of determination (r^2) of the resulting 5-parameter model decreased from 0.33 to 0.24, and the condition index decreased from 116 to 41. The interaction

between the predictors in 5-parameter model appeared to be decreased after removal of the temperature from the model, but the eigenvalue was still smaller than 0.01 (0.0023) and the proportion of variation of the stability (0.97) were still greater than 0.5.

The multicollinearity diagnostics described above exhibited divergent results. The largest condition index and proportion of variation indicated potential collinearity may exist in the predictors. However, the variance inflation factors were much smaller than 10 for all five parameter estimates indicating that the multicollinearity may not be a problem. Neter et al suggested (1996) that even though there is serious multicollinearity, the fitted model may be useful for estimating mean responses or making predictions, if the inferences of the fitted regression model are restricted to the same multicollinearity pattern as the data on which the regression model is based. Consequently, it was concluded that retaining all predictors included in the best-fitting model with its qualitative characteristics is more beneficial for explanatory observational purposes of this research than dropping the potentially intercorrelated predictors from the model. Similar considerations were used for the other compounds as well.

Table A-1. Results of the Best-fitting 7-Parameter Model for *m,p*-Xylene

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	35.9789	7.19578	12.81	<.0001
Error	177	99.43418	0.56178		
Corrected Total	182	135.4131			

Root MSE	0.74952	R-Square	0.2657
Dependent Mean	0.81562	Adjusted R-Square	0.2450
Coefficient of Variation	91.89477		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	5.56153	2.24241	2.48	0.0141
F14_1mInv	(Distance to FC14)$^{-1}$	1	14.56178	5.17879	2.81	0.0055
GS1mInv	(Distance to Gas Station)$^{-1}$	1	22.46222	8.56119	2.62	0.0095
Stab4	Atmospheric Stability	1	0.52630	0.14753	3.57	0.0005
K5	Temperature	1	-0.02472	0.00671	-3.69	0.0003
U4	Wind speed	1	-0.12254	0.05923	-2.07	0.0400

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Atmospheric Stability	0.1317	0.1317	27.9107	27.46	<.0001
2	(Distance to Gas Station)$^{-1}$	0.0460	0.1778	18.9388	10.08	0.0018
3	Temperature	0.0362	0.2140	12.3164	8.24	0.0046
4	(Distance to FC14)$^{-1}$	0.0340	0.2479	6.2173	8.04	0.0051
5	Wind Speed	0.0178	0.2657	3.9860	4.28	0.0400

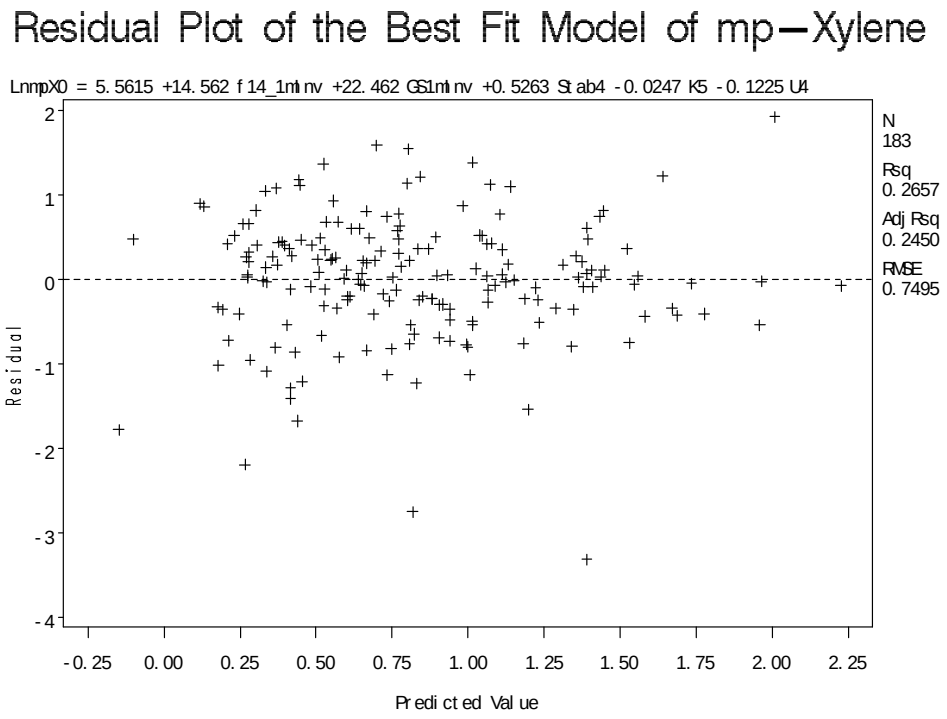

Figure A-1. Residual vs. Predicted Plot of the Best-fitting 7-Parameter Model of *m,p*-Xylene

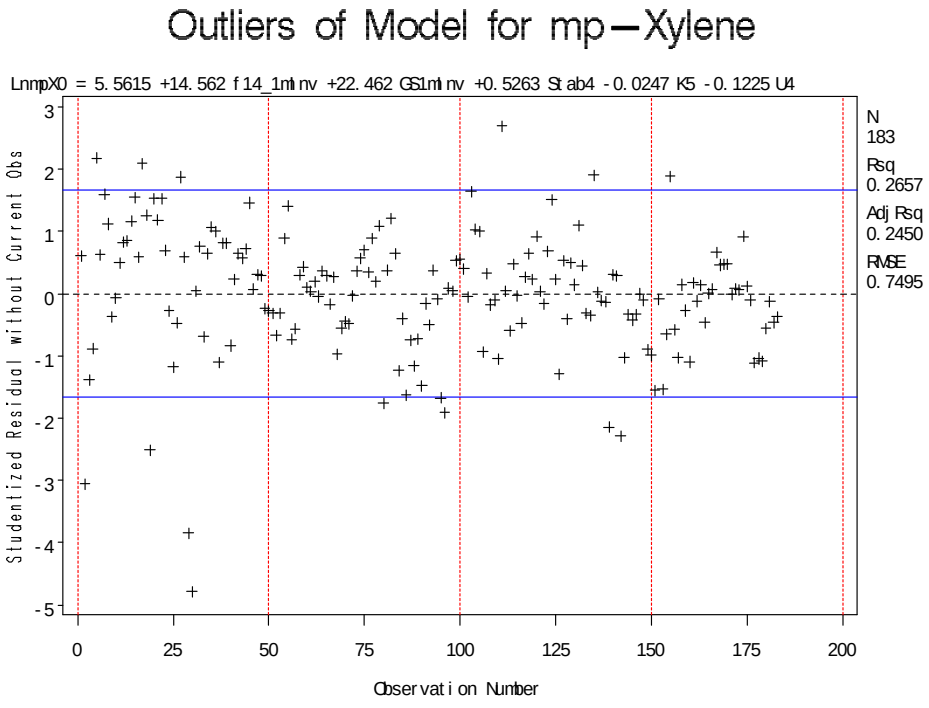

Figure.A-2. Outliers of 6-Parameter Model of *m,p*-Xylene

Table.A-2. Results of the Best-fitting 5-Parameter Model for *m,p*-Xylene after Removing the Twenty Outliers

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	22.98982	4.59796	15.99	<.0001
Error	162	46.58357	0.28755		
Corrected Total	167	69.57338			

Root MSE	0.53624	R-Square	0.3304
Dependent Mean	0.86365	Adjusted R-Square	0.3098
Coefficient of Variation	62.08976		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	4.94236	1.70161	2.9	0.0042
F14_1mInv	(Distance to FC14)$^{-1}$	1	7.94739	4.43103	1.79	0.0747
GS1mInv	(Distance to Gas Station)$^{-1}$	1	17.43615	6.29951	2.77	0.0063
Stab4	Atmospheric Stability	1	0.53744	0.11065	4.86	<.0001
K4	Temperature	1	-0.0232	0.00507	-4.58	<.0001
U4	Wind Speed	1	-0.0653	0.04438	-1.47	0.1431

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Atmospheric Stability	0.1813	0.1813	32.3859	36.76	<.0001
2	Temperature	0.0871	0.2684	13.4982	19.64	<.0001
3	(Distance to Gas Station)$^{-1}$	0.0385	0.3068	6.2732	9.10	0.0030
4	(Distance to FC14)$^{-1}$	0.0147	0.3215	4.7579	3.52	0.0624
5	Wind Speed	0.0090	0.3304	4.6109	2.17	0.1431

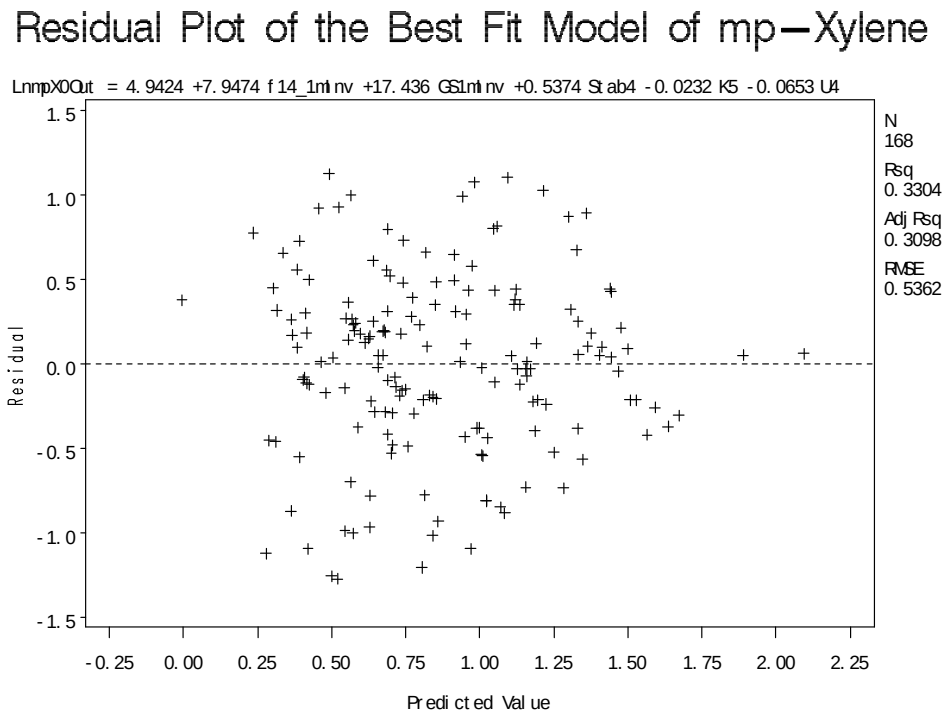

Figure A-3. Residual vs. Predicted Plot of the Best-fitting 6-Parameter Model of *m,p*-Xylene after Removing the Outliers

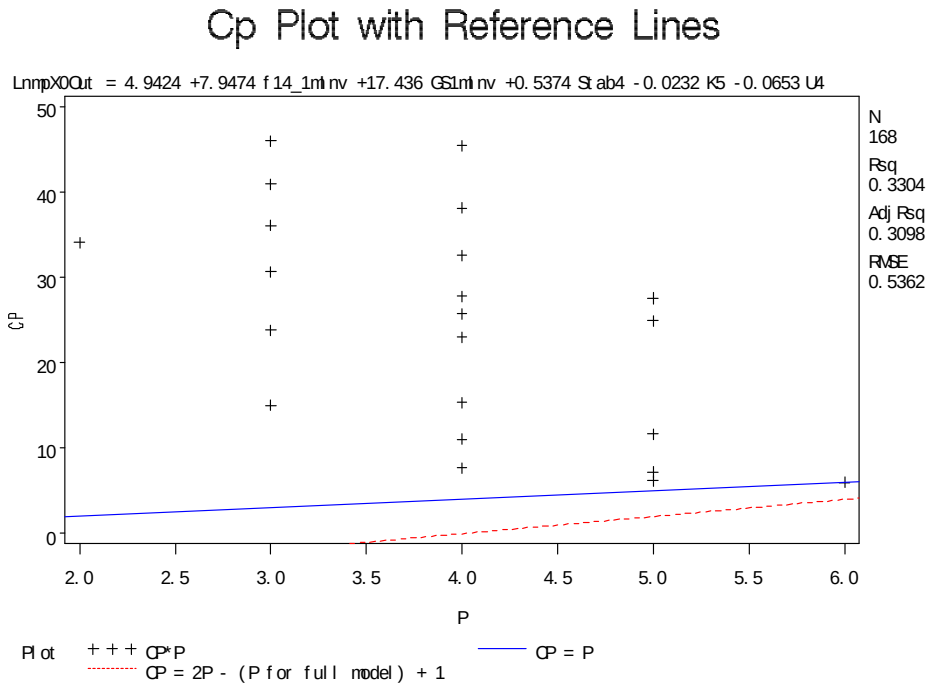

Figure A-4. Cp Plot of Model of *m,p*-Xylene after Removing the Outliers

o-Xylene

Bivariate Pearson Correlation

The correlation coefficients between ln-transformed *o*-xylene concentrations and the distance to urban interstate (FC11) roadways and the distance to urban major arterial (FC14) roadways were -0.147 ($p=0.048$) and -0.148 ($p=0.046$), respectively. The distance to the US Highway Route 1 also had a statistically significantly correlation coefficient of -0.266, $p=0.0003$. The correlation coefficient between ln-transformed ambient air concentration of *o*-xylene and the inverse distance to the closest gas station was 0.24 ($p=0.0011$). The refinery was the only point source whose distance to the residences had a statistically significant correlation with the ln-transformed *o*-xylene concentrations (0.174, $p=0.019$).

The meteorological variables that were statistically significantly correlated with *o*-xylene concentrations were wind speed (-0.30, $p<0.0001$), atmospheric stability (0.427, $p<0.0001$), mixing heights (-0.28, $p=0.0002$), relative humidity (0.16, $p=0.027$), and temperature (-0.11, $p<0.15$). Precipitation and atmospheric pressure were not correlated with the residential ambient air concentration of *o*-xylene.

Preliminary Selection of Predictors

A series of preliminary regression analyses for each group of variables were performed using the ln-transformed *o*-xylene concentrations to determine which variables to include in the model. The distances to the closest gas station, the refinery, and the urban major arterial roadways (FC14) were selected as important predictors among the variables that describe the distance between sources and residences. From the meteorological variables, wind speed, temperature, and stability were selected as predictor variables ($p<0.15$).

Selection of the Best-fitting Model

The predictor variables selected by the different regression model selection methods for the residential ambient air concentration of *o*-xylene were relatively consistent. The meteorological variables, which were consistently included in the series of regression model, were the atmospheric stability, temperature, and wind speed, in order of selection. The C(p) was 7, the same as the number of parameters included in model. The parameter estimates were significant ($p<0.05$), except for the intercept ($p=0.26$). The model statistics are summarized in Table A-3. As illustrated in Figure A-5, the residuals were distributed relatively random. The PP plot was nearly linear implying that the error term of the model followed a normal distribution. Twelve data points were identified as possible Outliers by using a test statistic of ±1.645 (0.95, df=175, Figure A-6). The analysis of the variance, parameter estimates, and the summary of model statistics for the best-fitting 7-parameter model for *o*-xylene after removal of outliers are listed in Table A-4. The selected model was statistically significant ($p<0.0001$). The residuals for the model with the outliers removed were randomly distributed without showing any obvious trend or any particular pattern (Figure A-7). The standardized residuals of the best-fitting model were close to a normal distribution and had the constant variances. Based on a visual diagnosis on residual plot, probability plot, and quantile plot, there was no evidence of a lack of fit or unequal error variance for the best-fitting 7-parameter regression model for the ambient residential *o*-xylene. The Mallows' C_p statistic associated with this particular subset of variables was determined at 7.0, indicating that the resulting model had the appropriate number of parameters.

Diagnostics of Equal Variances and Multicollinearity Diagnostics

To test the assumption of equal variance, the heteroscedasticity of the parameter estimates were tested. The chi-square was 28 with a probability of 0.104, a value greater than 0.05 (Appendix C). Therefore, the variances of the parameter estimates could be concluded as not being significantly different. As a consequence, the equal error variances in parameter estimates were assumed in the best-fitting 7-parameter model. The same considerations about parameter correlations expressed for *m*/*p* xylene also apply to *o* xylene.

Table A-3. Results of the Best-fitting 7-Parameter Model for *o*-Xylene

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	39.22208	7.84442	17.24	<.0001
Error	177	80.5563	0.45512		
Corrected Total	182	119.7784			

Root MSE	0.67463	R-Square	0.3275
Dependent Mean	-0.09397	Adjusted R-Square	0.3085
Coefficient of Variation	-717.895		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	2.28427	2.01835	1.13	0.2593
F14_1mInv	(Distance to FC14)$^{-1}$	1	20.31620	4.66133	4.36	<.0001
GS1mInv	(Distance to Gas Station)$^{-1}$	1	13.94798	7.70577	1.81	0.0720
Stab4	Atmospheric Stability	1	0.63575	0.13279	4.79	<.0001
K5	Temperature	1	-0.01848	0.00604	-3.06	0.0025
U4	Wind speed	1	-0.11480	0.05331	-2.15	0.0326

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Atmospheric Stability	0.1907	0.1907	31.3278	42.65	<.0001
2	(Distance to FC14)$^{-1}$	0.0802	0.2709	12.4773	19.81	<.0001
3	Temperature	0.0285	0.2994	7.0815	7.27	0.0077
4	Wind Speed	0.0156	0.3150	5.0189	4.06	0.0454
5	(Distance to Gas Station)$^{-1}$	0.0124	0.3275	3.7836	3.28	0.0720

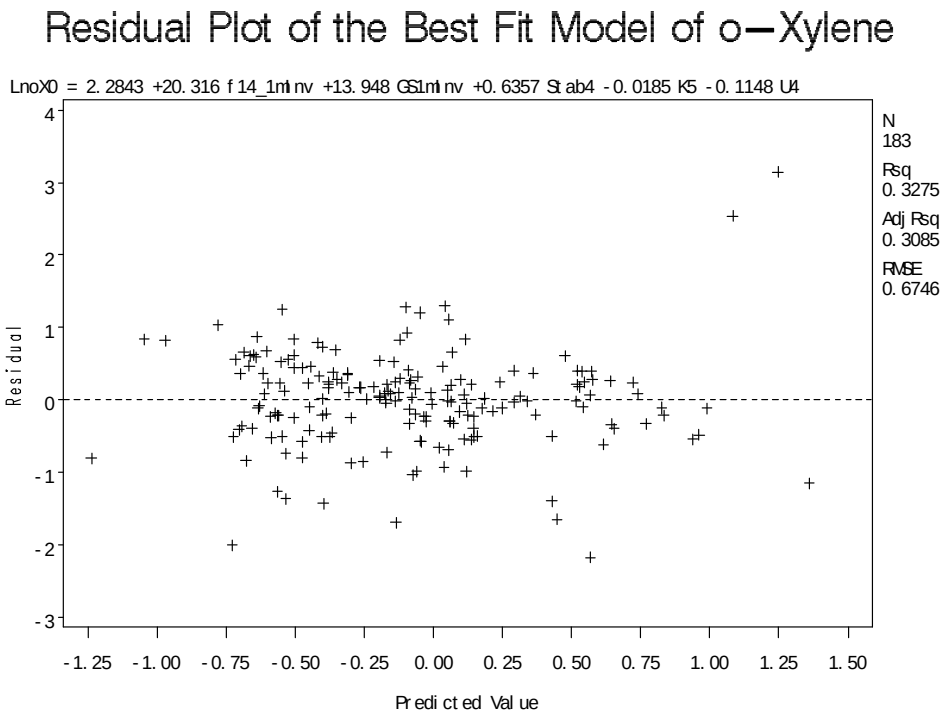

Figure A-5. Residual Plot of the Model of *o*-Xylene

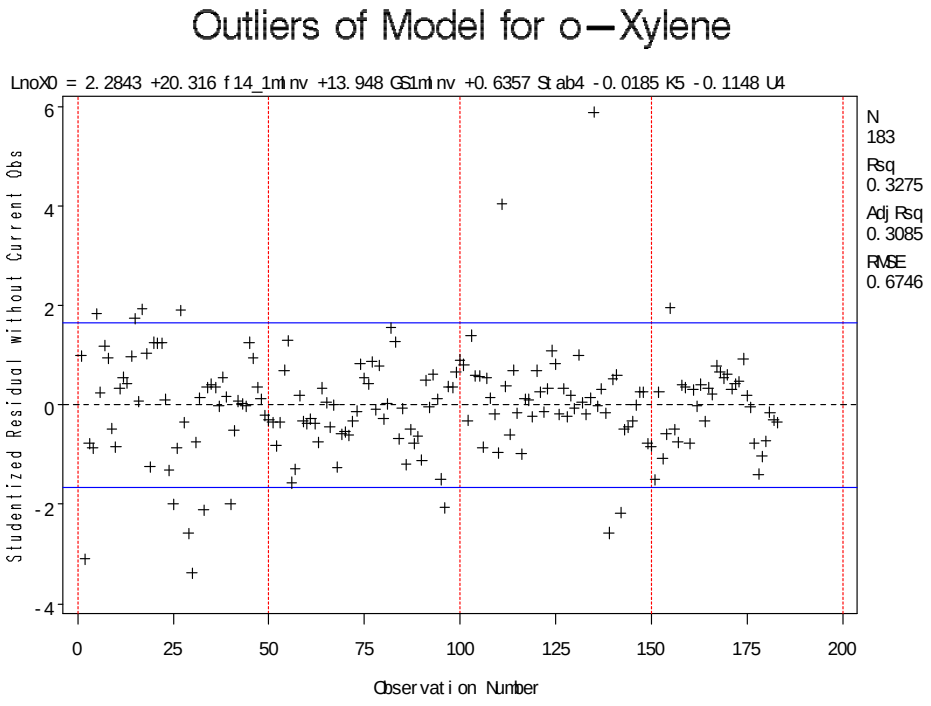

Figure A-6. Outliers of Model of *o*-Xylene

Table A-4 Results of the Best-fitting 7-Parameter Regression Model for *o*-Xylene after Removing of the Outliers

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	23.35913	4.67183	23.69	<.0001
Error	162	31.95114	0.19723		
Corrected Total	167	55.31027			

Root MSE	0.44410	R-Square	0.4223
Dependent Mean	-0.09736	Adjusted R-Square	0.4045
Coefficient of Variation	-456.16000		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	4.45813	1.40740	3.17	0.0018
F14_1mInv	(Distance to FC14)$^{-1}$	1	7.44373	4.48291	1.66	0.0988
GS1mInv	(Distance to Gas Station)$^{-1}$	1	9.54244	5.47996	1.74	0.0835
Stab4	Atmospheric Stability	1	0.52092	0.09234	5.64	<.0001
K5	Temperature	1	-0.02352	0.00419	-5.62	<.0001
U4	Wind speed	1	-0.12197	0.03697	-3.30	0.0012

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Atmospheric Stability	0.2717	0.2717	38.3416	61.92	<.0001
2	Temperature	0.0877	0.3594	15.9706	22.59	<.0001
3	Wind Speed	0.0353	0.3947	8.1589	9.57	0.0023
4	(Distance to Gas Station)$^{-1}$	0.0178	0.4125	5.2175	4.93	0.0277
5	(Distance to FC14)$^{-1}$	0.0098	0.4223	4.4861	2.76	0.0988

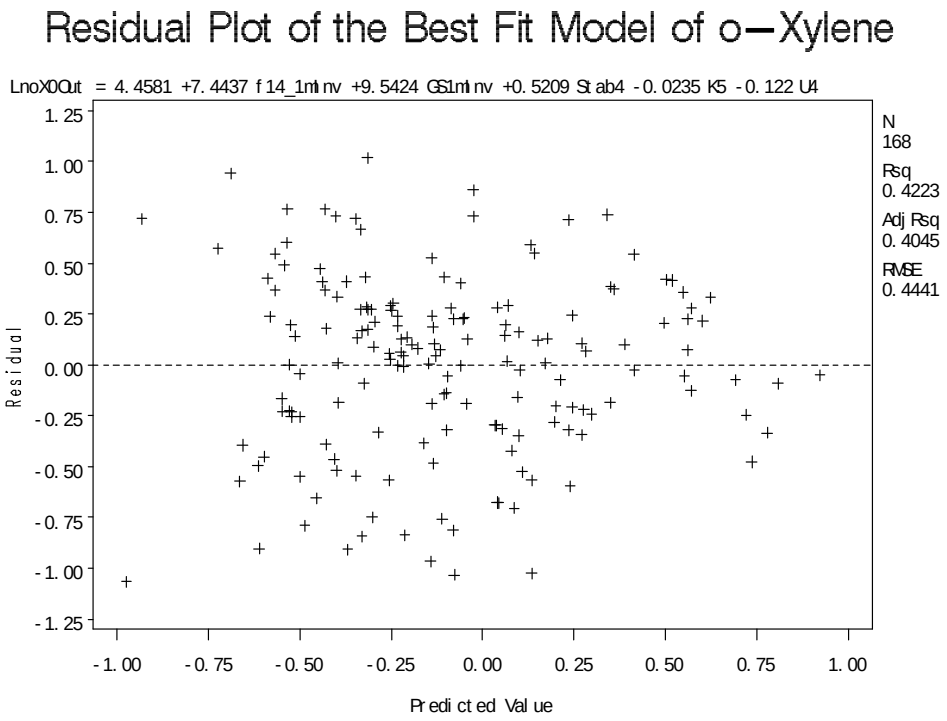

Figure A-7- Residual vs. Predicted Plot of the Best-fitting 7-Parameter Model of *o*-Xylene after Removing the Outliers

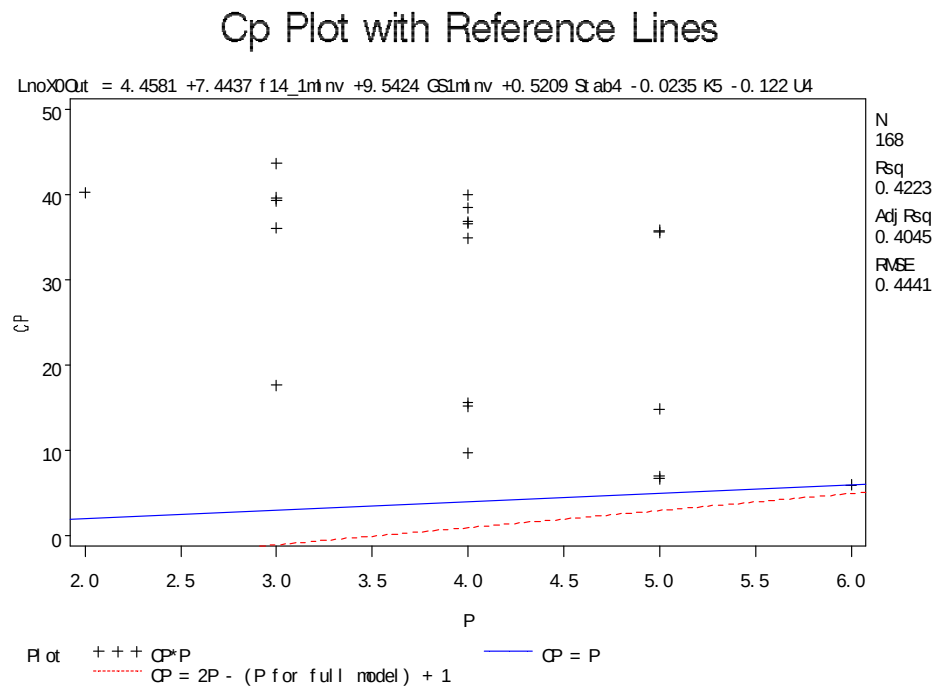

Table 4.A- Results of the Test of Multicollinearity of Predictor Variables Included in the Best-fitting Model for *o*-Xylene

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Tolerance	Variance Inflation
Intercept	1	5.19346	1.43598	3.62	0.0004	.	0
f14_1mInv	1	8.75248	4.30991	2.03	0.0439	0.92783	1.07778
GS1mInv	1	9.80788	5.25048	1.87	0.0636	0.91647	1.09114
Stab4	1	0.53003	0.08902	5.95	<.0001	0.72367	1.38185
K4	1	-0.02635	0.00443	-5.94	<.0001	0.93148	1.07355
U4	1	-0.13256	0.03548	-3.74	0.0003	0.69736	1.43397

Collinearity Diagnostics

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation: Intercept	f14_1mInv	GS1mInv	Stab4	K4	U4
1	4.83871	1.00000	0.00002176	0.01278	0.01262	0.00022105	0.00002821	0.00172
2	0.63347	2.76376	0.00004494	0.37172	0.34247	0.00041480	0.00005795	0.00400
3	0.46833	3.21432	3.48031E-9	0.60663	0.62097	0.00000123	4.828461E-8	0.00002712
4	0.05572	9.31914	0.00038414	0.00002814	0.00518	0.01731	0.00064391	0.57135
5	0.00347	37.36627	0.02094	0.00109	0.01645	0.88898	0.05375	0.27248
6	0.00030792	125.35642	0.97861	0.00776	0.00231	0.09308	0.94552	0.15042

Table 4.4.4. Results of the Test of Heteroscedasticity of Parameter Estimates Determined in the Best-fitting Model for *o*-Xylene

Consistent Covariance of Estimates

Variable	Intercept	f14_1mInv	GS1mInv	Stab4	K4	U4
Intercept	1.6675971894	0.6946045492	-1.338990966	-0.041187547	-0.004861092	-0.017907343
f14_1mInv	0.6946045492	12.512236841	-2.924001768	0.0253987847	-0.003166084	0.0004076095
GS1mInv	-1.338990966	-2.924001768	15.630366869	-0.007104998	0.0047403892	-0.010061741
Stab4	-0.041187547	0.0253987847	-0.007104998	0.0068550523	3.7178761E-7	0.0013929763
K4	-0.004861092	-0.003166084	0.0047403892	3.7178761E-7	0.0000168767	0.000016424
U4	-0.017907343	0.0004076095	-0.010061741	0.0013929763	0.000016424	0.0015020694

Test of First and Second Moment Specification

DF	Chi-Square	Pr > ChiSq
20	28.23	0.1040

4.5. Toluene

4.5.1. Bivariate Pearson Correlation

The correlation coefficients between ln-transformed toluene concentration and the distance to FC14 roadways and for inverse distance to FC14 were –0.177 ($p=0.018$) and 0.172 ($p=0.02$), respectively. The distance to the US Highway Route 1 also was statistically significantly correlated to the ln-transformed concentration of toluene in the residential ambient air (-0.162, $p=0.03$). The correlation coefficient between ln-transformed toluene concentration and the distance to the closest gas station was –0.18 ($p=0.015$). The refinery was the only point source whose distance from the residence to the facility had a statistically significant correlation for the ln-transformed toluene concentration in ambient air (–0.15, $p=0.04$).

The meteorological variables that correlated with toluene concentration were atmospheric stability (0.31, $p<0.0001$), wind speed (-0.198, $p<0.01$), mixing heights (-0.19, $p<0.01$), relative humidity (0.17, $p<0.05$), temperature (-0.135, $p<0.1$), and precipitation (0.115, $p<0.15$). Atmospheric pressure was not correlated with the residential ambient air concentration of toluene.

Preliminary Selection of Predictors

The distances to the refinery, and the distance to major urban arterial roadways (FC14) were selected as important predictors among the variables that describe the distance between sources to residences. From the meteorological variables, wind speed and atmospheric stability were selected as predictor variables ($p<0.15$).

Selection of the Best-fitting Model

The predictor variables selected by the different selection methods for regression model for the residential ambient air concentration of toluene from the proximity variables and the meteorological variables were consistent. The meteorological variables included in the regression model were the atmospheric stability, and temperature. The inverse distance to the closest major urban arterial roadway (FC14) was included in the model as a predictor among the proximity to the major roadway variables. The inverse distance to the refinery was included as significant predictor variables in the model among the distance variables to point source. The distance to the gas station was not selected as a predictor for the model of toluene. The model was statistically significant ($p<0.0001$) with an r^2 of 0.199 (adjusted r^2 of 0.181).

The parameter estimates for the meteorological predictors were significant at $p<0.01$ and the parameter estimates for the proximity variables in the model were significant at $p<0.15$. The model statistics are summarized in Table A-5. The residuals were distributed relatively random (Figure A-9). The error term of the model followed a normal distribution, based on the linearity observed in PP plot (Appendix B). The possible outliers were determined by using the test statistics of ± 1.645 (0.95, df. = 165) to improve by removing the less contributing Outliers (Figure A-10). The analysis of the variance, parameter estimates, and the summary of model statistics for the best-fitting 5-parameter model for toluene are listed in Table A-5. The removal of the outlier improved the r^2, from 0.20 to 0.33. The residuals were randomly distributed without showing any obvious trend or any particular pattern based on a visual examination (Figure A-11). The standardized residuals of the best-fitting model appear to be close to a normal distribution and had constant variances. The residual, PP, QQ plots, there appear to show no visual evidence of lack of fit or unequal error variance. The Mallows' C_p statistic associated with this particular subset of variables

was determined to be 5, indicating the resulting model had appropriate number of parameters.

Diagnostics of Equal Variances and Multicollinearity Diagnostics

To test the assumption of equal variance, the heteroscedasticity of the parameter estimates were tested as well as multicollinearity (Appendix C). The chi-square was 28.15 with a probability of 0.014, a value smaller than 0.05. Therefore, the variances of the parameter estimates were concluded as significantly different. As a consequence, the error variances in parameter estimates could not be assumed as equal for the best-fitting 5-parameter model of toluene. The bivariate Pearson correlations between pairs of predictors included in the model showed some statistically significant correlations were identified between 'the inverse distance to the refinery' and 'the closest distance to the urban major arterial roadways (FC14)' (-0.284, $p<0.0001$).

The variance inflations for predictor variables were close to 1 (1.01 ~ 1.11) which is not greater than 10. Based on the variation inflation factors, there was no significant collinearity between the predictors in the model. However, as a result of the collinearity diagnostics, the condition index was 107, which was greater than 100, and the eigenvalue was close to zero (0.00038), which was smaller than 0.01. The proportion of variation of intercept (0.98) and temperature (0.97) were greater than 0.5, indicating that the two parameters interacted. Therefore, there were possible co-dependences in the model, which might overspecify the model outcome.

Table A-5. Results of the 5-Parameter Multiple linear regression Model for Toluene

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	39.48395	9.87099	12.03	<.0001
Error	178	146.1108	0.82085		
Corrected Total	182	185.5947			

Root MSE	0.90601	R-Square	0.2127
Dependent Mean	1.52498	Adjusted R-Square	0.1951
Coefficient of Variation	59.41107		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	6.29278	2.41687	2.60	0.0100
f14_1mInv	(Distance to FC14)$^{-1}$	1	16.66451	6.14926	2.71	0.0074
Stab4	Atmospheric Stability	1	0.65082	0.16197	4.02	<.0001
K4	Temperature	1	-0.03208	0.00830	-3.87	0.0002
RH5	Relative Humidity	1	0.01554	0.00613	2.54	0.0120

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Atmospheric Stability	0.1162	0.1162	23.6494	23.8	<.0001
2	Temperature	0.0384	0.1546	16.8377	8.18	0.0047
3	(Distance to FC14)$^{-1}$	0.0296	0.1843	12.0408	6.50	0.0116
4	Relative Humidity	0.0285	0.2127	7.5148	6.44	0.0120

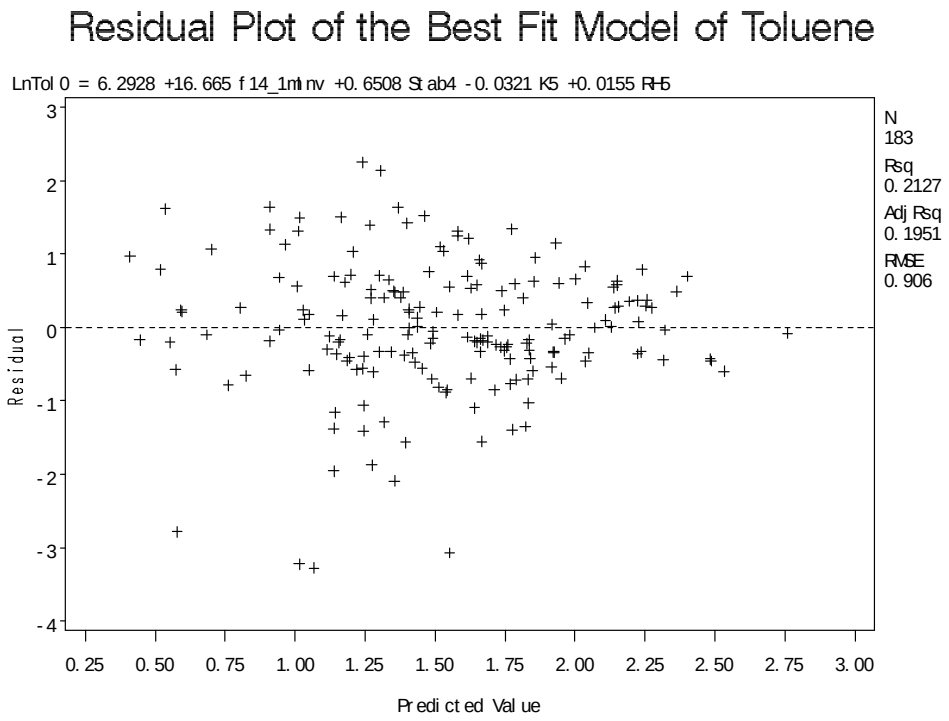

Figure A-9. Residual vs. Predicted Plot of the 5-Parameter Model of Toluene

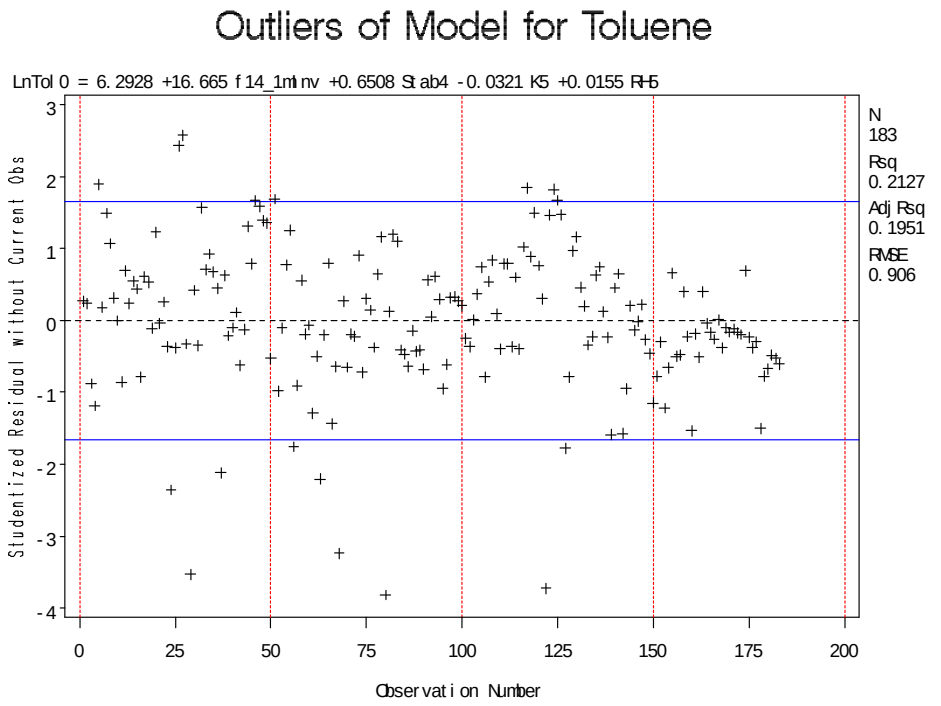

Figure A-10. Outliers of Model of Toluene

Table A-6. Results of the Best-fitting 5-Parameter Model for Toluene after Removing the Outliers

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	30.71102	7.67776	18.42	<.0001
Error	162	67.51842	0.41678		
Corrected Total	166	98.22944			

Root MSE	0.64559	R-Square	0.3126
Dependent Mean	1.60929	Adjusted R-Square	0.2957
Coefficient of Variation	40.11608		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	3.11017	1.84905	1.68	0.0945
f14_1mInv	(Distance to FC14)$^{-1}$	1	14.72149	4.43634	3.32	0.0011
Stab4	Atmospheric Stability	1	0.70584	0.12046	5.86	<.0001
K4	Temperature	1	-0.02067	0.00635	-3.25	0.0014
RH5	Relative Humidity	1	0.01116	0.00480	2.33	0.0212

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Atmospheric Stability	0.2245	0.2245	18.6321	47.76	<.0001
2	(Distance to FC14)$^{-1}$	0.0376	0.2620	11.8339	8.35	0.0044
3	Temperature	0.0276	0.2897	7.3616	6.34	0.0128
4	Relative Humidity	0.0230	0.3126	3.9805	5.42	0.0212

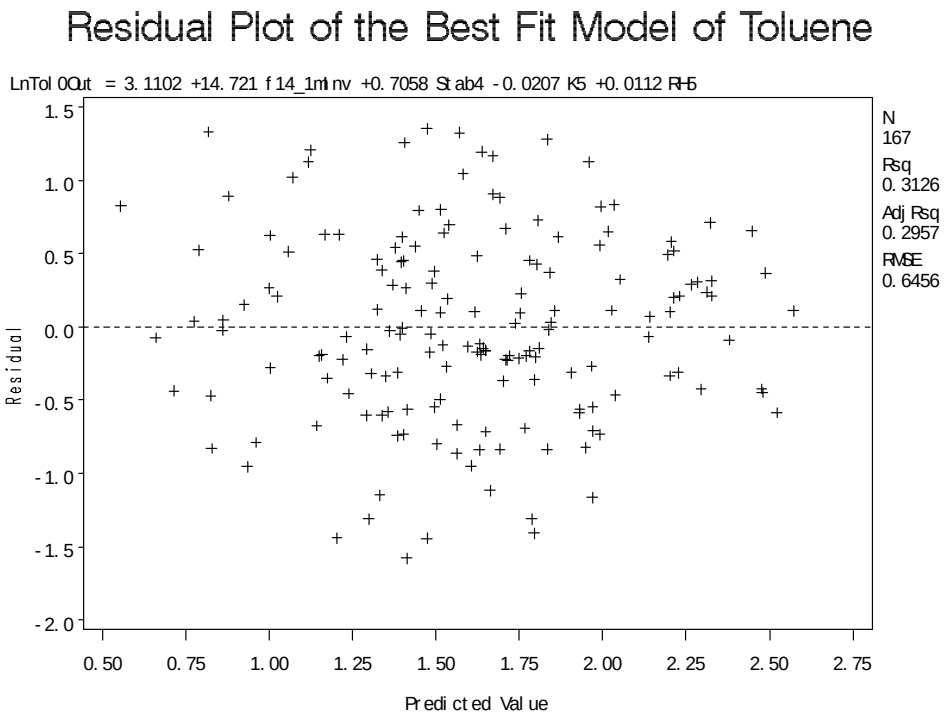

Figure A-11. Residual vs. Predicted Plot of the Best-fitting 5-Parameter Model of Toluene after Removing the Outliers

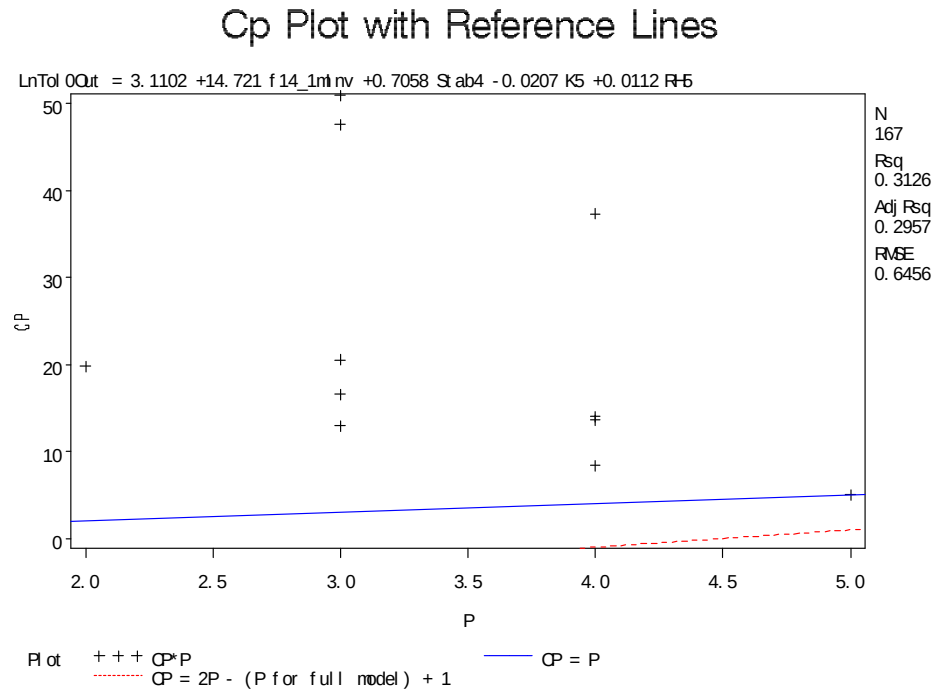

Figure A-12. Cp Plot for the Best-fitting 5-Parameter Model for Toluene after Removing the Outliers

4.6. Benzene

4.6.1. Bivariate Pearson Correlation

A negative correlation coefficient of -0.196(p=0.008) was determined for the distance to the US Highway Route 1 and natural ln-transformed benzene concentrations. The correlation coefficient between ln-transformed benzene concentration and the inverse distance to the closest gas station was significant (0.259, p=0.0004). Among the distances to the four identified point sources, no point source was significantly correlated with benzene concentration.

The meteorological variables that correlated with benzene concentrations were temperature (-0.392, p<0.0001), atmospheric stability (0.263, p=0.0003), mixing heights (-0.224, p=0.0024), wind speed (-0.167, p=0.025), and atmospheric pressure (0.165, p=0.026). Precipitation and relative humidity were not significantly correlated with the benzene concentration.

Preliminary Selection of Predictors

The distances to the closest gas station, the refinery, and the US highway Route 1 were selected as important predictors of ambient benzene concentration (p<0.15). Wind speed, temperature, and atmospheric stability were selected as predictor variables from the meteorological variables (p<0.15).

Selection of the Best-fitting Model

The predictor variables selected by the various selection methods in models for the residential ambient air concentration of benzene, were relatively consistent. The meteorological variables which were consistently included were: temperature, atmospheric

stability, and wind speed in order of selection. The C(p) suggested the 6-parameter model was appropriate. The model statistics and parameter estimates are summarized in Table A-7. The residuals were distributed relatively randomly (Figure A-13) suggesting that there was equal variance in residuals. The probability plot (Appendix B) was linear indicating that the error term of the model followed a normal distribution.

Possible outliers were identified based on a test statistic of ± 1.645 (0.95, df. = 175) (Figure A-14). After the removal of sixteen possible Outliers, the model became the 5-parameter model because the wind speed was not included in the best-fitting regression model. The analysis of the variance, parameter estimates, and the summary of model statistics for the best-fitting 5-parameter model for benzene are listed in Table A-8. A visual examination of the residuals indicated that they were randomly distributed without showing any obvious trend or any particular pattern (Figure A-15). The standardized residual of the best-fitting model was close to a normal distribution with constant variances. There was no visual evidence for the lack of a fit or of significant unequal error variance for the best-fitting 5-parameter regression model for the residential ambient air benzene concentration. The Mallows' C_p statistic associated with this particular subset of variables was determined at 5, suggesting the model result had appropriate number of parameters included.

Diagnostics of Equal Variances and Multicollinearity Diagnostics

To test the assumption of equal variance, the heteroscedasticity of the parameter estimates were tested as well as multicollinearity. The chi-square was 18.85 with a probability of 0.17, a value greater than 0.05. Therefore, the variances of the parameter estimates could be concluded as not significantly different. As a consequence, the equal error variances in parameter estimates were assumed in the best-fitting 5-parameter model. The bivariate

Pearson correlations between pairs of predictors included in the model showed statistically significant correlations between 'the inverse distance to the closest gas station' and 'the inverse distance to the closest urban major arterial roadways (FC14)' (0.184, p=0.013).

The variance inflations for predictor variables were close to 1 (1.01 ~ 1.06) which is not greater than 10. Based on the variation inflation factors, there was no significant collinearity between the predictors in the model. However, as a result of the collinearity diagnostics, the condition index was 103, which was greater than 100, and the eigenvalue was close to zero (0.00036), which was smaller than 0.01.

Table A-7. Results of the Best-fitting 6-Parameter Model for Benzene

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	26.86251	8.95417	20.62	<.0001
Error	179	77.73533	0.43428		
Corrected Total	182	104.5978			

Root MSE	0.6590	R-Square	0.2568
Dependent Mean	0.1278	Adjusted R-Square	0.2444
Coefficient of Variation	515.6321		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	11.46110	1.74199	6.58	<.0001
GS1mInv	(Distance to Gas Station)$^{-1}$	1	26.16590	7.30908	3.58	0.0004
K4	Temperature	1	-0.03743	0.00585	-6.4	<.0001
U4	Wind speed	1	-0.17239	0.04456	-3.87	0.0002

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Temperature	0.1407	0.1407	27.2654	29.64	<.0001
2	Wind Speed	0.0629	0.2036	14.1672	14.22	0.0002
3	(Distance to Gas Station)$^{-1}$	0.0532	0.2568	3.3947	12.82	0.0004

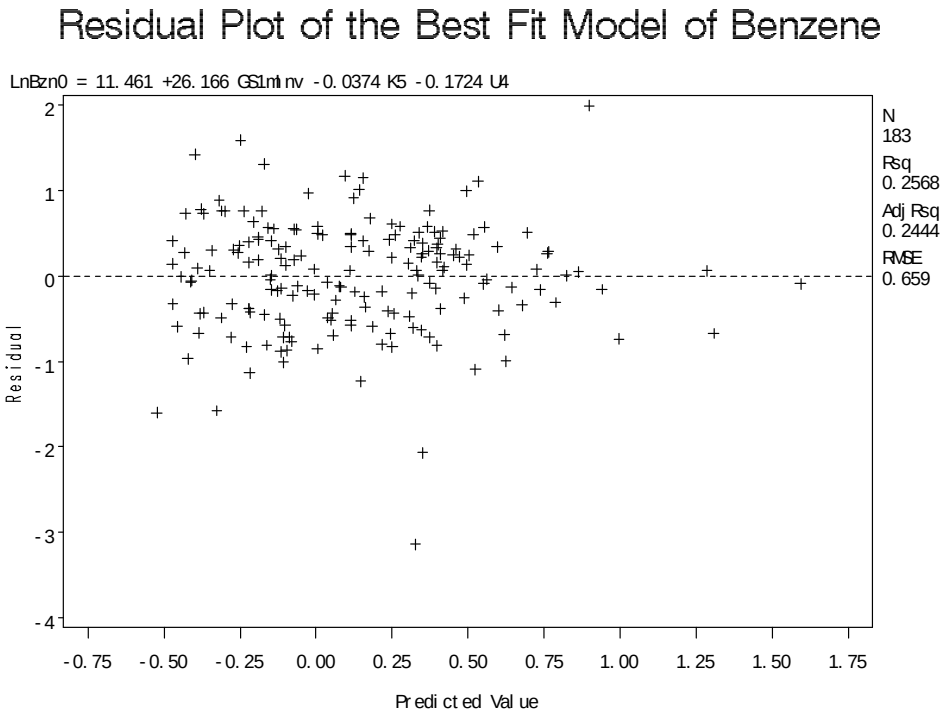

Figure A-13. Residual vs. Predicted Plot of the 6-Parameter Model of Benzene

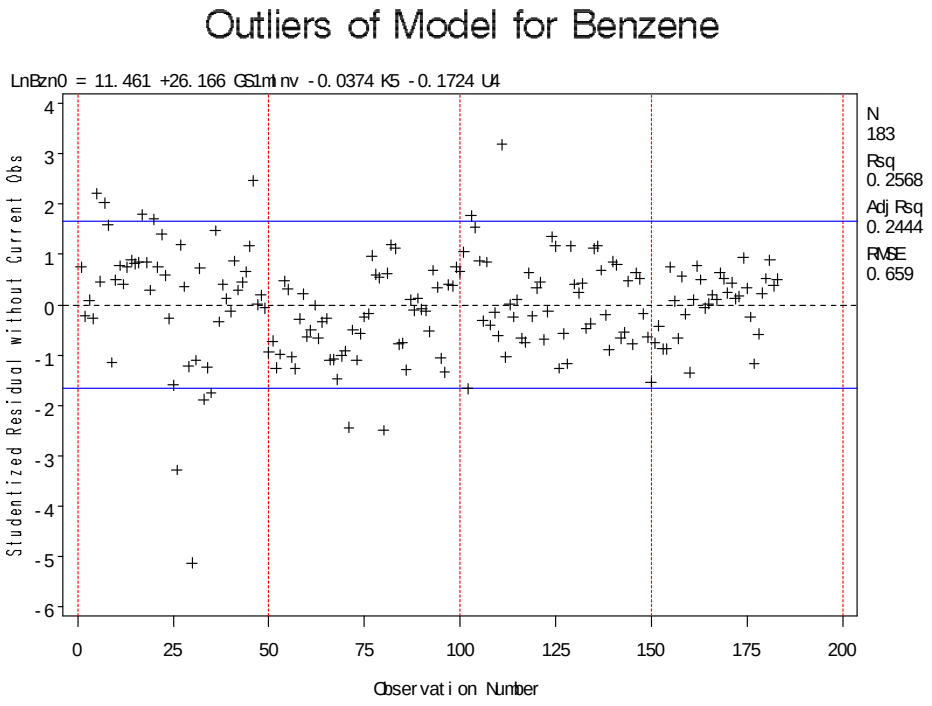

Figure A-14. Outliers of 6-Parameter Model of Benzene

Table A-8. Results of the Best-fitting 5-Parameter Model for Benzene after Removing the Outliers

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	26.0494	5.20987	22.8	<.0001
Error	163	37.2387	0.22846		
Corrected Total	168	63.2881			

Root MSE	0.4780	R-Square	0.4116
Dependent Mean	0.1438	Adjusted R-Square	0.3936
Coefficient of Variation	332.5000		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	10.07440	1.49805	6.73	<.0001
F14_1mInv	(Distance to FC14)$^{-1}$	1	5.49770	3.32981	1.65	0.1007
GS1mInv	(Distance to Gas Station)$^{-1}$	1	16.14780	5.57504	2.90	0.0043
Stab4	Atmospheric Stability	1	0.30356	0.09971	3.04	0.0027
K4	Temperature	1	-0.03914	0.00447	-8.76	<.0001
U4	Wind Speed	1	-0.08488	0.03966	-2.14	0.0338

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Temperature	0.2510	0.2510	41.1680	55.95	<.0001
2	Atmospheric Stability	0.0996	0.3506	15.7541	25.46	<.0001
3	(Distance to Gas Station)$^{-1}$	0.0340	0.3846	8.3926	9.12	0.0029
4	Wind Speed	0.0172	0.4018	5.6629	4.71	0.0314
5	(Distance to FC14)$^{-1}$	0.0098	0.4116	4.9545	2.73	0.1007

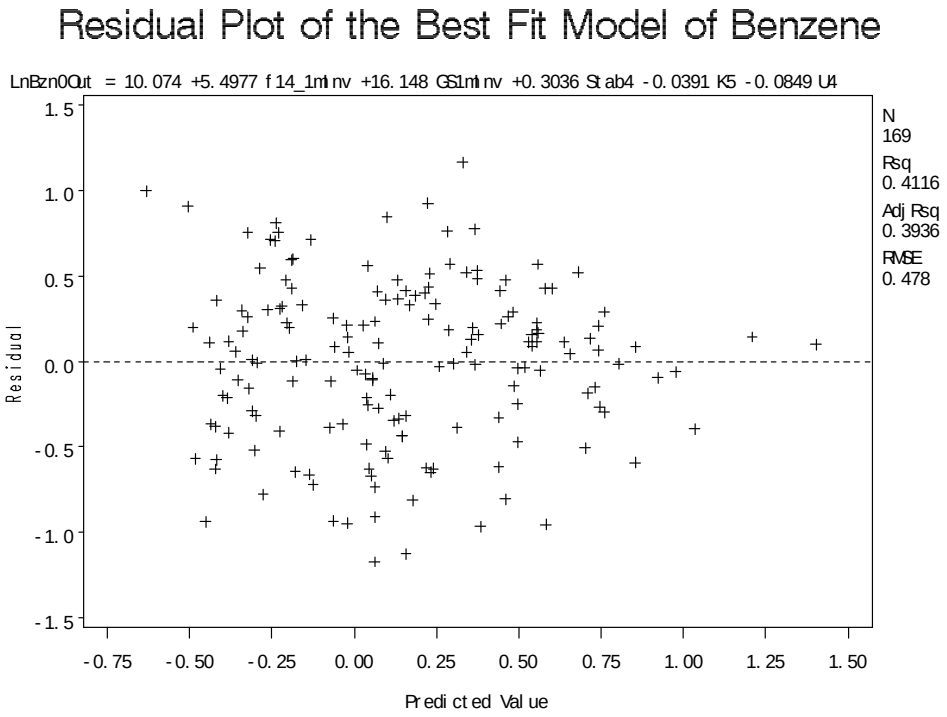

Figure A-15. Residual vs. Predicted Plot of the Best-fitting 5-Parameter Model of Benzene after Removing the Outliers

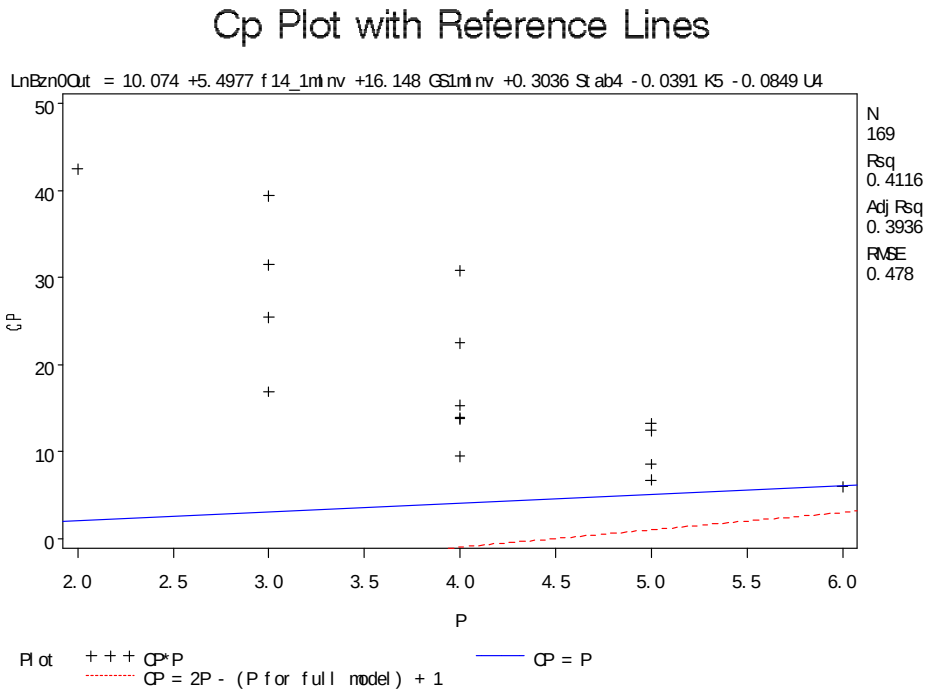

Figure A-16. Cp Plot for the Best-fitting 5-Parameter Model for Benzene after Removing the Outliers

4.7. Ethylbenzene

4.7.1. Bivariate Pearson Correlation

Ethylbenzene concentration was not correlated significantly with the any of the proximity variables of roadway classification directly or following any transformations. The inverse distance from the sampler location to US highway Route 1 (one of the individual roadways of FC14) showed significant correlation for un-transformed ambient air concentration of ethylbenzene (0.167, $p=0.024$). The inverse distance to the closest gas station had significant correlation for un-transformed ambient air concentration of ethylbenzene (0.206, $p=0.005$). There were two identified point sources of ethylbenzene in the study area, but only the distance to the refinery was correlated with ethylbenzene concentration ($p<0.1$).

The meteorological variables that correlated with ln-transformed ethylbenzene concentrations were atmospheric stability (0.27, $p=0.0003$), wind speed (-0.17, $p=0.024$), and temperature (-0.14, $p=0.06$). Mixing height, relative humidity, precipitation, and atmospheric pressure were not significantly correlated with the ambient air concentration of ethylbenzene.

4.7.2. Preliminary Selection of Predictors

A series of preliminary regression analyses for each group of variable were performed using the ln-transformed ethylbenzene concentrations to determine which variables to include in the model. The distances to the closest gas station was selected as an important predictor among the variables that describe the distance between sources and residences. From the meteorological variables, atmospheric stability was selected as a predictor variable ($p<0.15$).

4.7.3. Selection of the Best-fitting Model

The predictor variables selected in the models by the different selection methods for the residential ambient air concentration of ethylbenzene were relatively consistent. Atmospheric stability and temperature were selected among the meteorological variables as predictor variables. The inverse square distance to the urban major arterial roadways (FC14) was selected as a significant predictor in the model of ethylbenzene among the source proximity variables. The C(p) was 4.3, close to the number of parameters (4) included in model. The parameter estimates were significant at $p<0.05$ for the meteorological variables (atmospheric stability and temperature). The model statistics are summarized in Table A-9. The residuals were distributed relatively randomly (Figure A-17) suggesting that there was equal variance in residuals. The probability plot (Appendix C) was linear indicating that the error term of the model followed a normal distribution. Possible outliers were identified by using a test statistics of ± 1.645 (0.95, df. = 175) (Figure A-18). The analysis of the variance, parameter estimates, and the summary of model statistics for the best-fitting 4-parameter model for ethylbenzene after removal of outliers are listed in Table A-10. The removal of the seventeen outliers did not improve the r^2 as much as observed from the models of the other VOCs in this research. However, the probability of parameter estimates of the best-fitting model was improved for all variables ($p<0.05$).

The residuals for the model with the outliers removed were more randomly distributed (Figure A-11) compared to the distribution before removal (Figure A-9). The standardized residuals of the best-fitting model were close to a normal distribution and had constant variances. The probability plot showed the linearity of the error term followed a normal distribution. Based on a visual diagnosis, there was no evidence of a lack of fit or

unequal error variance for the best-fitting 4-parameter regression model for the ambient residential ethylbenzene. The Mallows' C_p statistic associated with this particular subset of variables was determined at 4 (Figure A-12), indicating that the resulting model had the appropriate number of parameters.

Diagnostics of Equal Variances and Multicollinearity Diagnostics

To test the assumption of equal variance, the heteroscedasticity of the parameter estimates were tested as well as the mulitcollinearity. The chi-square was 16.88 with a probability of 0.0506, a value slightly greater than 0.05. Therefore, the variances of the parameter estimates could be concluded as not significantly different. As a consequence, the equal error variances in parameter estimates were assumed in the best-fitting 4-parameter model. The bivariate Pearson correlations between pairs of predictors did not identify any statistically significant correlations between predictors in the model at □ =0.05.

The variance inflations for predictor variables were close to 1 (1.007 ~ 1.023) a value not greater than 10. Based on the variation inflation factors, there was no significant collinearity between the predictors in the model. As a result of the collinearity diagnostics, the condition index was 91, which was smaller than 100. However, the eigenvalue was close to zero (0.00037), which was smaller than 0.01. The proportion of variation of intercept (0.98) and temperature (0.97) were greater than 0.5, indicating that the two parameters were interacted. As described for *m,p* xylene some codependency among the variable exist..

Table A-9. Results of the Best-fitting 4-Parameter Model for Ethylbenzene

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	32.8909	8.22271	6.56	<.0001
Error	178	223.166	1.25374		
Corrected Total	182	256.057			

Root MSE	1.1197	R-Square	0.1285
Dependent Mean	-0.2723	Adjusted R-Square	0.1089
Coefficient of Variation	-411.2200		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	5.32543	3.34905	1.59	0.1136
F14_1mInv	(Distance to FC14)$^{-1}$	1	13.31980	7.59097	1.75	0.0810
Stab4	Atmospheric Stability	1	0.52795	0.21804	2.42	0.0165
K5	Temperature	1	-0.02653	0.00999	-2.66	0.0086
U4	Wind Speed	1	-0.17165	0.08826	-1.94	0.0534

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Atmospheric Stability	0.0741	0.0741	7.2130	14.49	0.0002
2	Temperature	0.0204	0.0945	5.1088	4.06	0.0455
3	Wind Speed	0.0189	0.1134	3.3142	3.81	0.0525
4	(Distance to FC14)$^{-1}$	0.0151	0.1285	2.2823	3.08	0.0810

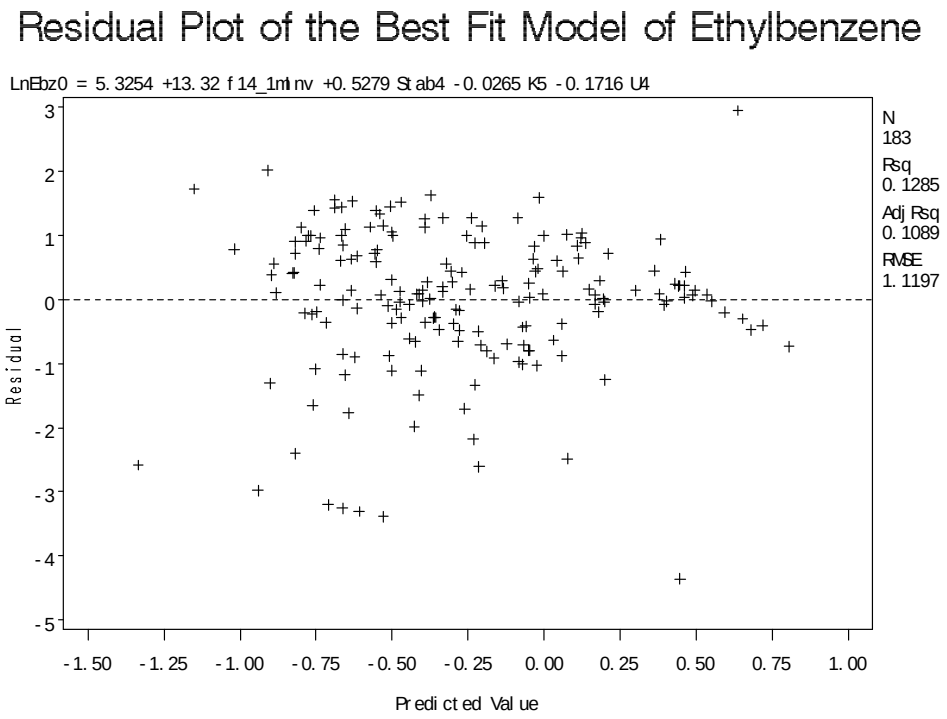

Figure A-17. Residual vs. Predicted Plot of the 4-Parameter Model of Ethylbenzene

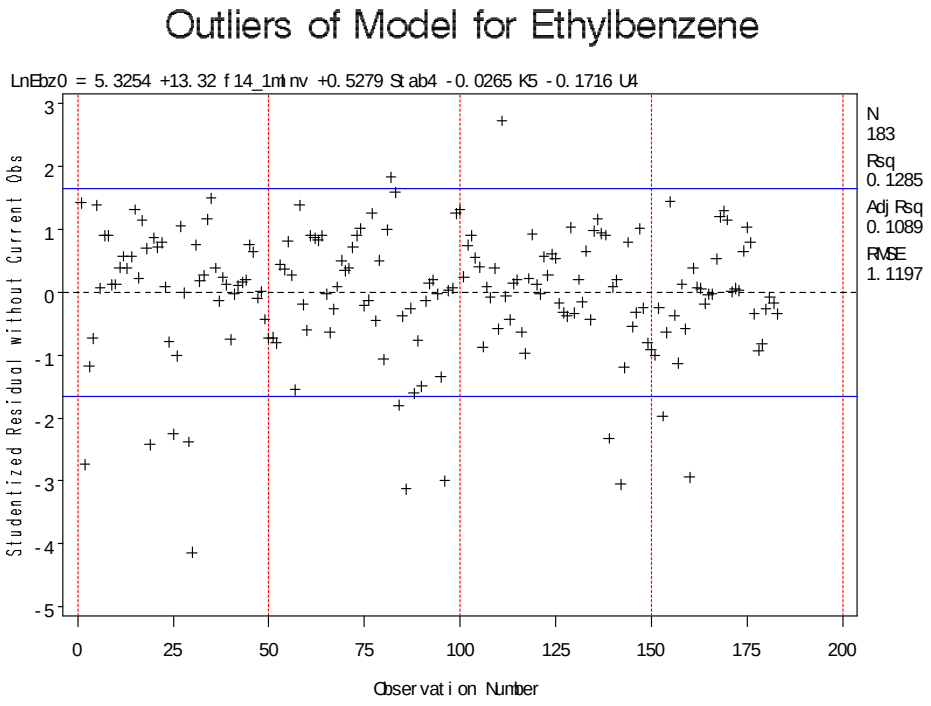

Figure A-18. Outliers of 4-Parameter Model of Ethylbenzene

Table A-10. Results of the Best-fitting 4-Parameter Multiple linear regression Model for Ethylbenzene After Removing the Outliers

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	20.4924	5.12309	7.96	<.0001
Error	164	105.564	0.64368		
Corrected Total	168	126.056			

Root MSE	0.8023	R-Square	0.1626
Dependent Mean	-0.1066	Adjusted R-Square	0.1421
Coefficient of Variation	-752.6100		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	5.98125	2.43806	2.45	0.0152
F14_1mInv	(Distance to FC14)$^{-1}$	1	9.68110	5.62338	1.72	0.0870
Stab4	Atmospheric Stability	1	0.43775	0.16287	2.69	0.0079
K4	Temperature	1	-0.02747	0.00732	-3.75	0.0002
U4	Wind Speed	1	-0.11372	0.06587	-1.73	0.0861

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Atmospheric Stability	0.0774	0.0774	13.4719	14.01	0.0002
2	Temperature	0.0552	0.1326	4.8002	10.56	0.0014
3	Wind Speed	0.0149	0.1474	3.9236	2.88	0.0917
4	(Distance to FC14)$^{-1}$	0.0151	0.1626	2.9959	2.96	0.0870

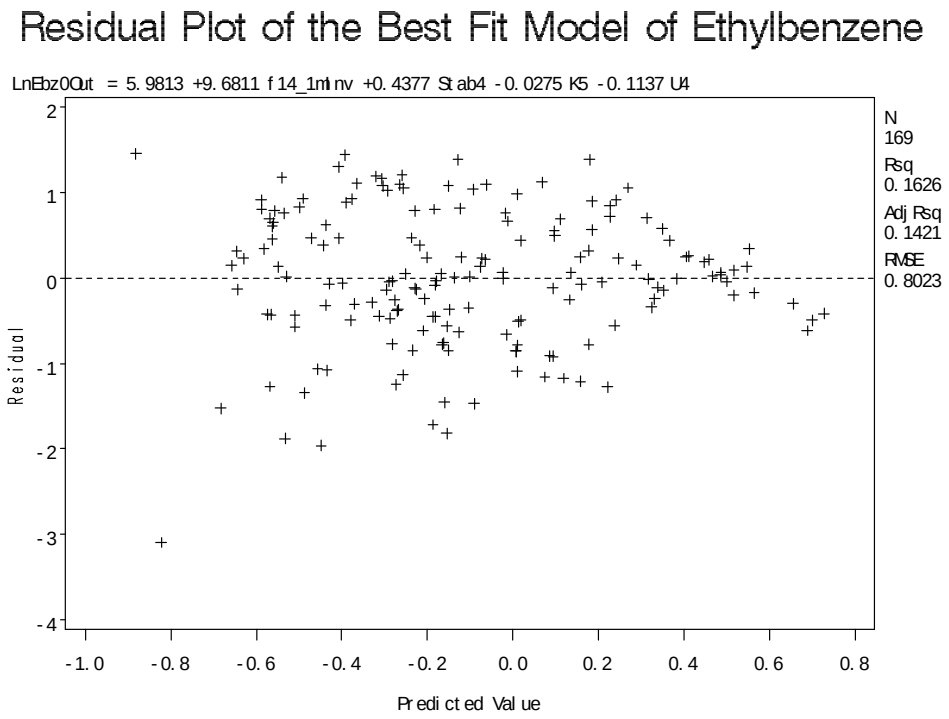

Figure A-19. Residual vs. Predicted Plot of the Best-fitting 4-Parameter Model of Ethylbenzene after Removing the Outliers

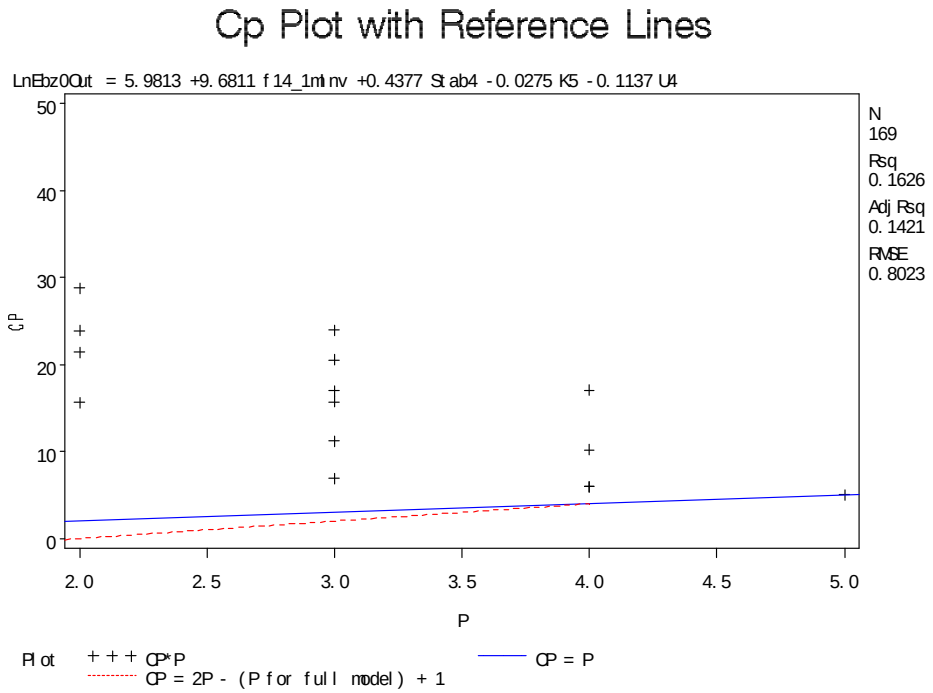

Figure A-20. Cp Plot for the Best-fitting 4-Parameter Model for Ethylbenzene after Removing the Outliers

Methyl tert Butyl Ether (MTBE)

Bivariate Pearson Correlation

The correlation coefficients between the untransformed ambient air concentration of MTBE and the distance and the inverse distance to the nearest FC 11 (interstate highways in urban) was 0.22, $p=0.0027$. The only identified point source of MTBE within 3 kilometers of Elizabeth, NJ, which generated more than 0.9 tons in 1999, was the refinery in Linden. However, the distance to the refinery was not significantly correlated to the MTBE air concentration.

The meteorological variables that were significantly correlated with the ln-transformed MTBE air concentrations were: atmospheric stability (0.296, $p<0.0001$), wind speed (-0.265, $p=0.0004$), relative humidity (0.196, $p=0.0094$), and temperature (0.173, $p=0.022$). The atmospheric pressure and precipitation were not correlated with the MTBE air concentrations.

Preliminary Selection of Predictors

The distances from the air sampler to the closest gas station was selected as a predictor of ambient air concentration of MTBE at $p<0.15$. The distance to the major roadways and refinery were not selected as significant predictors. Atmospheric stability was selected as predictors from the meteorological variables.

Selection of the Best-fitting Model

The predictor variables selected by the different regression model selection methods for the residential ambient air concentration of MTBE were relatively consistent. The meteorological variables, which were consistently included in the series of regression model,

were the atmospheric stability, temperature, and wind speed, in order of selection. The distance to the closest interstate roadways (FC11) and the distance to the closet major urban arterial roadways (FC14) were not selected as significant predictor variables in the model of MTBE. The distance to the closest gas station was included as a significant predictor variable in the model of MTBE. The model statistics are summarized in Table A-11. The residuals were relatively randomly distributed but had some irregular pattern in error variances (Figure A-21). The PP plot was nearly linear implying that the error term of the model followed a normal distribution (Appendix B). Fourteen data points were identified as possible Outliers were identified using test statistics of ± 1.655 at 0.95, df=165 (Figure A-22). The analysis of the variance, parameter estimates, and the summary of model statistics for the best-fitting 5-parameter model for MTBE after removing outliers are listed in Table A-12. After removing the Outliers, the parameter estimates became more significant for all variables. A visual examination of the residuals indicated that they were randomly distributed without showing any obvious trend or any particular pattern (Figure A-23). The standardized residuals of the best-fitting model were close to a normal distribution with constant variances. The PP plot was nearly linear implying that the error term of the model followed a normal distribution (Appendix C). There was no visual evidence for the lack of fit or unequal error variance for the best-fitting 5-parameter regression model for the residential ambient air MTBE concentrations. The Mallows' C_p statistic associated with this particular subset of variables was 5.0. Since the number of parameters (p) including the intercept in the best-fitting model was 5, the resulting model had the appropriate number of parameters.

Diagnostics of Equal Variances and Multicollinearity Diagnostics

To test the assumption of equal variance, the heteroscedasticity of the parameter estimates were tested as well as multicollinearity. The chi-square was 15.55 with a probability of 0.34, a value greater than 0.05. Therefore, the variances of the parameter estimates could be concluded as not significantly different. As a consequence, the equal error variances in parameter estimates were assumed in the best-fitting 5-parameter model. The bivariate Pearson correlations between pairs of predictors included in the model identified statistically significant correlations between the wind speed and atmospheric stability (-0.51, $p<0.0001$), and between the wind speed and temperature (-0.25, $p=0.0007$).

The variance inflations for seven predictor variables were close to 1 (1.03 ~ 1.46) which is not greater than 10. Based on the variation inflation factors, there was no significant collinearity between the predictors in the model. However, as a result of the collinearity diagnostics, the condition index was 115, which was greater than 100, and the eigenvalue was close to zero (0.00033), which was smaller than 0.01. The proportion of variation of intercept (0.98) and temperature (0.94) were greater than 0.5, indicating that the two parameters were interacted. Therefore, between some predictors, there were possible co-dependences, which might overspecify the model outcome.

Table A-11. Results of the Best-fitting 5-Parameter Model for MTBE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	23.4697	11.7349	8.23	0.0004
Error	180	256.619	1.42566		
Corrected Total	182	280.089			

Root MSE	1.1940	R-Square	0.0838
Dependent Mean	1.2495	Adjusted R-Square	0.0736
Coefficient of Variation	95.5620		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	1.97438	0.35393	5.58	<.0001
GS1mInv	(Distance to Gas Station)$^{-1}$	1	39.49380	13.21040	2.99	0.0032
U4	Wind speed	1	-0.21592	0.07758	-2.78	0.0060

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	(Distance to Gas Station)$^{-1}$	0.0444	0.0444	5.9898	8.40	0.0042
2	Wind Speed	0.0394	0.0838	0.3584	7.75	0.0060

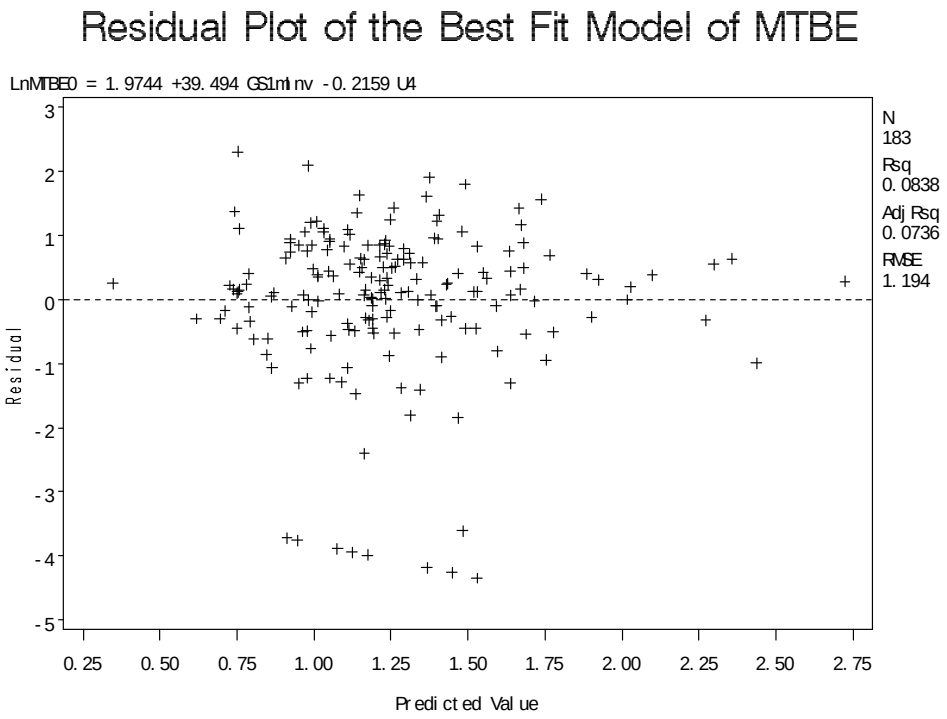

Figure A-21. Residual vs. Predicted Plot of the Best-fitting 5-Parameter Model of MTBE

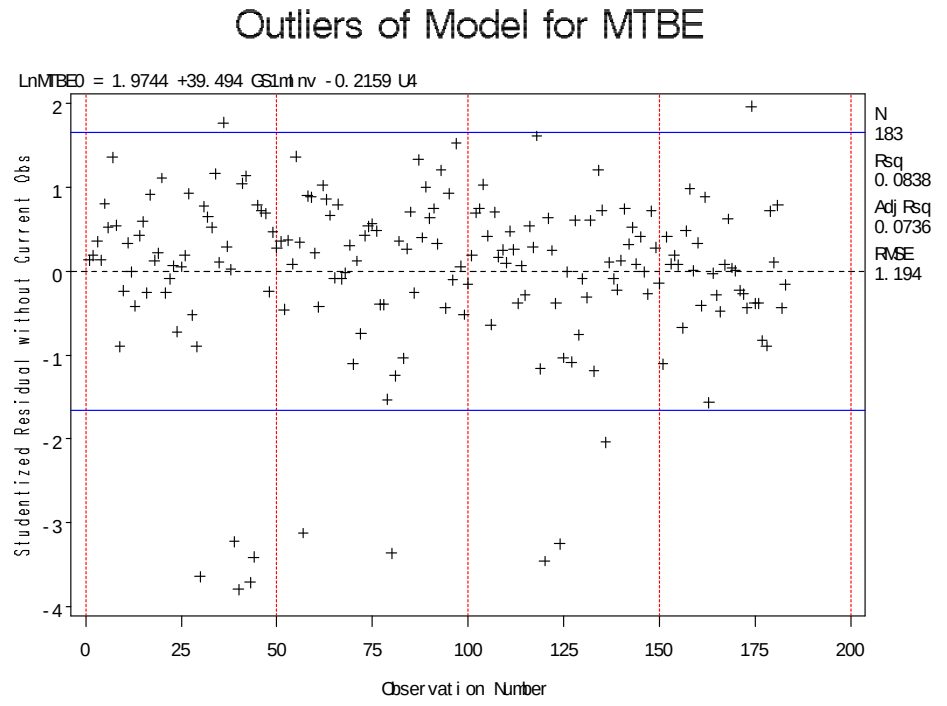

Figure A-22. Outliers of 5-Parameter Model of MTBE

Table A-12. Results of the Best-fitting 5-Parameter Model for MTBE after Removing the Outliers

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	29.1548	5.83095	10.97	<.0001
Error	165	87.6716	0.53134		
Corrected Total	170	116.826			

Root MSE	0.7289	R-Square	0.2496
Dependent Mean	1.4525	Adjusted R-Square	0.2268
Coefficient of Variation	50.1840		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	-2.74300	2.27071	-1.21	0.2288
F11_1mInv	(Distance to FC11)$^{-1}$	1	22.25470	14.30720	1.56	0.1217
GS1mInv	(Distance to Gas Station)$^{-1}$	1	33.56270	8.28221	4.05	<.0001
Stab4	Atmospheric Stability	1	0.24348	0.14963	1.63	0.1056
K5	Temperature	1	0.01239	0.00669	1.85	0.0659
U4	Wind speed	1	-0.18702	0.06022	-3.11	0.0022

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Wind Speed	0.1226	0.1226	23.1543	23.62	<.0001
2	(Distance to Gas Station)$^{-1}$	0.0897	0.2123	5.7156	19.13	<.0001
3	Temperature	0.0136	0.2259	4.7657	2.94	0.0885
4	Atmospheric Stability	0.0126	0.2386	4.0309	2.75	0.0991
5	(Distance to FC11)$^{-1}$	0.0110	0.2496	3.6458	2.42	0.1217

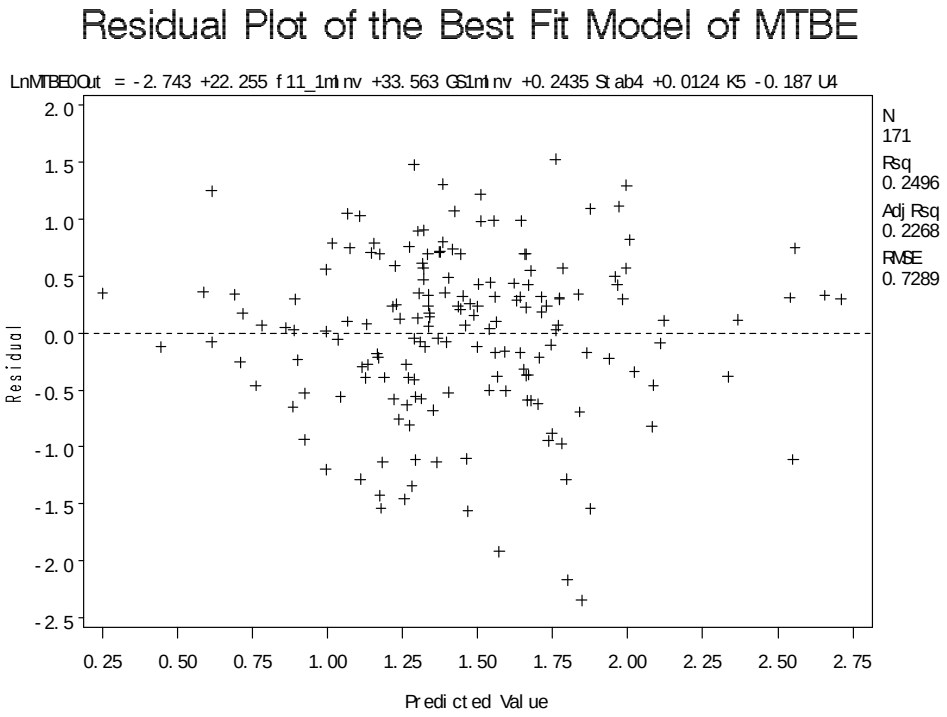

Figure A-23. Residual vs. Predicted Plot of the Best-fitting 5-Parameter Model of MTBE after Removing the Outliers

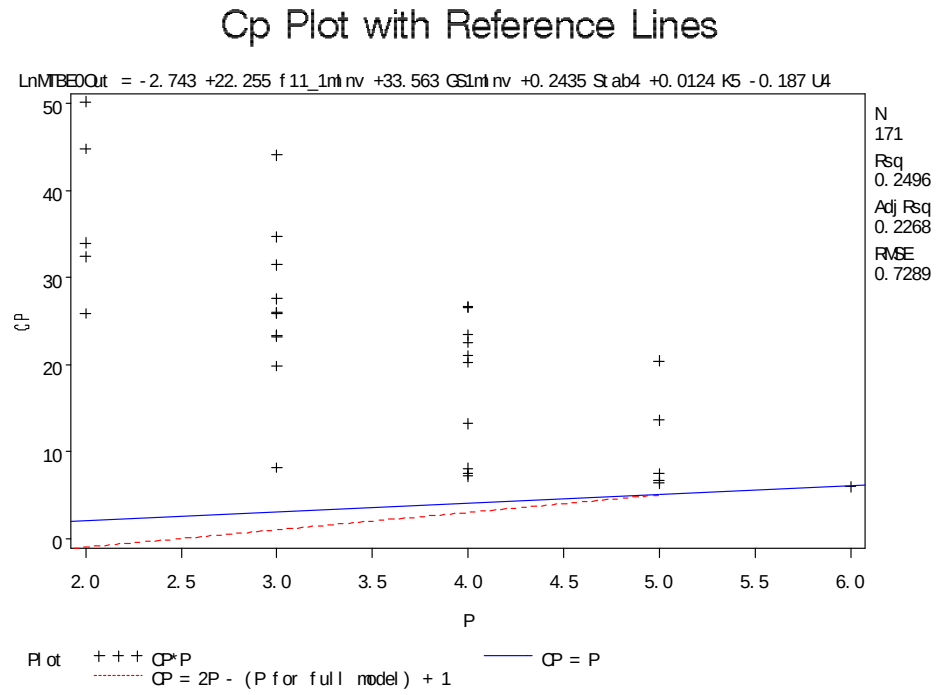

Figure A-24. Cp Plot for the Best-fitting 5-Parameter Model for MTBE after Removing the Outliers

4.9. Tetrachloroethylene (PCE)

4.9.1. Bivariate Pearson Correlation

The distance to the closest dry cleaning facility (DCF1) was statistically significantly correlated with the ln-transformed ambient air concentration of PCE at □ =0.01, regardless of form of transformation of the distance variable. The refinery was the only identified point source of the PCE with the combined annual generation of 1.0 ton. The distance to the refinery was not correlated significantly with PCE in any of the analysis. A statistically significant correlation was found at $p<0.01$ were the distance to the US Highway Route 1 and the distance to the closest gas station. Other proximity variables to roadway were not significantly correlated.

The meteorological variables that were significantly correlated with PCE concentrations were wind speed (-0.373, $p<0.0001$), relative humidity (0.313, $p<0.0001$), atmospheric stability (0.282, $p=0.0001$), mixing heights (-0.254, $p=0.0005$), and precipitation (0.235, $p=0.0013$). Temperature and atmospheric pressure was not correlated with the residential ambient air concentration of PCE.

4.9.2. Preliminary Selection of Predictors

A series of preliminary regression analyses were performed on the ln-transformed PCE concentration to determine which variables to include in the model. The distances to the closest dry cleaning facility was selected as an important predictor of ambient PCE concentration from the variables that describe the distance between sources and residences. The wind speed, precipitation, and stability were selected as predictor variables ($p<0.15$) from the meteorological variables.

Selection of the Best-fitting Model

The predictor variables selected by the different regression model selection methods for the residential ambient air concentration of PCE were relatively consistent. The meteorological variables, which were consistently included in the series of regression model, were the wind speed, temperature, atmospheric stability, and relative humidity in order of selection. The inverse distance to the closest dry cleaning facility was selected as a significant predictor variable in the best-fitting model of PCE, while the distance to major roadways or gas station were not selected as expected. The model statistics are summarized in Table A-13. The residuals were relatively randomly distributed but had few outliers in error variances (Figure A-25). The PP plot was nearly linear implying that the error term of the model followed a normal distribution. Eight data points were identified as possible Outliers were identified using test statistics of ± 1.655 at 0.95, df=165 (Figure A-26). The analysis of the variance, parameter estimates, and the summary of the model statistics for the best-fitting 6-parameter model for PCE after removing the outliers are listed in Table A-14. The selected model was statistically significant ($p<0.0001$).

A visual examination of the residuals indicated that they were randomly distributed without showing any obvious trend or any particular pattern (Figure A-26). The standardized residuals of the best-fitting model were close to a normal distribution with constant variances. The PP plot was nearly linear implying that the error term of the model followed a normal distribution. There was no visual evidence of lack of fit or unequal error variance for the best 6-parameter regression model for the residential ambient air PCE concentration. The Mallows' C_p statistic associated with this particular subset of variables was determined at 6.0 (Figure A-27). Since the number of parameters (p) including the intercept in the best-fitting model was 6, the resulting model was appropriate in number of parameters.

Diagnostics of Equal Variances and Multicollinearity Diagnostics

To test the assumption of equal variance, the heteroscedasticity of the parameter estimates were tested as well as multicollinearity. The chi-square was 18.1 with a probability of 0.58, a value greater than 0.05. Therefore, the variances of the parameter estimates could be concluded as not significantly different. As a consequence, the equal error variances in parameter estimates were assumed in the best-fitting 6-parameter model. The bivariate Pearson correlations showed statistically significant correlations were identified between the wind speed and atmospheric stability (-0.51, $p<0.0001$), and between the wind speed and temperature (-0.25, $p=0.0007$). Relative humidity was also significantly correlated with atmospheric stability (0.36, $p<0.0001$), and with wind speed (-0.43, $p<0.0001$).

The variance inflations for seven predictor variables were close to 1 (1.04 ~ 1.44) which is not greater than 10. Based on the variation inflation factors, there was no significant collinearity between the predictors in the model. However, as a result of the collinearity diagnostics, the condition index was 130, which was greater than 100, and the eigenvalue was close to zero (0.00033), which was smaller than 0.01

Table A-13. Results of the Best-fitting 6-Parameter Model for PCE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	25.689	6.42224	11.85	<.0001
Error	178	96.5014	0.54214		
Corrected Total	182	122.190			

Root MSE	0.7363	R-Square	0.2102
Dependent Mean	-0.3394	Adjusted R-Square	0.1925
Coefficient of Variation	-216.9400		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	-0.65455	0.86751	-0.75	0.4515
DCF1mInv	(Distance to DCF)$^{-1}$	1	54.38580	21.18230	2.57	0.0111
Stab4	Atmospheric Stability	1	0.21071	0.14312	1.47	0.1427
U4	Wind speed	1	-0.22587	0.05600	-4.03	<.0001
Precip5	Precipitation	1	0.01026	0.00388	2.65	0.0088

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Wind Speed	0.1389	0.1389	13.2138	29.20	<.0001
2	(Distance to DCF)$^{-1}$	0.0314	0.1704	8.1961	6.82	0.0098
3	Precipitation	0.0303	0.2006	3.4399	6.78	0.0100
4	Atmospheric Stability	0.0096	0.2102	3.2933	2.17	0.1427

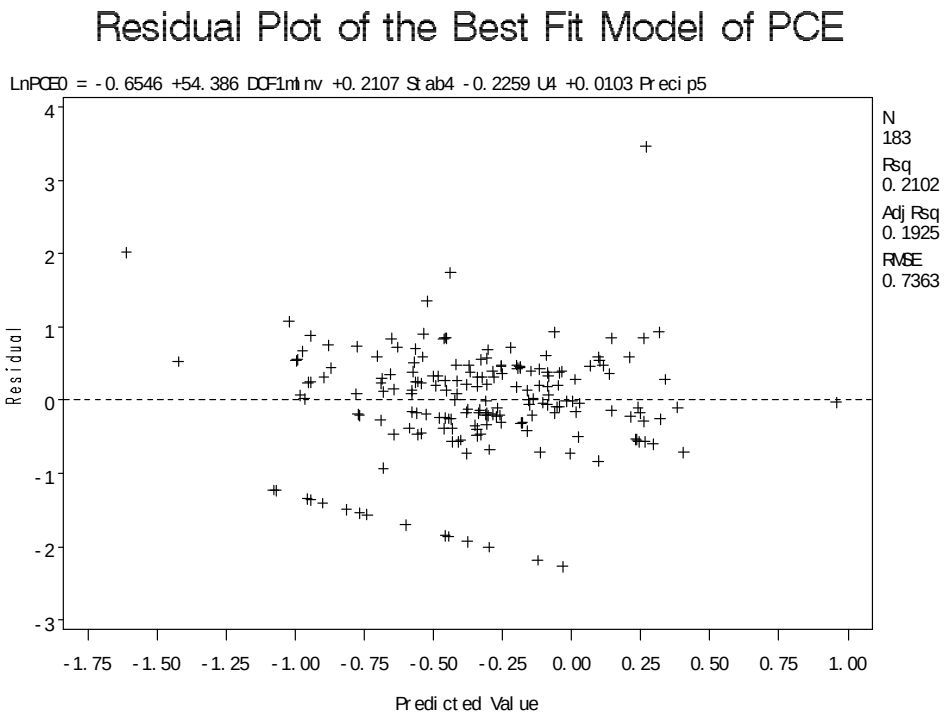

Figure A-25. Residual vs. Predicted Plot of the Best-fitting 6-Parameter Model of PCE

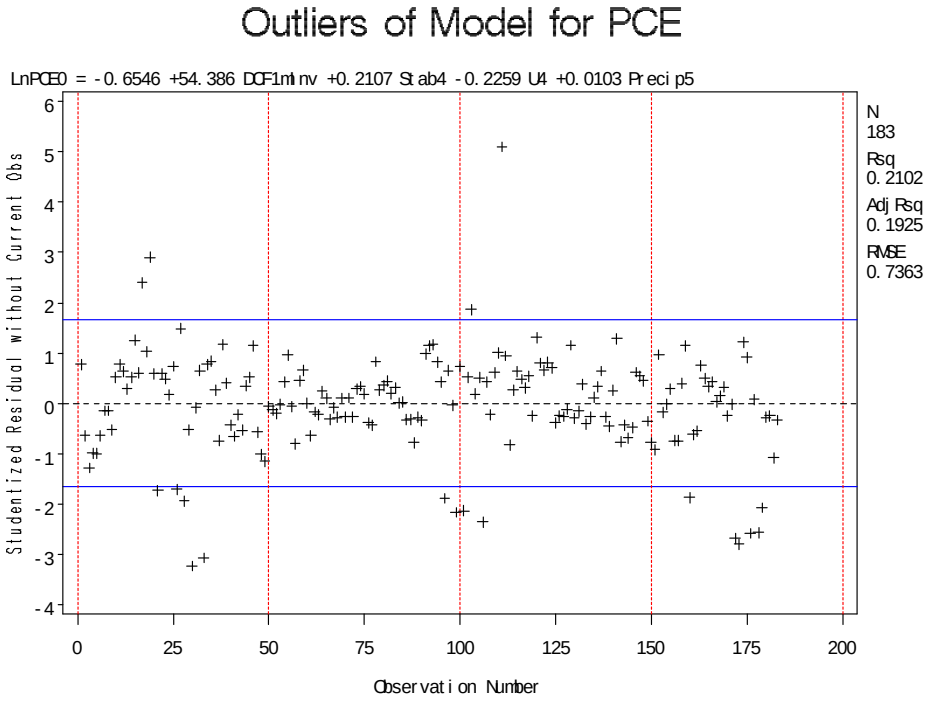

Figure A-26. Outliers of 6-Parameter Model for PCE

Table A-14. Results of the Best-fitting 6-Parameter Model for PCE after Removing the Outliers

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	12.5816	2.51633	14.36	<.0001
Error	158	27.6904	0.17526		
Corrected Total	163	40.2721			

Root MSE	0.4186	R-Square	0.3124
Dependent Mean	-0.2064	Adjusted R-Square	0.2907
Coefficient of Variation	-202.8800		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	2.49450	1.34715	1.85	0.0659
DCF1mInv	(Distance to DCF)$^{-1}$	1	32.67340	12.39640	2.64	0.0092
Stab4	Atmospheric Stability	1	0.14442	0.08831	1.64	0.1040
K5	Temperature	1	-0.01229	0.00416	-2.96	0.0036
U4	Wind speed	1	-0.14410	0.03588	-4.02	<.0001
RH4	Relative Humidity	1	0.00913	0.00301	3.04	0.0028

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Wind Speed	0.1847	0.1847	25.6505	36.70	<.0001
2	(Distance to DCF)$^{-1}$	0.0381	0.2228	18.9710	7.90	0.0056
3	Relative Humidity	0.0317	0.2545	13.7554	6.80	0.0100
4	Temperature	0.0463	0.3008	5.2151	10.53	0.0014
5	Atmospheric Stability	0.0116	0.3124	4.5652	2.67	0.1040

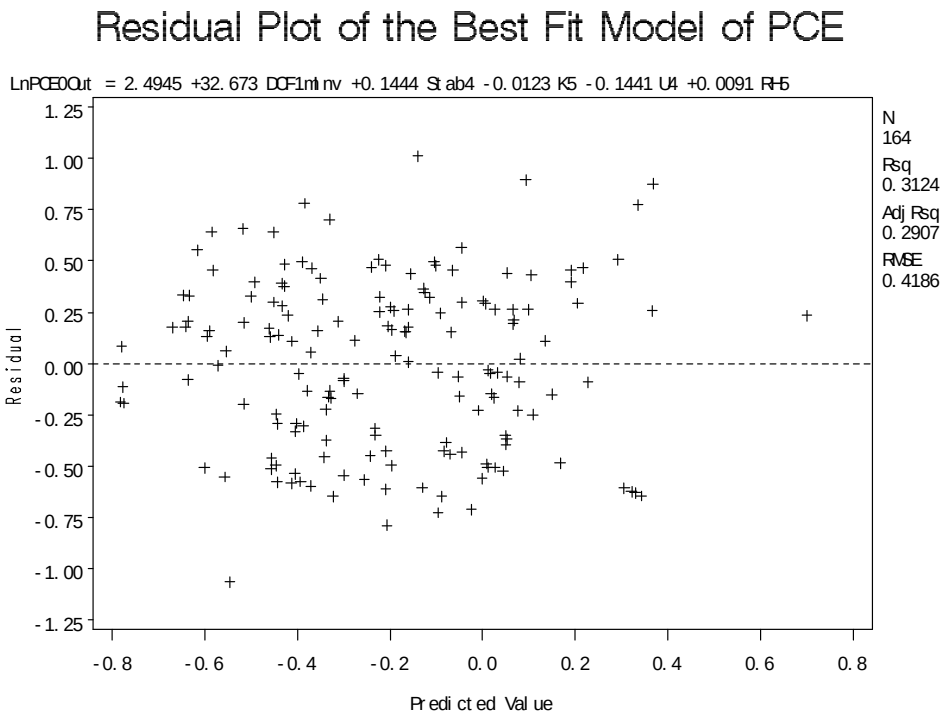

Figure A-27. Residual vs. Predicted Plot of the Best-fitting 6-Parameter Model of PCE after Removing the Outliers

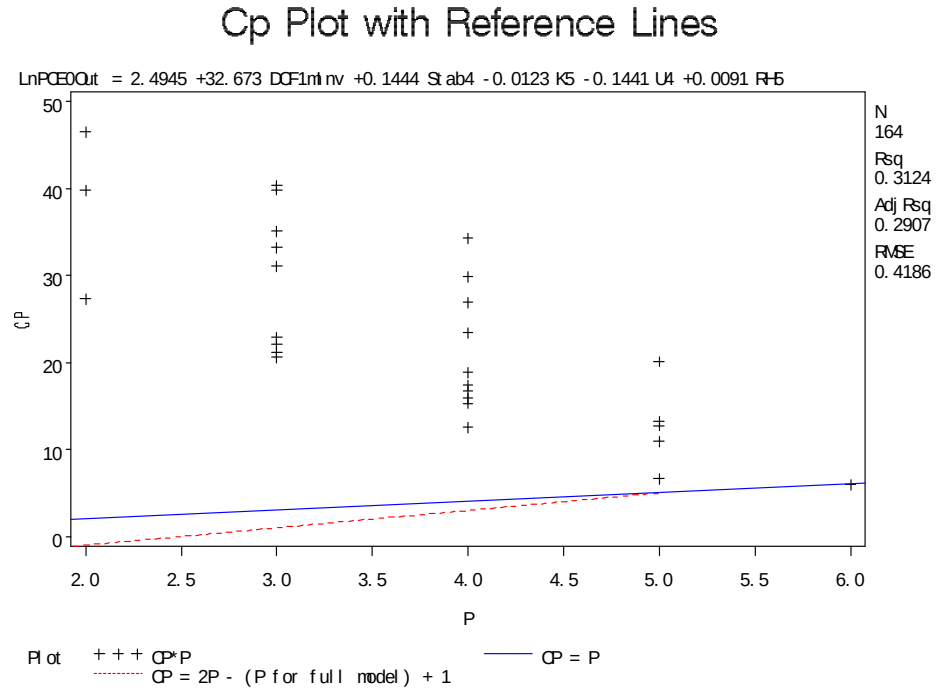

Figure A-28. Cp Plot for the Best-fitting 6-Parameter Model for PCE after Removing the Outliers

$PM_{2.5}$ Mass

Bivariate Pearson Correlation

The correlation coefficients of the ln-transformed $PM_{2.5}$ Mass and the distance to the inverse of urban interstate (FC11) roadways, minor arterial roads (F16) and local roads (F19) were 0.21 (p=0.03), 0.22 (p=0.03) and 0.23 (0.02), respectively.

The Pearson correlation coefficients of the meteorological variables and the ln-transformed $PM_{2.5}$ Mass that were statistically significantly correlated were: atmospheric stability, 0.56 (p<0.0001); mixing height, -0.26 (p=0.01); wind speed, -0.50 (p<.0001); and relative humidity, 0.39 (p<0.0001).

Preliminary Selection of Predictors

The preliminary regression analysis was performed on the ln-transformed $PM_{2.5}$ Mass to determine the relative importance of variables within the same types (proximity and meteorological) of independent variables. The distances to the urban interstates (FC11), and local roadways (FC19) were included in the linear regression model (p<0.15). Inverse distance to a truck loading/unloading area (PM03) was also selected. Among the meteorological variables, atmospheric stability, temperature, atmospheric pressure and wind speed were selected.

Selection of the Best-fitting Model

The variables selected by the different regressions methods were relatively consistent. Atmospheric stability was the most important factor in the regression model with partial r^2 of 0.318. The wind speed, temperature and atmospheric stability were also included as predictors in the model. The model also included the inverse distances to the

major roadways (FC11), local roadways (F19) and truck loading area (PM03). The parameters and analysis of variance of the regression equations for the $PM_{2.5}$ Mass ambient air concentration for the best-fitting model with 6 variables selected are given in Table A-15. The C(p), which is Mallows' C_p statistic, associated with this particular subset of variables was determined to be 8.0. The resulting model was appropriate in number of parameters, because the number of parameters (p) including the intercept in the best-fitting model match to the C(p) value. The diagnostic plots, the residual plot against the predicted values, and the normal probability-probability (PP) plot were generated and visually examined (Figures A-30 and Appendix B). The residuals were randomly distributed without showing any obvious trend or any particular pattern (Figure A-30.) and close to a normal distribution and the constant variances. The PP plot was nearly linear so it could be considered the error term of the model follows a normal distribution. Based on the visual diagnosis, there was no significant evidence of lack of fit or of significant unequal error variance for the best 8-parameter regression model.

No evidence of outliers was found.

Diagnostics of Equal Variances and Multicollinearity Diagnostics

The standardized residuals of the "best-fitting" model are close to a normal distribution and have constant variances. A Shapiro-Wilk W test for normality was also performed, and the p-value was very large (0.61), indicating that we cannot reject that the residuals are normally distributed. Based on these results, there was no evidence of a lack of fit or unequal error variance for this 8-parameter regression model.

The multicollinearity of the eight predictor variables was tested by checking their variance inflation factors (vif), which varied between 0 and 1.35 and were never higher

than 10 (reference value). This indicates that there was no significant collinearity between the predictors in the model. However, the collinearity diagnostic tests showed a condition index of 651 (much higher than 30, the reference value) and an eigenvalue of 0.00002 (much smaller than 0.01, the reference value). The proportion of variation for the intercept (0.99) and for the pressure (0.98) were greater than 0.5 (reference value), indicating that the two parameters were probably interacting, and that the 8-parameters model could be overly specified. However, to a certain extent, this might be unavoidable because it is extremely unlikely for all parameters to be completely independent (non correlated) to each other.

In order to lessen the degree of multicollinearity diagnosed, the atmospheric pressure, which showed some interaction with the intercept, was removed from the predictor variables and a new multiple regression analysis was run. The coefficient of determination (r^2) of the resulting "best-fitting"-7-parameter model was decreased from 0.50 to 0.49, and the condition index was decreased from 651 to 127.3. The interaction between the predictors in this new 7-parameter model appeared to be decreased after the removal of the pressure from the model, but the eigenvalue was still smaller than 0.01 (0.0003), and the proportion of variation of the intercept (0.98) and the temperature (0.94) were still greater than 0.5. Therefore, the temperature was removed from the 7-parameter model and a new multiple regression analysis was run again. This time, the coefficient of determination (r^2) for the resulting 6-parameter model was decreased from 0.49 to 0.47, the condition index decreased from 127.3 to 43.2 (much closer to 30, the reference value), and the eigenvalue increased to 0.002 (much closer to 0.01, the reference value). However, the proportion of variations of the intercept (0.949) and that

of the atmospheric stability (0.839) were still greater than 0.5. Eliminating the stability parameter from the model resulted in a drastic decrease of the coefficient of determination (r^2); thus, we concluded that probably the 6-parameters regression equation better describes the ln-transformed outdoor concentration of $PM_{2.5}$ (Table 16).

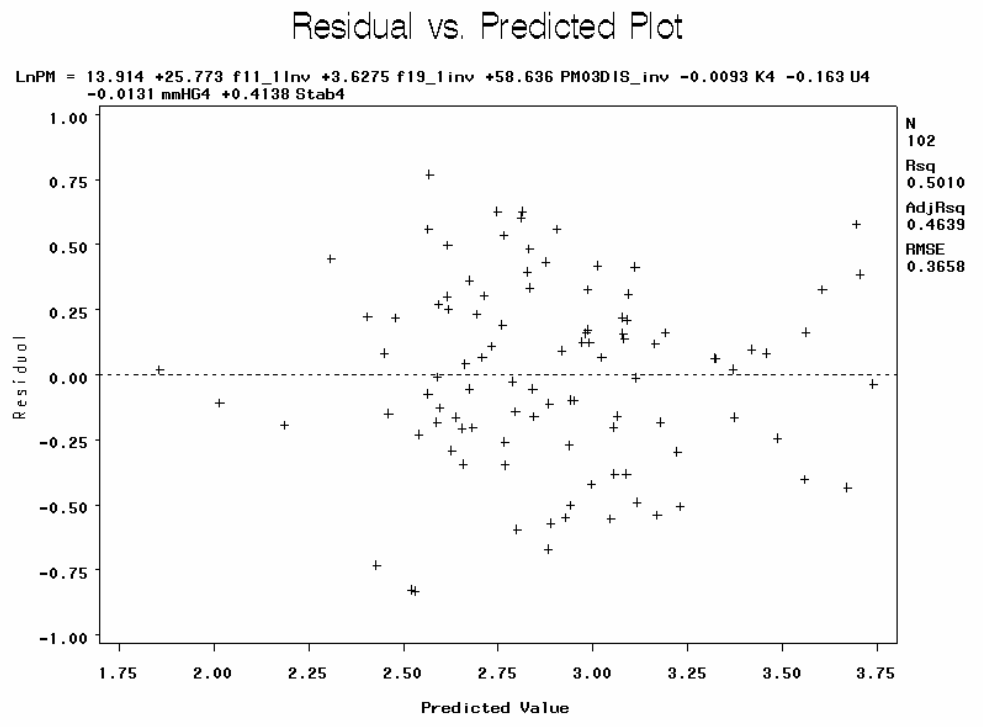

Figure A-30. Residual vs. Predicted Plot of the Best-fitting 6-Parameter Model of $PM_{2.5}$ Mass

Table.A-15. Results of the Best-fitting 7-Parameter Model for **LnPM2.5**

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	12.62973	1.80425	13.48	<.0001
Error	94	12.57688	0.13380		
Corrected Total	101	25.20661			

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	13.91429	6.84008	4.14	0.0447
F11_1Inv	Distance to FC11	1	25.77270	11.93020	4.67	0.0333
F19_1Inv	Distance to FC19	1	3.62748	1.65616	4.80	0.0310
PM03DIS_Inv		1	58.63622	29.32827	4.00	0.0485
K4	Temperature	1	-0.00927	0.00476	3.79	0.0545
U4		1	-0.16301	0.03855	17.88	<.0001
mmHG4		1	-0.01311	0.00827	2.52	0.1160
Stab4	Atmospheric Stability	1	0.41383	0.09362	19.54	<.0001

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Stab4	0.3184	0.3184	30.4114	46.71	<.0001
2	U4	0.0655	0.3839	20.0725	10.52	0.0016
3	F19_1Inv	0.0524	0.4363	12.1924	9.12	0.0032
4	F11_1Inv	0.0210	0.4573	10.2394	3.75	0.0557
5	PM03DIS_Inv	0.0163	0.4736	9.1699	2.97	0.0880
6	K4	0.0141	0.4877	8.5164	2.61	0.1094
7	mmGH4	0.0134	0.5010	8.0000	2.52	0.1160

Table.A-16 Results of the Best-fitting 5-Parameter Model for **PM2.5** after Removing Atmospheric Pressure and Temperature

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	11.93801	2.38760	17.27	<.0001
Error	96	13.26860	0.13821		
Corrected Total	101	25.20661			

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	1.06000	0.57297	3.42	0.0674
F11_1Inv	Distance to FC11	1	25.27425	12.03839	4.41	0.0384
F19_1Inv	Distance to FC19	1	4.19851	1.65726	6.42	0.0129
PM03DIS_Inv		1	50.94389	29.55414	2.97	0.0880
U4		1	-0.13037	0.03636	12.86	0.0005
Stab4		1	0.42820	0.09491	20.35	<.0001

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Stab4	0.3184	0.3184	26.3067	46.71	<.0001
2	U4	0.0655	0.3839	16.3623	10.52	0.0016
3	F19_1Inv	0.0524	0.4363	8.7980	9.12	0.0032
4	F11_1Inv	0.0210	0.4573	6.9713	3.75	0.0557
5	PM03DIS_Inv	0.0163	0.4736	6.0000	2.97	0.0880

Elemental Carbon

Bivariate Pearson Correlation

The correlation coefficients of the ln-transformed elemental carbon concentration and the distance to the inverse of urban interstate (FC11) roadways and minor arterial roads (F16) were 0.28 ($p=0.04$) and 0.29 ($p=0.03$), respectively. Distance to hamburger restaurants where broiling of meats occur had a correlation coefficient of 0.35($p=0.01$).

The Pearson correlation coefficients of the meteorological variables and the ln-transformed elemental carbon concentration that were statistically significantly correlated were: atmospheric stability, 0.43 ($p<0.0001$); mixing height, -0.34 ($p=0.01$); wind speed, -0.33 ($p=.02$), relative humidity, 0.49 ($p<0.0001$) and precipitation 0.29($p=0.03$).

Preliminary Selection of Predictors

The preliminary regression analysis was performed on the ln-transformed elemental carbon concentration to determine the relative importance of variables within the same types (proximity and meteorological) of independent variables. The distances to the urban major arterial roadways (FC14) was included in the resulting linear regression model ($p<0.15$). Inverse distance to a truck loading/unloading area seaport area (PM02) was also selected. Among the meteorological variables, atmospheric stability, relative humidity were selected.

Selection of the Best-fitting Model

The variables selected by the different regressions methods were relatively consistent. Atmospheric stability and relative humidity were included as predictors in the model. The model also included the inverse distances to the major roadways (FC14) and truck loading/sea port area (PM02). The parameters and analysis of variance of the

regression equations for elemental carbon ambient air concentration for the best-fitting model with 6 variables selected are given in Table A-17. The C(p), which is Mallows' C_p statistic, associated with this particular subset of variables was determined to be 5. The resulting model was appropriate in number of parameters, because the number of parameters (*p*) including the intercept in the best-fitting model match to the C(p) value. The diagnostic plots, the residual plot against the predicted values, and the normal probability-probability (PP) plot were generated and visually examined (Figures A- and Appendix B). The residuals were randomly distributed without showing any obvious trend or any particular pattern (Figure A-31.) and close to a normal distribution and the constant variances. The PP plot was nearly linear so it could be considered the error term of the model follows a normal distribution. Based on the visual diagnosis, there was no significant evidence of lack of fit or of significant unequal error variance for the best 8-parameter regression model.

No evidence of outliers was found.

Diagnostics of Equal Variances and Multicollinearity Diagnostics

The standardized residuals of the "best-fitting" model are close to a normal distribution and have constant variances. A Shapiro-Wilk W test for normality showed a very large p-value (0.65), indicating that the residuals are normally distributed.

From the figure above (PP-plot) we see that the distribution of the residuals doesn't seem heteroscedastic and, therefore, we accept the hypothesis of homogeneity of variance of the residuals in the 5-parameter model. Also in this case we checked the multicollinearity of the five predictor variables, which varied between 0 and 1.29, indicating that there was no significant collinearity between the predictors in the model.

The collinearity diagnostic tests showed a condition index of 37 (a little higher than 30) and an eigenvalue of 0.003 (smaller than 0.01). The proportion of variation for the intercept (0.92) and for the atmospheric stability (0.95) were greater than 0.5. However, because of the very low vif values we concluded that there was no significant collinearity between the predictors in the model and the 5-parameters regression equation shown before is basically adequate to describe the variation in outdoor concentration of LnEC.

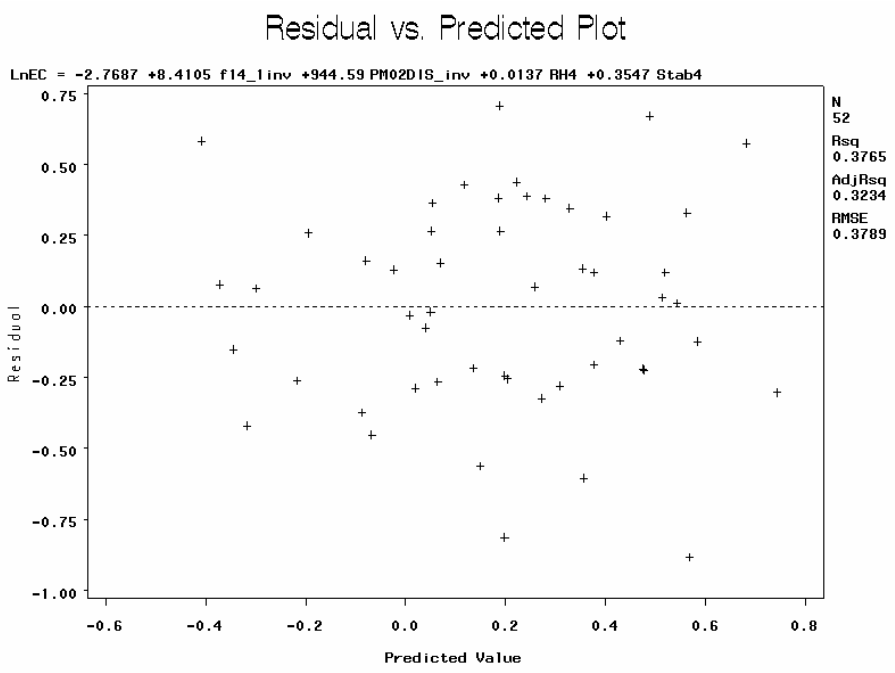

Figure A-31. Residual vs. Predicted Plot of the Best-fitting 6-Parameter Model of Elemental Carbon

Table.A-17. Results of the Best-fitting 4-Parameter Model for **Elemental Carbon**

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4.07456	1.01864	7.09	0.0001
Error	47	6.74801	0.14357		
Corrected Total	51	10.82257			

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	-2.76874	0.68589	16.30	0.0002
F14_1Inv	Distance to FC14	1	8.41047	3.84635	4.78	0.0338
PM02DIS_Inv		1	944.59427	494.54016	3.65	0.0622
RH4		1	0.01371	0.00498	7.60	0.0083
Stab4	Atmospheric Stability	1	0.35474	0.14301	6.15	0.0168

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	RH4	0.2357	0.2357	9.6140	15.42	0.0003
2	Stab4	0.0562	0.2919	7.3748	3.89	0.0542
3	F14_1Inv	0.0362	0.3281	6.6483	2.58	0.1145
4	PM02DIS_Inv	0.0484	0.3765	5.0000	3.65	0.0622

Organic Carbon

Bivariate Pearson Correlation

The correlation coefficients of the ln-transformed organic carbon concentration and the distance to the inverse of minor urban arterials (FC16) roadways was 0.29 ($p=0.04$). The Pearson correlation coefficients of the meteorological variables and the ln-transformed elemental carbon concentration that were statistically significantly correlated at $\alpha=0.05$, were atmospheric stability, 0.51 ($p<0.0001$) and relative humidity, 0.49 ($p<0.0001$).

Preliminary Selection of Predictors

The preliminary regression analysis was performed on the ln-transformed organic carbon concentration to determine the relative importance of variables within the same types (proximity and meteorological) of independent variables. The distances to the interstate roadways (FC11) was included in the resulting linear regression model ($p<0.15$). Atmospheric stability was included from the meteorological variables.

Selection of the Best-fitting Model

The variables selected by the different regressions methods were relatively consistent. Atmospheric stability was included as a predictor in the model. The model included the inverse distances to the interstate (FC11). The parameters and analysis of variance of the regression equations for elemental carbon ambient air concentration for the best-fitting model with 6 variables selected are given in Table A-18. The C(p), which is Mallows' C_p statistic, associated with this particular subset of variables was determined to be 3. The resulting model was appropriate in number of parameters, because the number of parameters (p) including the intercept in the best-fitting model match to the

C(p) value. The diagnostic plots, the residual plot against the predicted values, and the normal probability-probability (PP) plot were generated and visually examined (Figures A-32 and Appendix B). The residuals were randomly distributed without showing any obvious trend or any particular pattern (Figure A-32.) and close to a normal distribution and the constant variances. The PP plot was nearly linear so it could be considered the error term of the model follows a normal distribution. Based on the visual diagnosis, there was no significant evidence of lack of fit or of significant unequal error variance for the best 8-parameter regression model.

No evidence of outliers was found.

Diagnostics of Equal Variances and Multicollinearity Diagnostics

The standardized residuals of the "best-fitting" model are close to a normal distribution and have constant variances. A Shapiro-Wilk W test for normality showed a very large p-value (0.79), indicating that the residuals are normally distributed. From the figure above (PP plot) we see that also in this case we can accept the hypothesis of homogeneity of variance of the residuals.

The vif values of the three predictor variables varied between 0 and 1.006 (never higher than 10, the reference value). The collinearity diagnostic tests showed a condition index of 25.86 (lower than 30, the reference value) but an eigenvalue of 0.004 (a little smaller than 0.01, the reference value). Even though the proportion of variation for the intercept and for the atmospheric stability were greater than 0.5 (they both were 0.998) we still can assume that there was no significant collinearity between the predictors in the model because of the extremely low vif values and, therefore, the 3-parameters regression

equation shown before is basically adequate to describe the variation in outdoor concentration of LnOC.

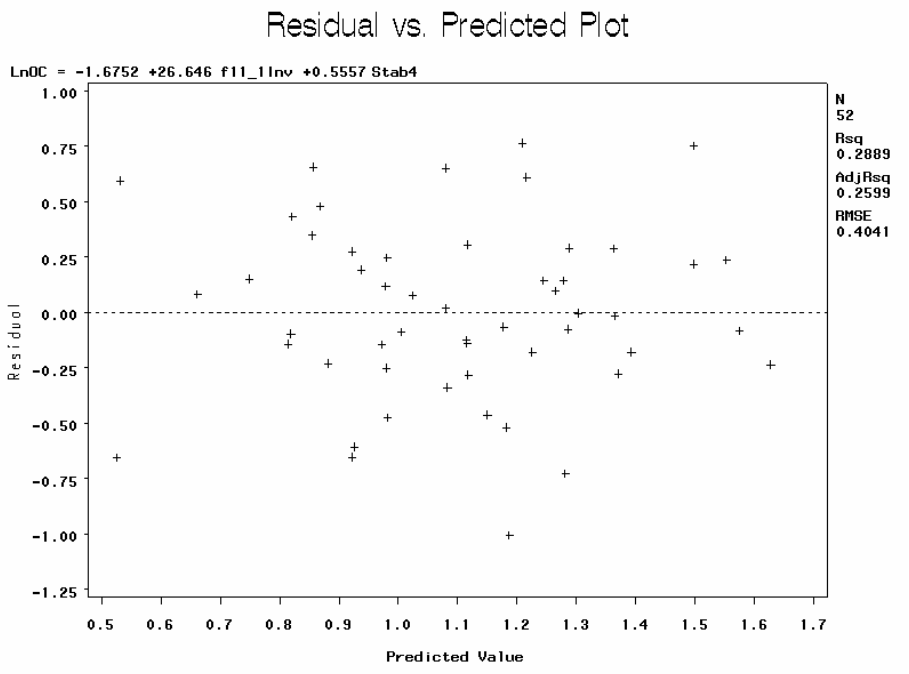

Figure A-32. Residual vs. Predicted Plot of the Best-fitting 6-Parameter Model of Organic Carbon

Table.A-18. Results of the Best-fitting 3-Parameter Model for **Organic Carbon**

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.25109	1.62555	9.95	0.0002
Error	49	8.00212	0.16331		
Corrected Total	51	11.25321			

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	-1.67517	0.66211	6.40	0.0147
F11_1Inv	Distance to FC11	1	26.64584	19.36632	1.89	0.1751
Stab4	Atmospheric Stability	1	0.55571	0.13474	17.01	0.0001

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	Stab4	0.2614	0.2614	2.8931	17.70	0.0001
4	F11_1Inv	0.0275	0.2889	3.000	1.89	0.1751

Coronene and Benzo-ghi-Pyrene

Bivariate Pearson Correlation

The correlation coefficients of the ln-transformed Coronene and Benzo-ghi-Pyrene concentrations and the distance to the inverse of urban collectors (FC17) roadways was 0.44 ($p=0.04$) for B-ghi-p and 0.42($P<.0001$) for COR. The Pearson correlation coefficients of the meteorological variables and the ln-transformed Coronene and Benzo-ghi-Pyrene concentrations that were statistically significantly correlated at $\alpha=0.05$, were atmospheric stability, 0.40 (B-ghi-P) and 0.44 (COR) ($p<0.0001$), temperature -0.42 (both) ($P<0.0001$) and mixing height -0.29 (B-ghi-P) and -0.31 (COR) ($p<0.04$).

Preliminary Selection of Predictors

The preliminary regression analysis was performed on the ln-transformed organic carbon concentration to determine the relative importance of variables within the same types (proximity and meteorological) of independent variables. The distances to the interstate roadways (FC11) and to Newark Airport (PM01) were included in the resulting linear regression model ($p<0.15$). Atmospheric stability, temperature and precipitation were included from the meteorological variables.

Selection of the Best-fitting Model

The variables selected by the different regressions methods were relatively consistent. Atmospheric stability was included as a predictor in the model. The model included the inverse distances to the interstate (FC11) and PM01. The parameters and analysis of variance of the regression equations for Coronene and Benzo-ghi-Pyrene ambient air concentration for the best-fitting model with 5 variables selected are given in

Table A-19 and 20. The C(p), which is Mallows' C_p statistic, associated with this particular subset of variables was determined to be 6. The resulting model was appropriate in number of parameters, because the number of parameters (p) including the intercept in the best-fitting model match to the C(p) value. The diagnostic plots, the residual plot against the predicted values, and the normal probability-probability (PP) plot were generated and visually examined (Figures A-33 and Appendix B). The residuals were randomly distributed without showing any obvious trend or any particular pattern (Figure A-33.) and close to a normal distribution and the constant variances. The PP plot was nearly linear so it could be considered the error term of the model follows a normal distribution. Based on the visual diagnosis, there was no significant evidence of lack of fit or of significant unequal error variance for the best 8-parameter regression model.

No evidence of outliers was found.

Diagnostics of Equal Variances and Multicollinearity Diagnostics

The standardized residuals of the "best-fitting" model are close to a normal distribution and have constant variances. A Shapiro-Wilk W test for normality indicated a very large p-values (0.72 and 0.75 for LnB-ghi-P and COR, respectively), suggesting that the residuals are normally distributed.

From the residual versus predicted values plots shown above we see that the distributions of the residuals doesn't seem overly heteroscedastic and, therefore, we can accept the hypothesis of homogeneity of variance of the residuals in both 6-parameter models.

The vif values varied between 0 and 1.27 for both PAHs, which indicates that there was no significant collinearity between the predictors in the two models. Once

again, the collinearity diagnostic tests showed condition indexes and eigenvalues that were, respectively, slightly higher and lower than the reference values (30 and 0.01), and proportion of variations for two of the predictor variables that were higher than 0.5 (the reference value). However, because of the very low vif values obtained we concluded that there was significant collinearity between the predictors in neither of the two models and that the two 6-parameters regression equations shown before are basically adequate to describe the variations in outdoor concentrations of LnB-ghi-P and LnCOR.

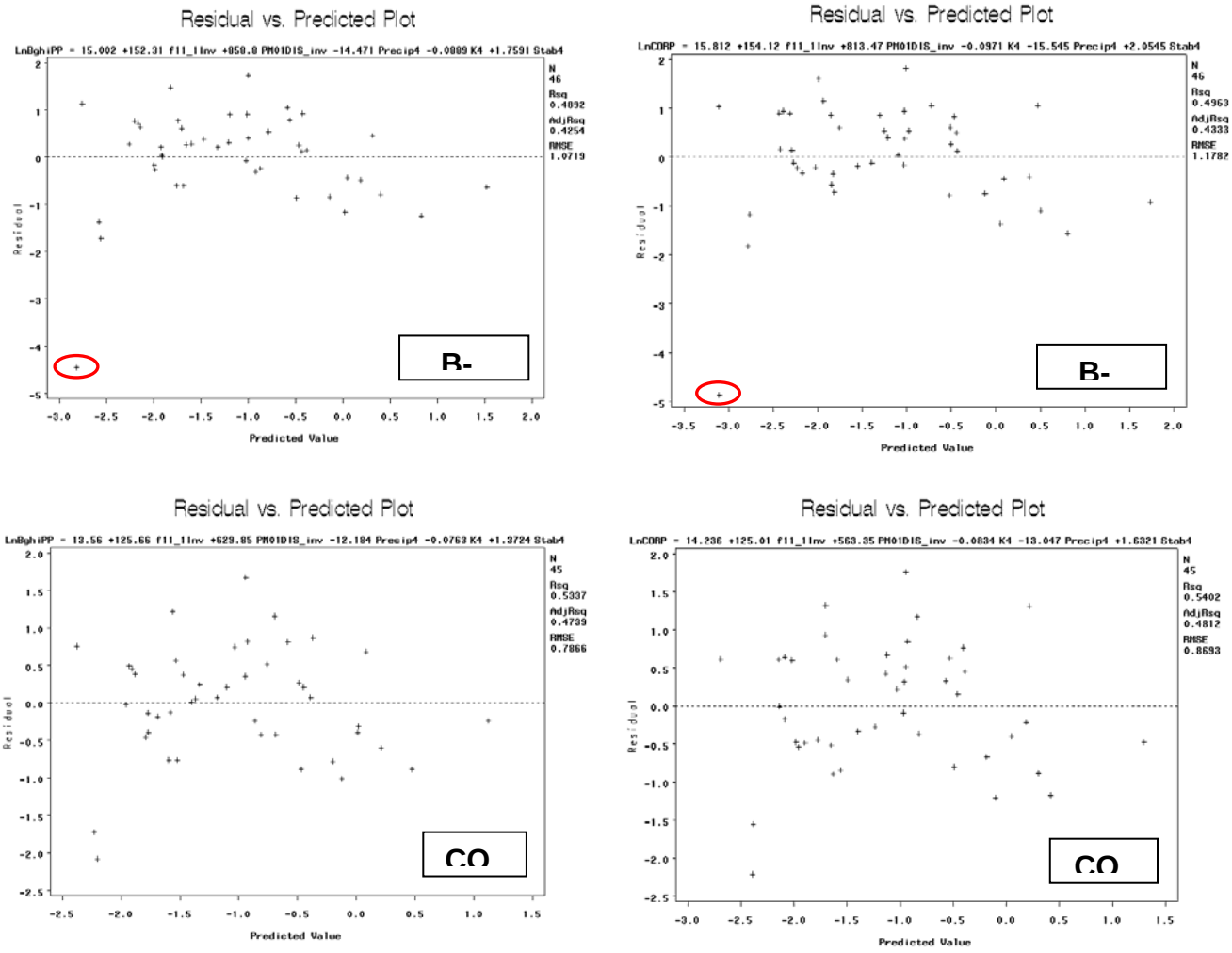

Figure A-33. Residual vs. Predicted Plot of the Best-fitting 6-Parameter Model of COR and B-hgi-P

Table.A-19. Results of the Best-fitting 5-Parameter Model for **Corenen**

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	14.23612	4.72659	9.07	0.0045
F11_1Inv	Distance to FC11	1	125.00731	53.04655	5.55	0.0236
PM01DIS_Inv		1	563.35265	371.56318	2.30	0.1375
Precip4		1	-13.04686	5.65791	5.32	0.0265
K4		1	-0.08337	0.01603	27.04	<.0001
Stab4		1	1.63208	0.35474	21.17	<.0001

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	K4	0.2140	0.2140	25.6691	11.71	0.0014
2	Stab4	0.1966	0.4106	10.9925	14.01	0.0005
3	F11_1Inv	0.0588	0.4694	8.0059	4.54	0.0391
4	Precip4	0.0437	0.5131	6.2988	3.59	0.0654
5	PM01DIS_Inv	0.0271	0.5402	6.0000	2.30	0.1375

Table.A-20. Results of the Best-fitting 5-Parameter Model for **Benzo-ghi-Pyrene**

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	Intercept	1	13.55955	4.27716	10.05	0.0030
F11_1Inv	Distance to FC11	1	125.65797	48.00264	6.85	0.0125
PM01DIS_Inv		1	629.84523	336.23325	3.51	0.0685
Predip4		1	-12.18382	5.11993	5.66	0.0223
K4	Temperature	1	-0.07630	0.01451	27.65	<.0001
Stab4	Atmospheric Stability	1	1.37241	0.32101	18.28	0.0001

Summary of Stepwise Selection

Step	Variable Entered	Partial R-Square	Model R-Square	Cp	F Value	Pr>F
1	K4	0.2160	0.2160	24.5663	11.85	0.0013
2	Stab4	0.1613	.03773	13.0761	10.88	0.0020
3	F11_1Inv	0.0719	0.4493	9.0590	5.36	0.0257
4	Precip4	0.0424	0.4917	7.5090	3.34	0.0751
5	PM01DIS_Inv	0.0420	0.5337	6.0000	3.51	0.0685

APPENDIX B. PP (Probability), QQ (Quantile) Plots

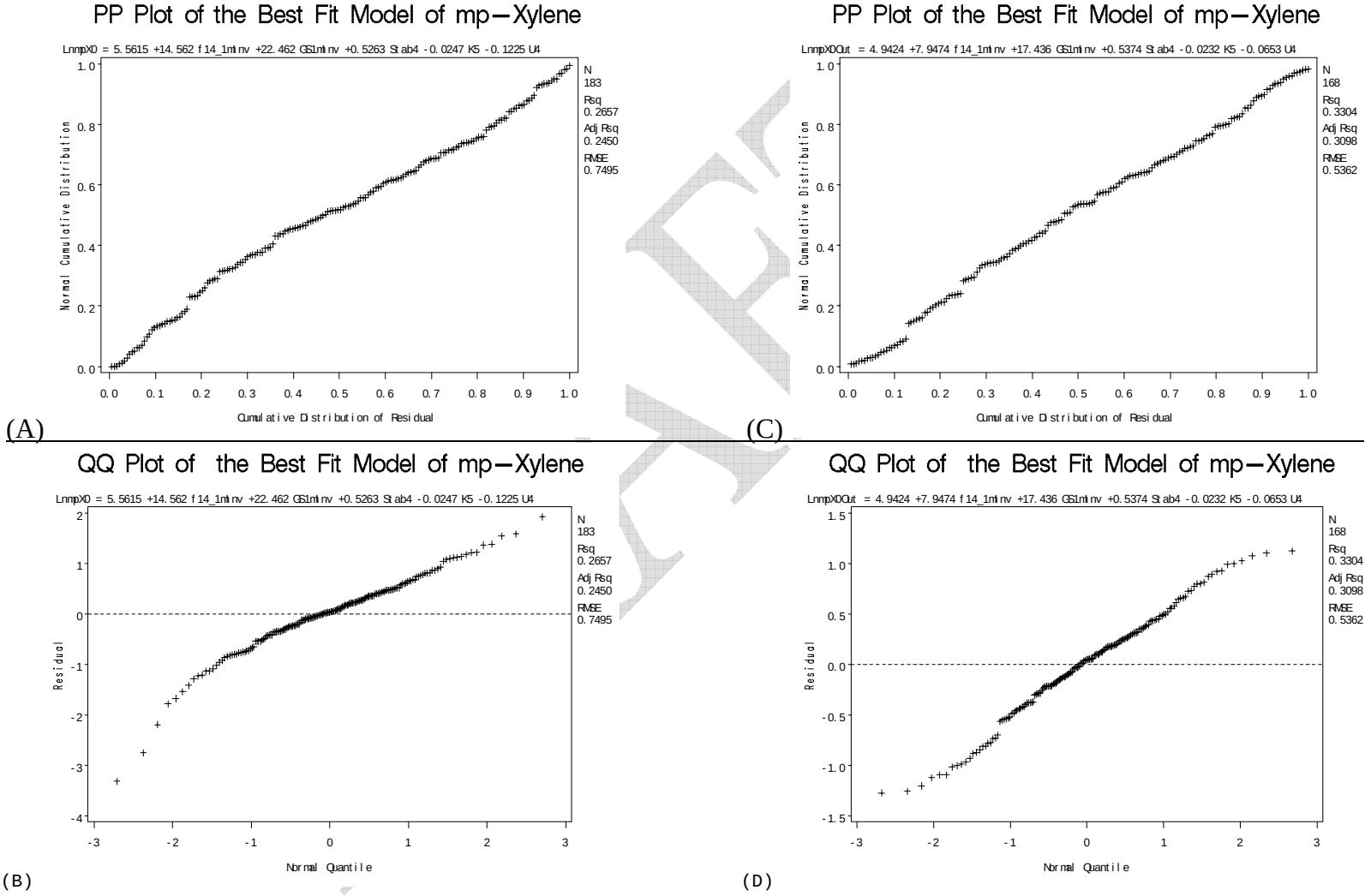

Figure B-1. Plots for Model of m,p-Xylene before (A,B) and after (C,D) removal of outliers

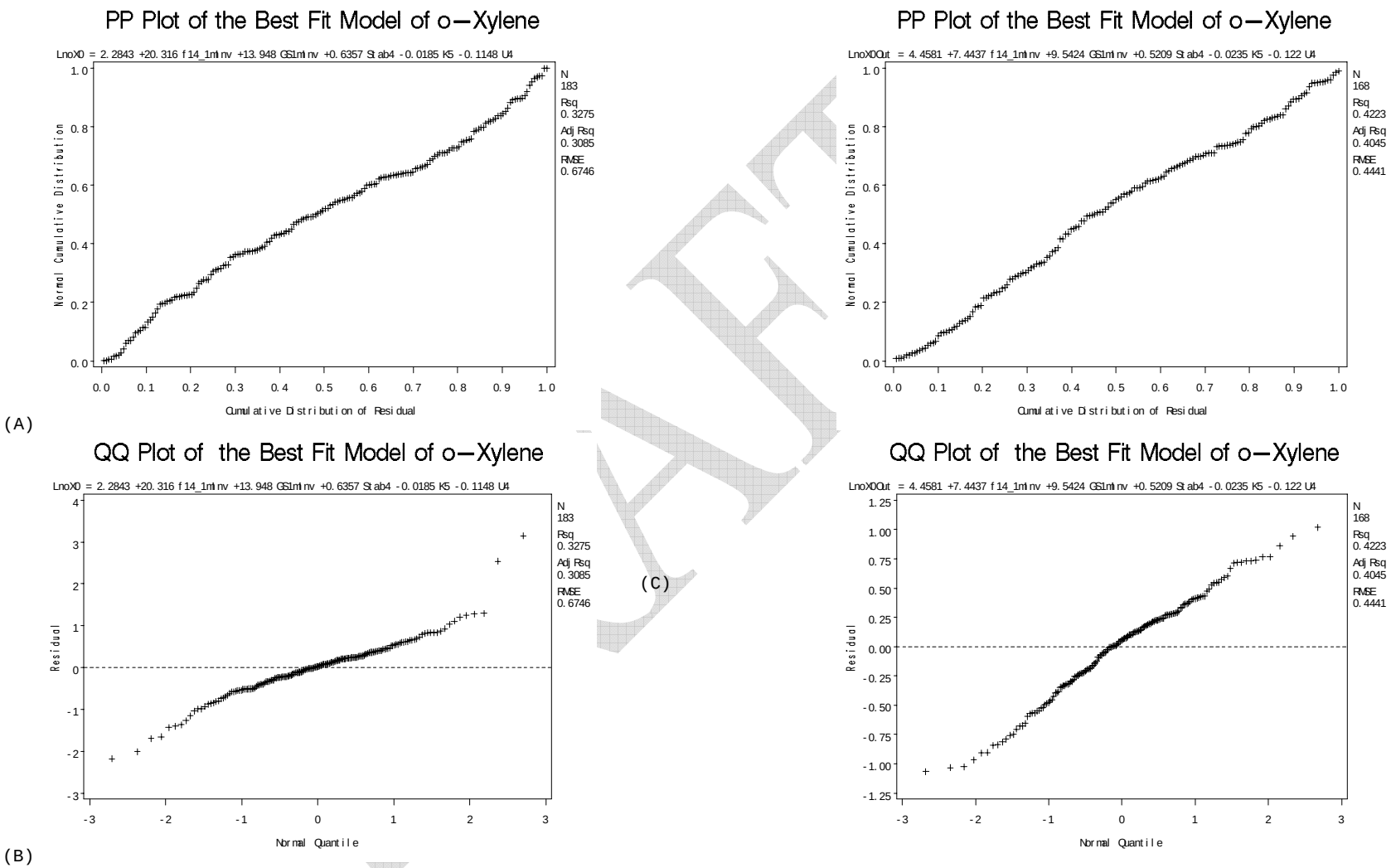

Figure B-2. Plots for Model of o-Xylene before (A,B) and after (C,D) removal of outliers

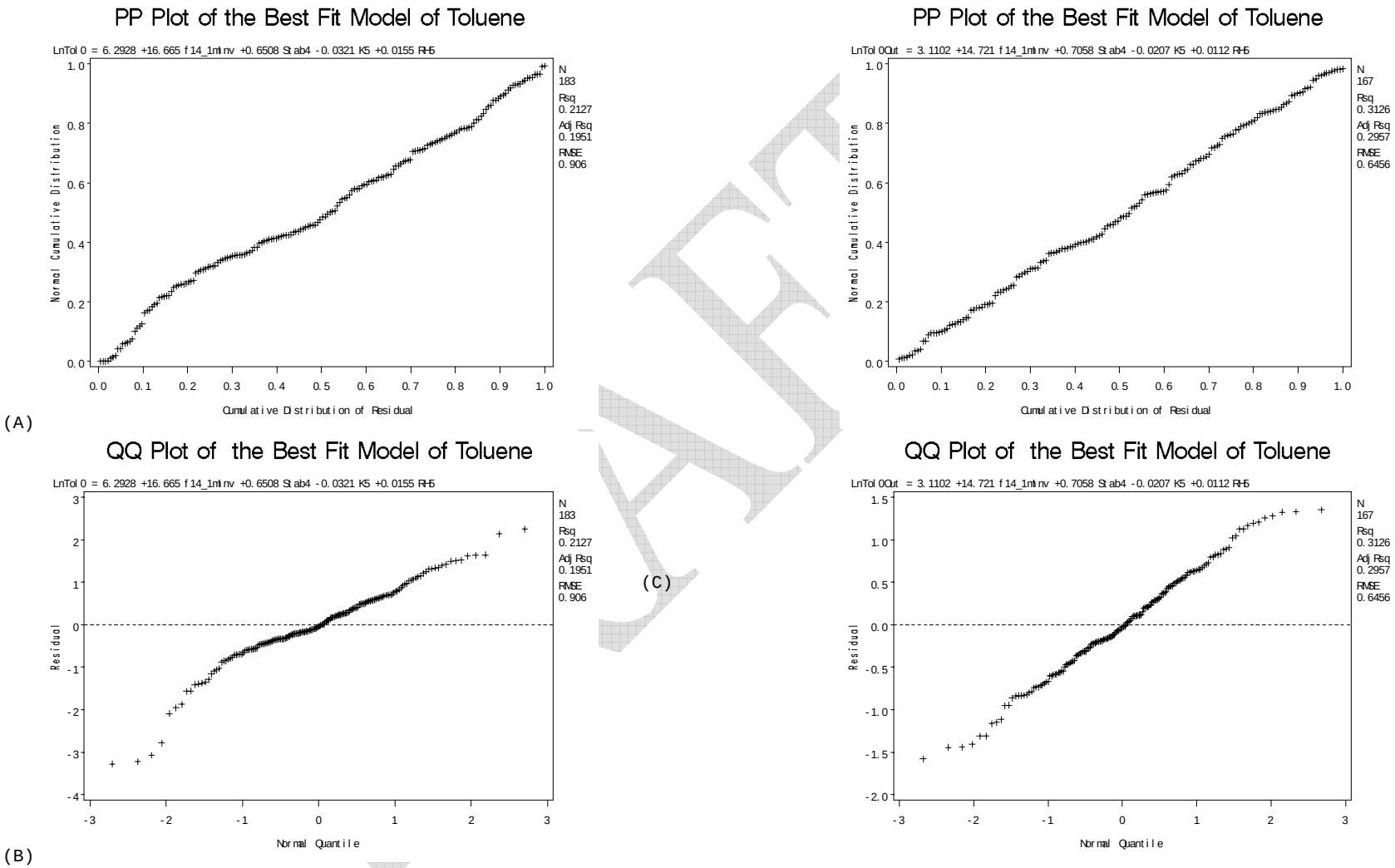

Figure B-3. Plots for Model of Toluene before (A,B) and after (C,D) removal of outliers

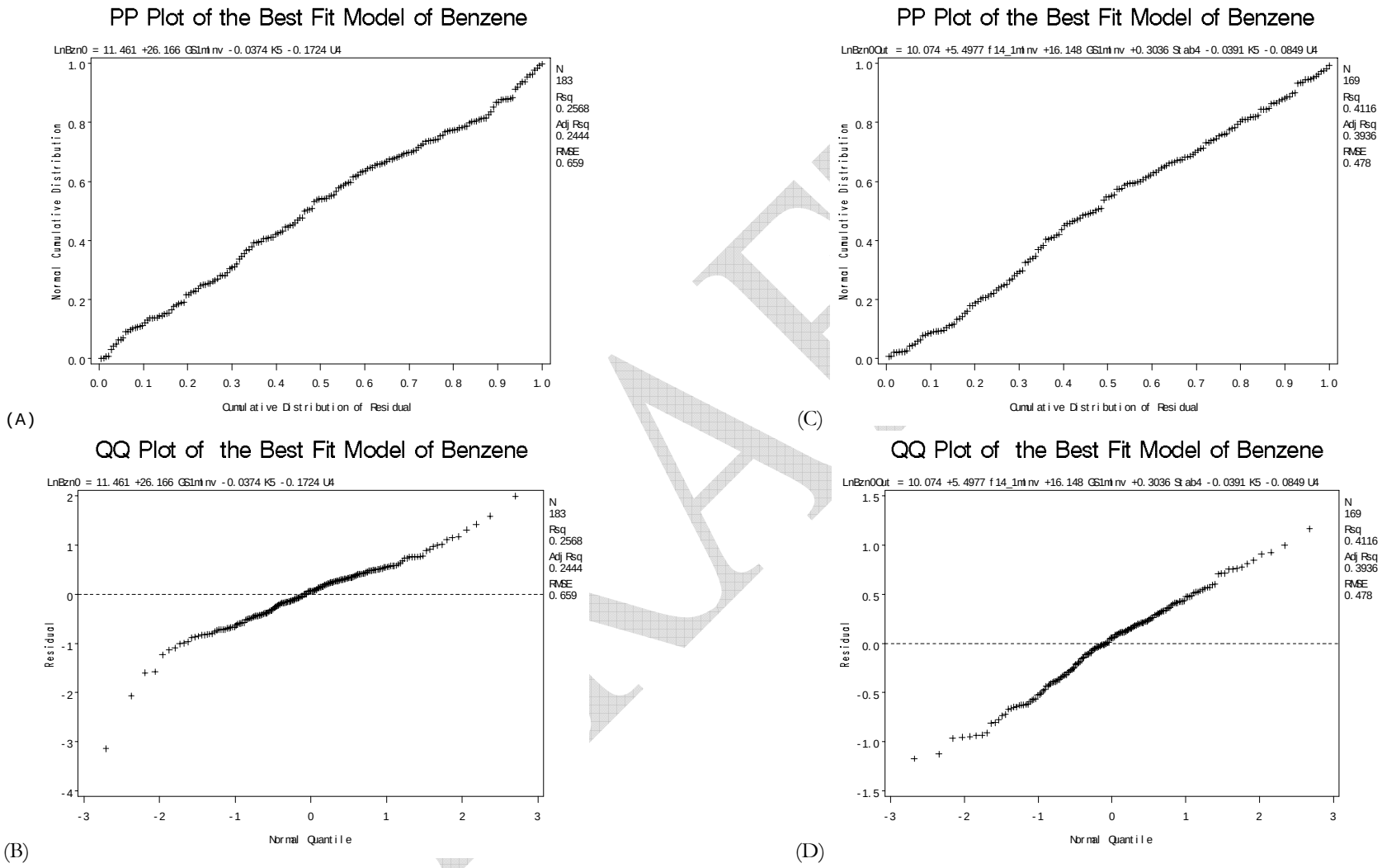

Figure B-4. Plots for Model of Benzene before (A,B) and after (C,D) removal of outliers

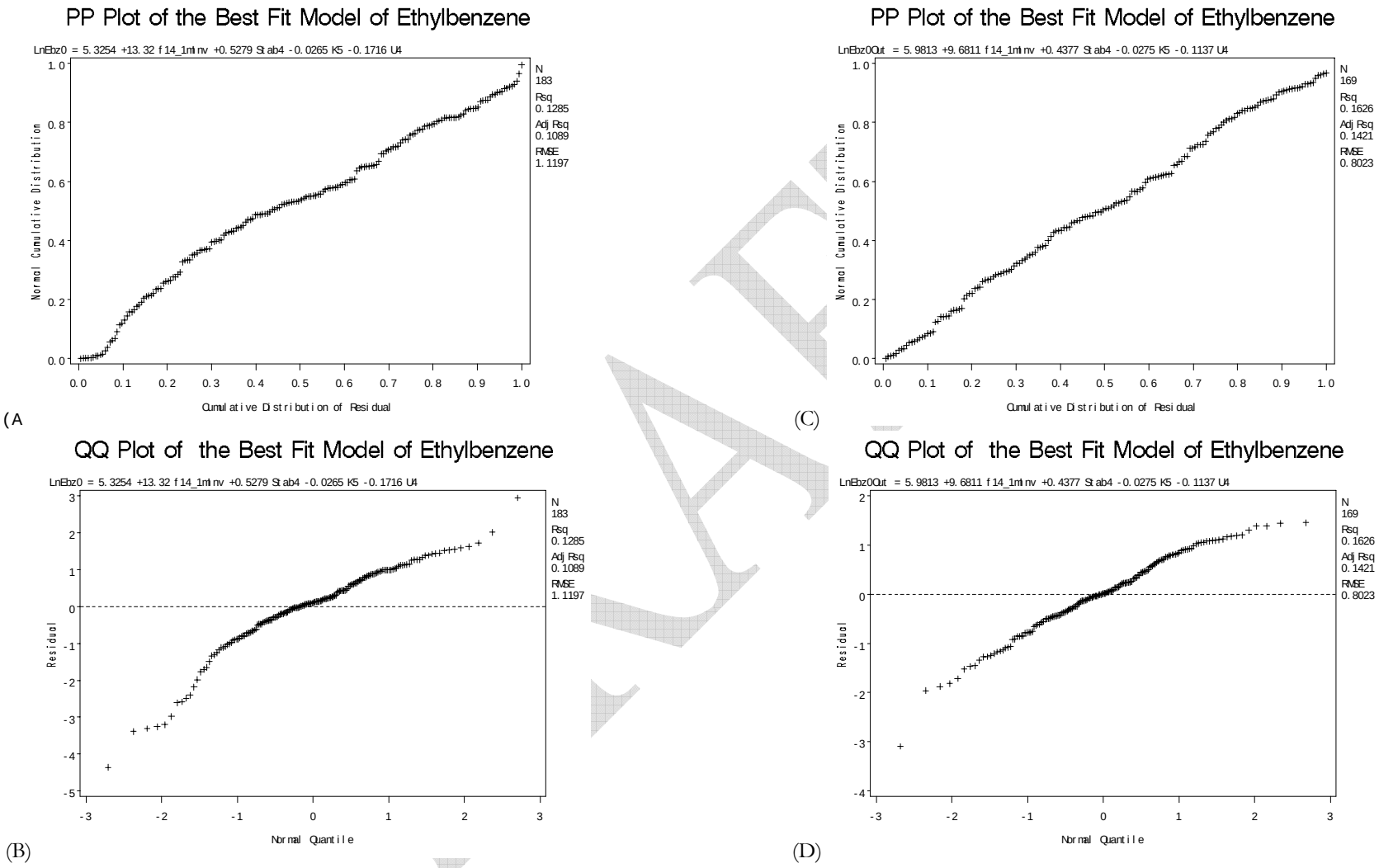

Figure B-5. Plots for Model of Ethylbenzene before (A,B) and after (C,D) removal of outliers

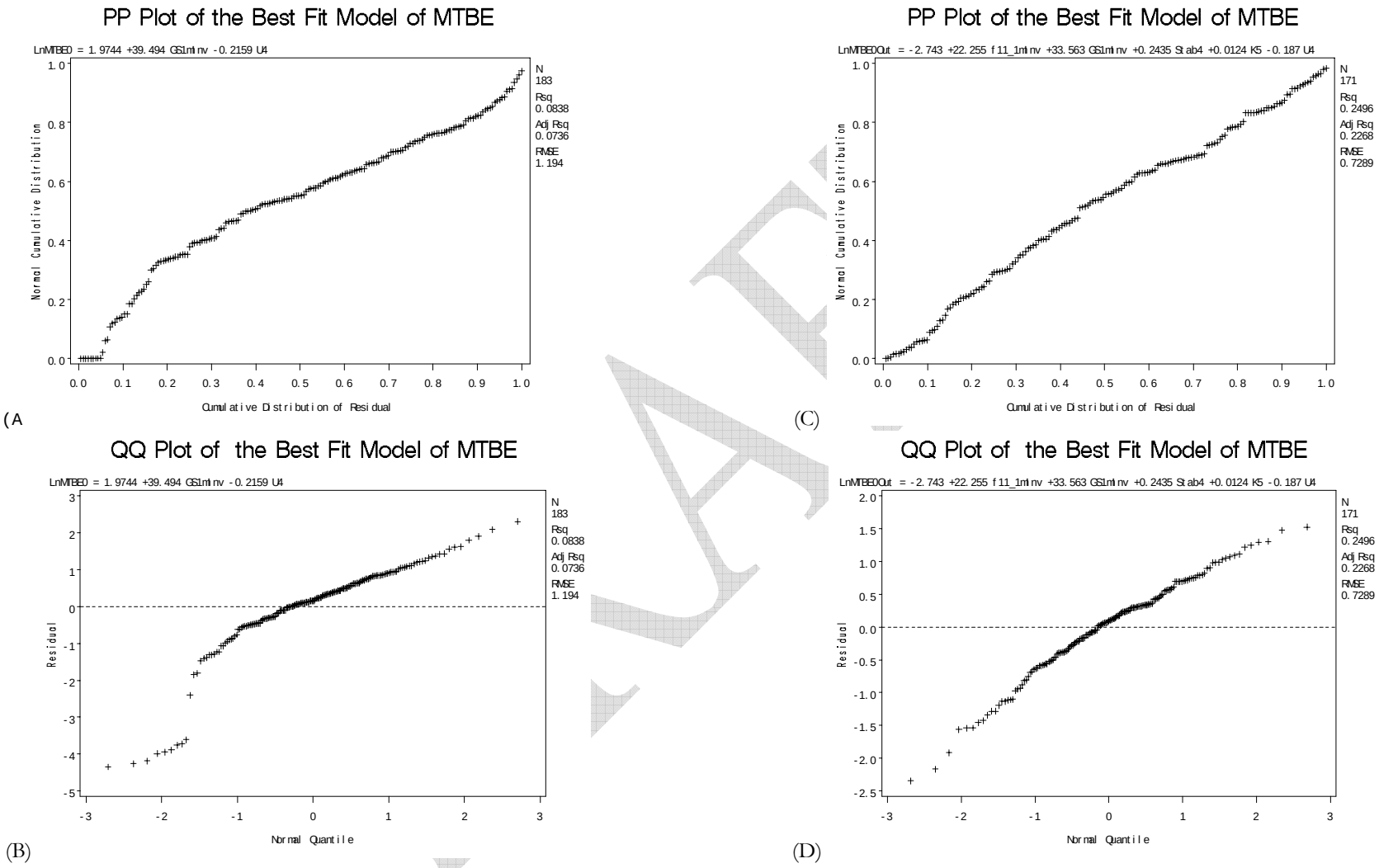

Figure B-6. Plots for Model of MTBE before (A,B) and after (C,D) removal of outliers

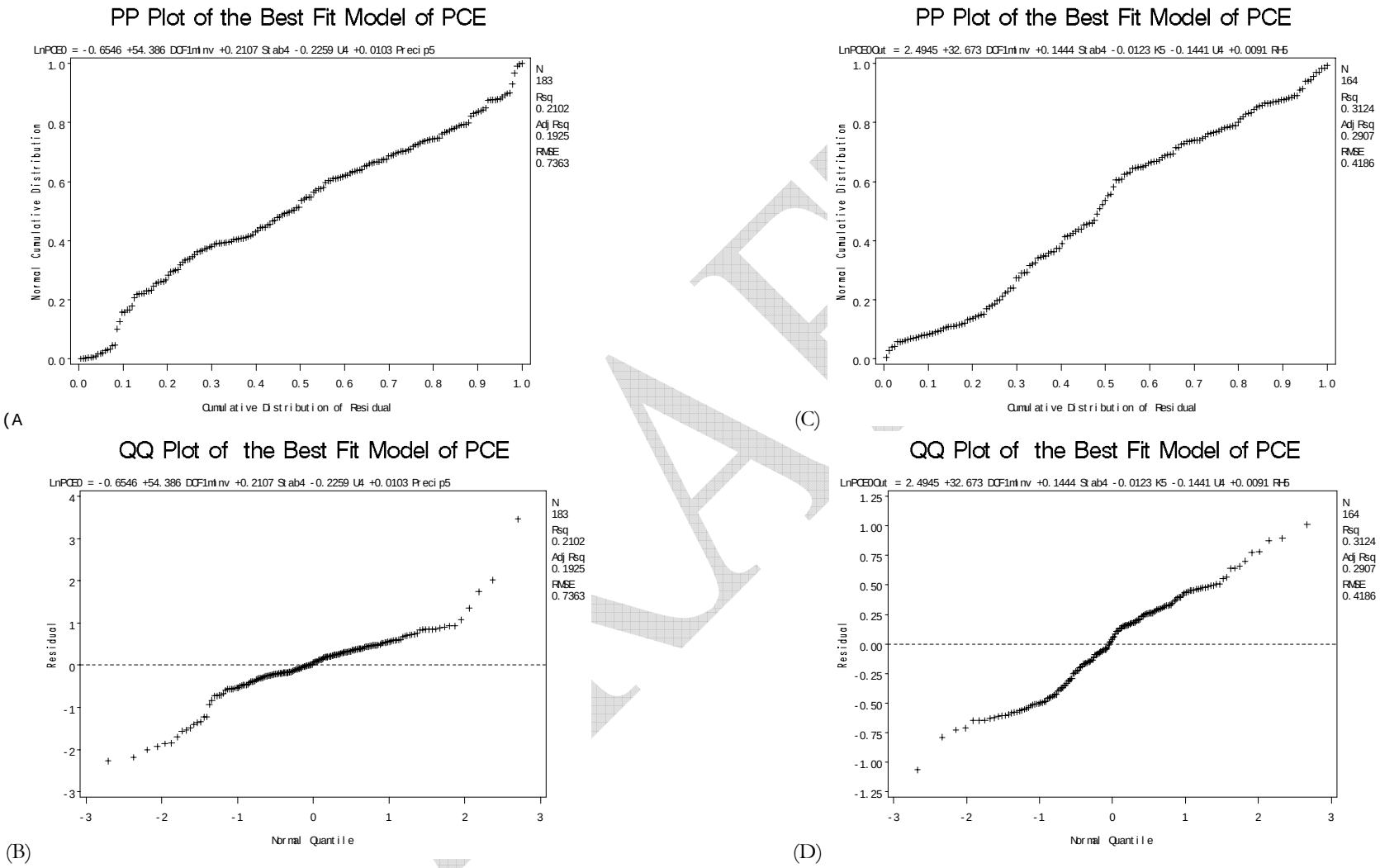

Figure B-7. Plots for Model of PCE before (A,B) and after (C,D) removal of outliers

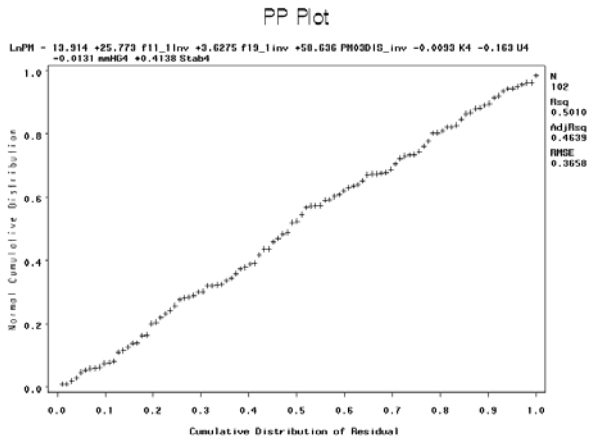

Figure B-7. Plots for Model of PM2.5 Mass

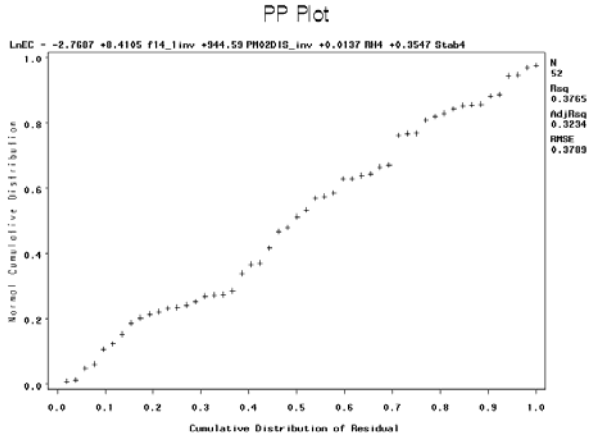

Figure B-7. Plots for Model of Elemental Carbon

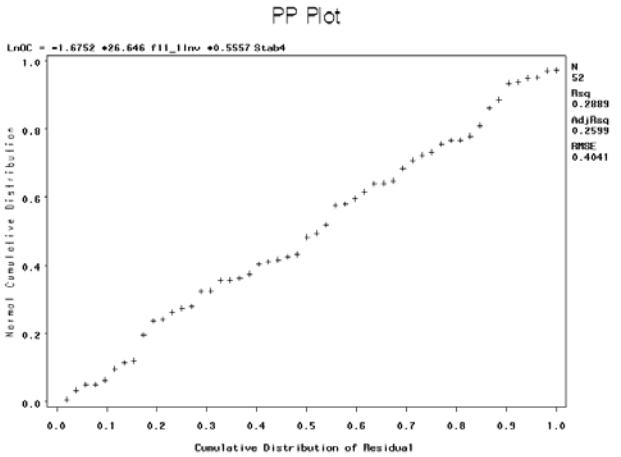

Figure B-7. Plots for Model of Organic Carbon

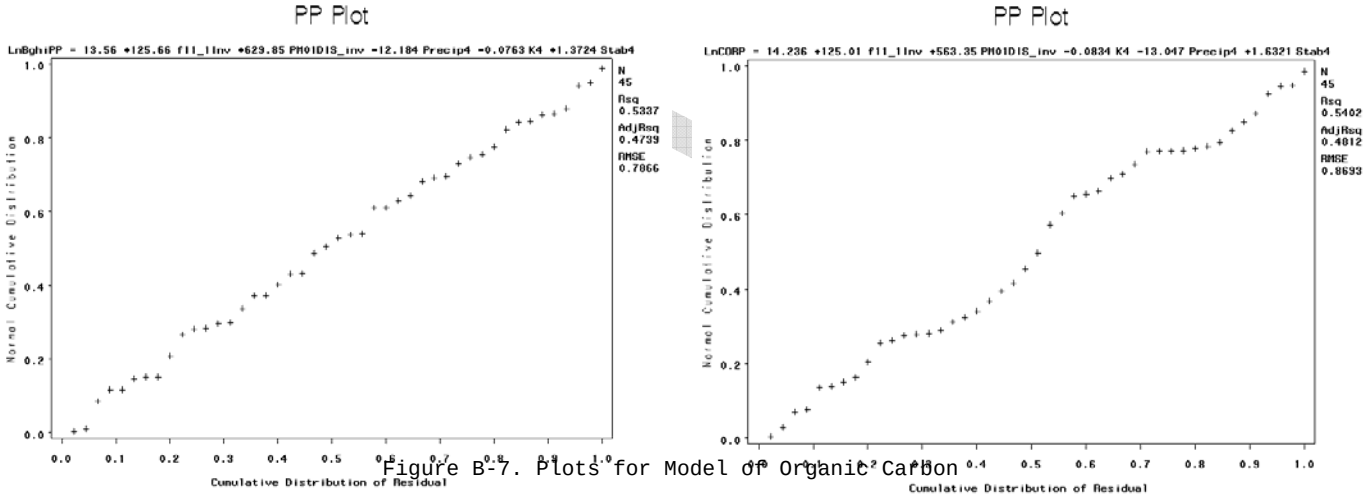

Figure B-7. Plots for Model of Organic Carbon

APPENDIX C. Diagnostic Results of Equal Variance and Multicollinearity

M,p-Xylene

Consistent Covariance of Estimates

Variable	Intercept	f14_1mInv	GS1mInv	Stab4	K5	U4
Intercept	2.4504	0.30152	-0.86165	-0.08012	-0.00668	-0.03337
f14_1mInv	0.30152	10.3825	-2.90185	0.04758	-0.00222	0.00483
GS1mInv	-0.86165	-2.90185	35.1448	-0.05396	0.00323	0.01211
Stab4	-0.08012	0.04758	-0.05396	0.00971	7.5E-05	0.00231
K5	-0.00668	-0.00222	0.00323	7.5E-05	2.1E-05	4.8E-05
U4	-0.03337	0.00483	0.01211	0.00231	4.8E-05	0.00197

Test of First and Second Moment Specification

DF	Chi-Square	Pr > ChiSq
20	23.57	0.2616

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	4.94236	1.70161	2.9	0.0042	.	0
F14_1mInv	1	7.94739	4.43103	1.79	0.0747	0.95826	1.04356
GS1mInv	1	17.4362	6.29951	2.77	0.0063	0.93148	1.07356
Stab4	1	0.53744	0.11065	4.86	<.0001	0.74522	1.34188
K5	1	-0.0232	0.00507	-4.58	<.0001	0.91088	1.09784
U4	1	-0.0653	0.04438	-1.47	0.1431	0.70158	1.42535

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation					
			Intercept	f14_1mInv	GS1mInv	Stab4	K5	U4
1	4.77579	1	2.5E-05	0.01274	0.0131	0.00023	3.4E-05	0.00179
2	0.64666	2.7176	4.1E-05	0.66273	0.14372	0.00036	5.6E-05	0.0034
3	0.51787	3.03677	7.3E-06	0.31247	0.81381	4.6E-05	1.1E-05	0.0003
4	0.05567	9.26198	0.00043	0.00407	0.00016	0.01681	0.00081	0.58238
5	0.00365	36.1517	0.0199	0.00799	0.02775	0.86081	0.06304	0.22528
6	0.00035	116.442	0.9796	1.3E-06	0.00147	0.12174	0.93606	0.18686

***o*-Xylene**

Consistent Covariance of Estimates

Variable	Intercept	f14_1mInv	GS1mInv	Stab4	K5	U4
Intercept	1.67737	0.71942	-1.41352	-0.05168	-0.00462	-0.01991
f14_1mInv	0.71942	13.7652	-3.15844	0.00684	-0.00276	-0.00947
GS1mInv	-1.41352	-3.15844	15.3048	0.00274	0.00464	-0.0015
Stab4	-0.05168	0.00684	0.00274	0.00738	2.7E-05	0.00141
K5	-0.00462	-0.00276	0.00464	2.7E-05	1.5E-05	2.1E-05
U4	-0.01991	-0.00947	-0.0015	0.00141	2.1E-05	0.0016

Test of First and Second Moment Specification

DF	Chi-Square	Pr > ChiSq
20	26.89	0.1384

Results of Multicollinearity Test on the Final Model of *o*-Xylene

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	4.45813	1.4074	3.17	0.0018	.	0
f14_1mInv	1	7.44373	4.48291	1.66	0.0988	0.93057	1.07461
GS1mInv	1	9.54244	5.47996	1.74	0.0835	0.91381	1.09431
Stab4	1	0.52092	0.09234	5.64	<.0001	0.72058	1.38777
K5	1	-0.02352	0.00419	-5.62	<.0001	0.9089	1.10023
U4	1	-0.12197	0.03697	-3.3	0.0012	0.68247	1.46527

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation					
			Intercept	f14_1mInv	GS1mInv	Stab4	K5	U4
1	4.83844	1	2.4E-05	0.01282	0.01258	0.00022	3.4E-05	0.0017
2	0.63267	2.76544	5E-05	0.36289	0.35101	0.00041	6.9E-05	0.00403
3	0.46846	3.21379	1.97E-10	0.61709	0.60919	1.9E-06	3.02E-08	1.3E-05
4	0.05647	9.25612	0.00042	2.5E-05	0.00486	0.01685	0.00079	0.5583
5	0.0036	36.6669	0.02017	0.00148	0.01935	0.85845	0.06507	0.24109
6	0.00035	116.918	0.97934	0.00569	0.00301	0.12406	0.93403	0.19487

Toluene

Consistent Covariance of Estimates

Variable	Intercept	f14_1mInv	Stab4	K5	RH5
Intercept	2.53139	0.99584	-0.03324	-0.00899	0.00266
f14_1mInv	0.99584	15.3306	-0.0417	-0.00354	0.00176
Stab4	-0.03324	-0.0417	0.00931	-1.1E-05	-0.00015
K5	-0.00899	-0.00354	-1.1E-05	3.5E-05	-1.2E-05
RH5	0.00266	0.00176	-0.00015	-1.2E-05	2E-05

Test of First and Second Moment Specification

DF	Chi-Square	Pr > ChiSq
14	26.13	0.0249

Results of Multicollinearity Test on the Final Model of Toluene

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	3.11017	1.84905	1.68	0.0945	.	0
f14_1mInv	1	14.7215	4.43634	3.32	0.0011	0.98854	1.01159
Stab4	1	0.70584	0.12046	5.86	<.0001	0.84859	1.17842
K5	1	-0.02067	0.00635	-3.25	0.0014	0.84105	1.18898
RH5	1	0.01116	0.0048	2.33	0.0212	0.73064	1.36867

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation				
			Intercept	f14_1mInv	Stab4	K5	RH5
1	4.31674	1	3.8E-05	0.01532	0.00035	4E-05	0.00122
2	0.65659	2.56408	2.5E-05	0.97236	0.00024	2.5E-05	0.00092
3	0.02114	14.291	0.00462	0.00271	0.01229	0.00315	0.81896
4	0.00516	28.9175	0.01539	0.0002	0.91675	0.02567	0.03874
5	0.00038	106.883	0.97993	0.00941	0.07038	0.97112	0.14016

Benzene

Consistent Covariance of Estimates

Variable	Intercept	f14_1mInv	GS1mInv	Stab4	K5	U4
Intercept	1.92094	0.77079	0.1021	-0.08161	-0.00501	-0.02127
f14_1mInv	0.77079	7.55125	-2.74611	-0.02564	-0.00231	-0.0074
GS1mInv	0.1021	-2.74611	28.7272	-0.07682	0.00096	-0.03097
Stab4	-0.08161	-0.02564	-0.07682	0.0103	7.3E-05	0.00224
K5	-0.00501	-0.00231	0.00096	7.3E-05	1.6E-05	1.3E-05
U4	-0.02127	-0.0074	-0.03097	0.00224	1.3E-05	0.00161

Test of First and Second Moment Specification

DF	Chi-Square	Pr > ChiSq
20	25.52	0.1824

Results of Multicollinearity Test on the Final Model of Benzene

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	10.0744	1.49805	6.73	<.0001	.	0
f14_1mInv	1	5.4977	3.32981	1.65	0.1007	0.95886	1.0429
GS1mInv	1	16.1478	5.57504	2.9	0.0043	0.94282	1.06065
Stab4	1	0.30356	0.09971	3.04	0.0027	0.7092	1.41003
K5	1	-0.03914	0.00447	-8.76	<.0001	0.9031	1.10729
U4	1	-0.08488	0.03966	-2.14	0.0338	0.66435	1.50522

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation					
			Intercept	f14_1mInv	GS1mInv	Stab4	K5	U4
1	4.74689	1	2.6E-05	0.01248	0.01339	0.00023	3.5E-05	0.00175
2	0.66641	2.66891	3.5E-05	0.7678	0.07717	0.00029	4.6E-05	0.00289
3	0.52523	3.00627	1.3E-05	0.20451	0.88566	8.8E-05	1.9E-05	0.00068
4	0.0576	9.0781	0.0004	0.00464	8.8E-06	0.01655	0.00077	0.54143
5	0.00351	36.755	0.02129	0.00542	0.02252	0.86748	0.0673	0.25551
6	0.00036	114.882	0.97823	0.00515	0.00125	0.11536	0.93183	0.19774

Ethylbenzene

Consistent Covariance of Estimates

Variable	Intercept	f14_1mInv	Stab4	K5	U4
Intercept	6.25969	0.16325	-0.23772	-0.01664	-0.07101
f14_1mInv	0.16325	31.8971	0.0485	-0.00188	-0.02487
Stab4	-0.23772	0.0485	0.02746	0.00027	0.00441
K5	-0.01664	-0.00188	0.00027	5.2E-05	8.2E-05
U4	-0.07101	-0.02487	0.00441	8.2E-05	0.00625

Test of First and Second Moment Specification

DF	Chi-Square	Pr > ChiSq
14	22.44	0.0700

Table E.10. Results of Multicollinearity Test on the Final Model of Ethylbenzene

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	5.98125	2.43806	2.45	0.0152	.	0
f14_1mInv	1	9.6811	5.62338	1.72	0.087	0.99172	1.00835
Stab4	1	0.43775	0.16287	2.69	0.0079	0.75403	1.32621
K5	1	-0.02747	0.00732	-3.75	0.0002	0.92634	1.07952
U4	1	-0.11372	0.06587	-1.73	0.0861	0.71056	1.40735

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation				
			Intercept	f14_1mInv	Stab4	K5	U4
1	4.29372	1	3.4E-05	0.01575	0.0003	4.6E-05	0.00227
2	0.64669	2.57673	2.4E-05	0.9744	0.0002	3E-05	0.0019
3	0.05557	8.79047	0.00046	0.00241	0.01763	0.00084	0.58435
4	0.00365	34.3084	0.0224	0.00158	0.87486	0.06731	0.23986
5	0.00038	106.199	0.97709	0.00585	0.10702	0.93178	0.17163

Methyl tert Butyl Ether (MTBE)

Consistent Covariance of Estimates

Variable	Intercept	f11_1mInv	GS1mInv	Stab4	K5	U4
Intercept	4.96409	0.29828	-1.29803	-0.13099	-0.01409	-0.06632
F11_1mInv	0.29828	210.102	-18.3631	0.08589	-0.00205	-0.09811
GS1mInv	-1.29803	-18.3631	61.7254	-0.34534	0.00883	0.03525
Stab4	-0.13099	0.08589	-0.34534	0.0222	3.8E-05	0.0028
K5	-0.01409	-0.00205	0.00883	3.8E-05	4.6E-05	0.00014
U4	-0.06632	-0.09811	0.03525	0.0028	0.00014	0.00261

Test of First and Second Moment Specification

DF	Chi-Square	Pr > ChiSq
20	18.20	0.5745

Results of Multicollinearity Test on the Final Model of MTBE

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	-2.743	2.27071	-1.21	0.2288	.	0
f11_1mInv	1	22.2547	14.3072	1.56	0.1217	0.97446	1.02621
GS1mInv	1	33.5627	8.28221	4.05	<.0001	0.95972	1.04197
Stab4	1	0.24348	0.14963	1.63	0.1056	0.70196	1.42459
K5	1	0.01239	0.00669	1.85	0.0659	0.89917	1.11213
U4	1	-0.18702	0.06022	-3.11	0.0022	0.6496	1.53942

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation					
			Intercept	f11_1mInv	GS1mInv	Stab4	K5	U4
1	4.66055	1	2.7E-05	0.01139	0.01397	0.00024	3.8E-05	0.00183
2	0.73182	2.52358	2E-05	0.94942	0.00242	0.00018	2.8E-05	0.00096
3	0.54484	2.92472	2.6E-05	0.0308	0.95844	0.00018	3.7E-05	0.00209
4	0.0588	8.90275	0.00038	0.00666	0.00263	0.01602	0.00077	0.53461
5	0.00364	35.8027	0.01965	0.00015	0.02173	0.85101	0.06764	0.24151
6	0.00036	113.396	0.9799	0.00157	0.00081	0.13237	0.93148	0.21899

Tetrachloroethylene (PCE)

Consistent Covariance of Estimates

Variable	Intercept	DCF1mInv	Stab4	K5	U4	RH5
Intercept	1.73111	-0.03794	-0.0554	-0.00463	-0.02507	-0.00017
DCF1mInv	-0.03794	200.107	-0.25514	0.0035	-0.01538	-0.00311
Stab4	-0.0554	-0.25514	0.00855	2.7E-05	0.00145	-8.6E-06
K5	-0.00463	0.0035	2.7E-05	1.5E-05	3.5E-05	-1.5E-06
U4	-0.02507	-0.01538	0.00145	3.5E-05	0.00135	3.1E-05
RH5	-0.00017	-0.00311	-8.6E-06	-1.49E-06	3.1E-05	7.86E-06

Test of First and Second Moment Specification

DF	Chi-Square	Pr > ChiSq
20	11.97	0.9170

Results of Multicollinearity Test on the Final Model of PCE

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	2.4945	1.34715	1.85	0.0659	.	0
DCF1mInv	1	32.6734	12.3964	2.64	0.0092	0.97021	1.0307
Stab4	1	0.14442	0.08831	1.64	0.104	0.74611	1.34028
K5	1	-0.01229	0.00416	-2.96	0.0036	0.85224	1.17338
U4	1	-0.1441	0.03588	-4.02	<.0001	0.71002	1.40842
RH5	1	0.00913	0.00301	3.04	0.0028	0.75523	1.3241

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation					
			Intercept	DCF1mInv	Stab4	K5	U4	RH5
1	5.53937	1	1.9E-05	0.0088	0.00017	2.4E-05	0.00132	0.0008
2	0.371	3.86407	3.4E-05	0.94315	0.00023	4.5E-05	0.00492	0.0016
3	0.06942	8.933	1.7E-05	0.01382	0.00338	6.8E-05	0.42245	0.0973
4	0.01614	18.5246	0.00354	0.02428	0.06688	0.00359	0.23735	0.84095
5	0.00374	38.4987	0.02173	0.00461	0.79268	0.05781	0.22567	0.00917
6	0.00034	127.603	0.97466	0.00533	0.13666	0.93846	0.10829	0.05018

4.10. PM2.5 Mass

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Stab4		Stab4	1	0.3184	0.3184	30.4114	46.71	<.0001
2	U4		U4	2	0.0655	0.3839	20.0725	10.52	0.0016
3	f19_1inv		f19_1inv	3	0.0524	0.4363	12.1924	9.12	0.0032
4	f11_1Inv		f11_1Inv	4	0.0210	0.4573	10.2394	3.75	0.0557
5	PM03DIS_inv		PM03DIS_inv	5	0.0163	0.4736	9.1699	2.97	0.0880
6	K4		K4	6	0.0141	0.4877	8.5164	2.61	0.1094
7	mmHG4		mmHG4	7	0.0134	0.5010	8.0000	2.52	0.1160

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	12.62973	1.80425	13.48	<.0001
Error	94	12.57688	0.13380		
Corrected Total	101	25.20661			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13.91429	6.84008	0.55366	4.14	0.0447
f11_1Inv	25.77270	11.93020	0.62441	4.67	0.0333
f19_1inv	3.62748	1.65616	0.64188	4.80	0.0310
PM03DIS_inv	58.63622	29.32827	0.53482	4.00	0.0485
	-0.00927	0.00476	0.50718	3.79	0.0545
K4	-0.16301	0.03855	2.39254	17.88	<.0001
U4mmHG4	-0.01311	0.00827	0.33669	2.52	0.1160
Stab4	0.41383	0.09362	2.61421	19.54	<.0001

Bounds on condition number: 1.5607, 58.934

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Stab4		Stab4	1	0.3184	0.3184	26.3067	46.71	<.0001
2	U4		U4	2	0.0655	0.3839	16.3623	10.52	0.0016
3	f19_1inv		f19_1inv	3	0.0524	0.4363	8.7980	9.12	0.0032

4	f11_1Inv	f11_1Inv	4	0.0210	0.4573	6.9713	3.75	0.0557
5	PM03DIS_inv	PM03DIS_inv	5	0.0163	0.4736	6.0000	2.97	0.0880

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	11.93801	2.38760	17.27	<.0001
Error	96	13.26860	0.13821		
Corrected Total	101	25.20661			

Find The Best Fitted Model for PM
Stepwise Selection: Step 5

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.06000	0.57297	0.47304	3.42	0.0674
f11_1Inv	25.27425	12.03839	0.60922	4.41	0.0384
f19_1inv	4.19851	1.65726	0.88708	6.42	0.0129
PM03DIS_inv	50.94389	29.55414	0.41068	2.97	0.0880
	-0.13037	0.03636	1.77726	12.86	0.0005
U4Stab4	0.42820	0.09491	2.81332	20.35	<.0001

Bounds on condition number: 1.3453, 28.976

C -11 Elemental Carbon

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	RH4		RH4	1	0.2357	0.2357	9.6140	15.42	0.0003
2	Stab4		Stab4	2	0.0562	0.2919	7.3748	3.89	0.0542
3	f14_1inv		f14_1inv	3	0.0362	0.3281	6.6483	2.58	0.1145
4	PM02DIS_inv		PM02DIS_inv	4	0.0484	0.3765	5.0000	3.65	0.0622

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4.07456	1.01864	7.09	0.0001
Error	47	6.74801	0.14357		
Corrected Total	51	10.82257			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-2.76874	0.68589	2.33955	16.30	0.0002
f14_1inv	8.41047	3.84635	0.68647	4.78	0.0338
PM02DIS_inv	944.59427	494.54016	0.52380	3.65	0.0622
RH4	0.01371	0.00498	1.09096	7.60	0.0083
Stab4	0.35474	0.14301	0.88334	6.15	0.0168

Bounds on condition number: 1.2892, 19.405

C-12 Organic Carbon

Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Stab4	Stab4	1	0.2614	0.2614	2.8931	17.70	0.0001
2	f11_1Inv	f11_1Inv	2	0.0275	0.2889	3.0000	1.89	0.1751

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.25109	1.62555	9.95	0.0002
Error	49	8.00212	0.16331		
Corrected Total	51	11.25321			

Find The Best Fitted Model for PM
Forward Selection: Step 2

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-1.67517	0.66211	1.04536	6.40	0.0147
f11_1Inv	26.64584	19.36632	0.30915	1.89	0.1751
Stab4	0.55571	0.13474	2.77773	17.01	0.0001

Bounds on condition number: 1.0061, 4.0244

FOR B-ghiP

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	K4		K4	1	0.2160	0.2160	24.5663	11.85	0.0013
2	Stab4		Stab4	2	0.1613	0.3773	13.0761	10.88	0.0020
3	f11_1Inv		f11_1Inv	3	0.0719	0.4493	9.0590	5.36	0.0257
4	Precip4		Precip4	4	0.0424	0.4917	7.5090	3.34	0.0751
5	PM01DIS_inv		PM01DIS_inv	5	0.0420	0.5337	6.0000	3.51	0.0685

The REG Procedure
Model: MODEL1
Dependent Variable: LnBghiPP LnBghiPP

Stepwise Selection: Step 5

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13.55955	4.27716	6.21906	10.05	0.0030
f11_1Inv	125.65797	48.00264	4.24028	6.85	0.0125
PM01DIS_inv	629.84523	336.23325	2.17136	3.51	0.0685
Precip4	-12.18382	5.11993	3.50416	5.66	0.0223
K4	-0.07630	0.01451	17.10874	27.65	<.0001
Stab4	1.37241	0.32101	11.31041	18.28	0.0001

Bounds on condition number: 1.3725, 30.547

FOR COR

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	K4		K4	1	0.2140	0.2140	25.6691	11.71	0.0014
2	Stab4		Stab4	2	0.1966	0.4106	10.9925	14.01	0.0005
3	f11_1Inv		f11_1Inv	3	0.0588	0.4694	8.0059	4.54	0.0391
4	Precip4		Precip4	4	0.0437	0.5131	6.2988	3.59	0.0654
5	PM01DIS_inv		PM01DIS_inv	5	0.0271	0.5402	6.0000	2.30	0.1375

The REG Procedure
Model: MODEL1
Dependent Variable: LnCORP LnCORP

Stepwise Selection: Step 5

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	14.23612	4.72659	6.85515	9.07	0.0045
f11_1Inv	125.00731	53.04655	4.19648	5.55	0.0236
PM01DIS_inv	563.35265	371.56318	1.73710	2.30	0.1375
Precip4	-13.04686	5.65791	4.01818	5.32	0.0265
	-0.08337	0.01603	20.43059	27.04	<.0001
K4*Stab4	1.63208	0.35474	15.99544	21.17	<.0001

Bounds on condition number: 1.3725, 30.547

Appendix C Correlation Matrix

Variables: f11_1 f12_1 f14_1 f16_1 f17_1 f19_1 GS1 DCF1 Tol_PS1

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
f11_1	183	1529	1052	279832	37.00000	3698	f11_1
f12_1	183	2529	1162	462872	24.00000	5578	f12_1
f14_1	183	499.56284	537.71156	91420	13.00000	2489	f14_1
f16_1	183	192.56284	166.63085	35239	5.00000	782.00000	f16_1
f17_1	183	288.22404	216.06297	52745	20.00000	967.00000	f17_1
f19_1	183	33.83060	20.68096	6191	2.00000	130.00000	f19_1
GS1	183	0.36053	0.20877	65.97700	0.02600	1.01200	GS1
DCF1	183	0.55103	0.38998	100.83800	0.05600	1.68700	DCF1
Tol_PS1	183	2.97733	1.12050	544.85226	0.83513	5.76225	Tol_PS1

Pearson Correlation between X Variables

	f12_1	f14_1	f16_1	f17_1	f19_1	GS1	DCF1
f12_1	0.22843 0.0019						
f14_1	-0.38184 <.0001	-0.46291 <.0001					
f16_1	0.16981 0.0216	-0.00242 0.9741	-0.26237 0.0003				
f17_1	0.14760 0.0462	0.14330 0.0530	0.10758 0.1472	0.07338 0.3235			
f19_1	0.09606 0.1958	0.21084 0.0042	-0.16341 0.0271	-0.04592 0.5370	-0.19577 0.0079		
GS1	0.12175 0.1006	0.05717 0.4420	0.18549 0.0119	-0.05098 0.4931	0.15279 0.0389	0.08628 0.2455	
DCF1	0.16410 0.0264	-0.09566 0.1977	0.38244 <.0001	0.02551 0.7318	0.50087 <.0001	-0.04480 0.5471	0.24865 0.0007
Tol_PS1	0.64536 <.0001	-0.50721 <.0001	0.22698 0.0020	0.12178 0.1006	0.21401 0.0036	-0.12433 0.0936	0.16325 0.0272	0.41727 <.0001

Prob > |r| under H0: Rho=0

The CORR Procedure

Simple Statistics

Pearson Correlation Coefficients, N = 183

Variables: f11_1mInv f12_1mInv f14_1mInv f16_1mInv f17_1mInv f19_1mInv GS1mInv DCF1mInv

9

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
f11_1mInv	183	0.00213	0.00412	0.38987	0.0002704	0.02703	f11_1mInv
f12_1mInv	183	0.00117	0.00467	0.21427	0.0001793	0.04167	f12_1mInv
f14_1mInv	183	0.00709	0.01098	1.29732	0.0004018	0.07692	f14_1mInv
f16_1mInv	183	0.01411	0.02342	2.58193	0.00128	0.20000	f16_1mInv
f17_1mInv	183	0.00824	0.00978	1.50805	0.00103	0.05000	f17_1mInv
f19_1mInv	183	0.04439	0.05320	8.12332	0.00769	0.50000	f19_1mInv
GS1mInv	183	0.00535	0.00670	0.97829	0.0009881	0.03846	GS1mInv
DCF1mInv	183	0.00307	0.00261	0.56212	0.0005928	0.01786	DCF1mInv
Tol_PS1mInv	183	0.0003985	0.0001863	0.07293	0.0001735	0.00120	Tol_PS1mInv

Pearson Correlation between X Variables

	f12_1m Inv	f14_1m Inv	f16_1m Inv	f17_1m Inv	f19_1m Inv	GS1mInv	DCF1mInv
f12_1mInv	-0.05668 0.4460						
f14_1mInv	-0.07284 0.3271	-0.06225 0.4025					
f16_1mInv	0.01199 0.8720	-0.05323 0.4742	-0.15548 0.0356				
f17_1mInv	-0.12023 0.1050	-0.09115 0.2197	0.04778 0.5206	-0.08090 0.2763			
f19_1mInv	0.51815 <.0001	-0.01704 0.8189	-0.09819 0.1860	-0.02093 0.7785	-0.08636 0.2451		
GS1mInv	0.10888 0.1423	-0.02444 0.7426	0.18385 0.0127	-0.04583 0.5378	0.17560 0.0174	0.02202 0.7673	
DCF1mInv	-0.17322 0.0190	-0.13674 0.0649	0.09242 0.2134	-0.06999 0.3465	0.08633 0.2452	0.03046 0.6823	-0.07006 0.3460
Tol_PS1mInv	0.41755 <.0001	-0.18906 0.0104	0.41235 <.0001	-0.07904 0.2875	0.00958 0.8976	-0.03329 0.6546	0.22006 0.0028	0.07961 0.2841

Pearson Correlation Coefficients, N = 183

Prob > |r| under H0: Rho=0

Variables: Stab4 U4 mixH4 mmHG4 K5 RH5 Precip5

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Stab4	183	5.02818	0.44446	920.15703	3.86667	6.06250	Stab4
U4	183	4.33517	1.14087	793.33679	1.93343	8.03340	U4
mixH4	183	1.02749	0.36186	188.03003	0.41398	2.09860	mixH4
mmHG4	183	762.25466	4.48193	139493	750.31733	773.08003	mmHG4
K5	183	286.53547	8.71882	52436	265.46528	303.27187	K5
RH5	183	66.51612	12.55339	12172	42.72000	91.79167	RH5
Precip5	183	6.60955	14.20507	1210	0	84.07400	Precip5

Pearson Correlation between X Variables

	U4	mixH4	mmHG4	K5	RH5
Stab4	-0.50892 <.0001				
U4	-0.39826 <.0001	0.29172 <.0001			
mixH4	0.05926 0.4255	-0.26353 0.0003	-0.23397 0.0014		
mmHG4	0.05779 0.4371	-0.27693 0.0001	0.17769 0.0161	-0.22479 0.0022	
K5	0.35230 <.0001	-0.42815 <.0001	-0.44864 <.0001	-0.01821 0.8067	0.35335 <.0001
RH5	-0.08714 0.2408	0.12823 0.0837	-0.20130 0.0063	-0.11505 0.1209	0.19164 0.0094	0.44748 <.0001

Prob > |r| under H0: Rho=0

The CORR Procedure
Simple Statistics
Pearson Correlation Coefficients, N = 183

Variables: mpX0 oX0 Tol0 Bzn0 Ebz0 MTBE0 PCE0 CCl40

	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
	183	3.24940	4.29215	594.64000	0.14500	51.21000	mpX0
	183	1.70768	6.51401	312.50500	0.06500	80.98000	oX0
	183	6.82175	5.83181	1248	0.11000	32.88000	Tol0
	183	1.50131	1.54081	274.74000	0.06000	18.06000	Bzn0
	183	1.34519	2.74160	246.17000	0.02000	36.24000	Ebz0
	183	5.74945	5.33444	1052	0.06000	27.17000	MTBE0
	183	1.10563	3.08295	202.33000	0.10000	41.82000	PCE0
	183	0.84019	2.27013	153.75500	0.16500	31.23000	CCl40

8
Variable
mpX0
oX0
Tol0
Bzn0
Ebz0
MTBE0
PCE0
CCl40

Pearson Correlation between Y Variables

	oX0	Tol0	Bzn0	Ebz0	MTBE0	PCE0
mpX0	0.61564					
	<.0001					
oX0	0.43479	0.25316				
	<.0001	0.0005				
Tol0	0.90230	0.37392	0.36673			
	<.0001	<.0001	<.0001			
Bzn0	0.91252	0.47358	0.30002	0.86511		
	<.0001	<.0001	<.0001	<.0001		
Ebz0	0.32187	0.13644	0.22987	0.33195	0.21578	
	<.0001	0.0655	0.0017	<.0001	0.0034	
MTBE0	0.86129	0.40914	0.24650	0.82914	0.94832	0.21629
	<.0001	<.0001	0.0008	<.0001	<.0001	0.0033
PCE0	0.83145	0.41009	0.20346	0.79918	0.94896	0.16167	0.98110
	<.0001	<.0001	0.0057	<.0001	<.0001	0.0288	<.0001

Prob > |r| under H0:Rho=0

CCl40
CCl40

The CORR Procedure

mpX0
Simple Statistics

Pearson Correlation Coefficients, N = 183

Variables: LnmpX0 LnoX0 LnTol0 LnBzn0 LnEbz0 LnMTBE0 LnPCE0 LnCCl40

	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
LnmpX0	183	0.81562	0.86257	149.25934	-1.93102	3.93593	LnmpX0
LnoX0	183	-0.09397	0.81125	-17.19703	-2.73337	4.39420	LnoX0
LnTol0	183	1.52498	1.00983	279.07109	-2.20727	3.49286	LnTol0
LnBzn0	183	0.12780	0.75810	23.38804	-2.81341	2.89370	LnBzn0
LnEbz0	183	-0.27229	1.18613	-49.82883	-3.91202	3.59016	LnEbz0
LnMTBE0	183	1.24947	1.24054	228.65228	-2.81341	3.30211	LnMTBE0
LnPCE0	183	-0.32811	0.81600	-60.04496	-2.30259	3.73337	LnPCE0
LnCCl40	183	-0.45530	0.53675	-83.32010	-1.80181	3.44138	LnCCl40

Variable

LnmpX0
LnoX0
LnTol0
LnBzn0
LnEbz0
LnMTBE0
LnPCE0
LnCCl40

Pearson Correlation between Y Variables

	LnoX0	LnTol0	LnBzn0	LnEbz0	LnMTBE0	LnPCE0
LnmpX0	0.90077					
	<.0001					
LnoX0	0.62676	0.59952				
	<.0001	<.0001				
LnTol0	0.73978	0.70564	0.50258			
	<.0001	<.0001	<.0001			
LnBzn0	0.83877	0.77242	0.46285	0.61731		
	<.0001	<.0001	<.0001	<.0001		
LnEbz0	0.22934	0.23674	0.14432	0.22802	0.18678	
	0.0018	0.0013	0.0513	0.0019	0.0114	
LnMTBE0	0.46364	0.49825	0.27263	0.37960	0.39547	0.26369
	<.0001	<.0001	0.0002	<.0001	<.0001	0.0003
LnPCE0	0.27904	0.30135	0.14601	0.24379	0.28540	0.07292	0.33776
	<.0001	<.0001	0.0486	0.0009	<.0001	0.3266	<.0001

Prob > |r| under H0: Rho=0

General Comments:

Overall, the report reflects a substantial amount of high-quality work, and reflects good practices in ensuring the quality of geographic data for use in subsequent analysis. The statistical approaches are supported by independent data, including relative vapor pressures for BTEX species.

General comments follow, with detailed in-line comments following.

1.
The choice of ordinary (multiple) linear regression for analysis of this data set is acceptable, but several caveats are in order. While a full description of the RIOPA data collection protocol has not yet been published, it is EPA's understanding that several homes were monitored concurrently for 48 hr (say n per subset), after which a new set of homes was monitored. To conduct monitoring at 100 homes, 100/n = p different rounds of home data collection would have to be undertaken. This process introduces an issue of non-independence of data for homes collected during the same of each of the p rounds of data collection. During the same 48 hr period of monitoring, the homes being monitored shared the same meteorological data (used in the current analysis). As such, these data may be analogous to the "clustering" phenomenon in surveys (associated with a loss of sampling efficiency). While this is unlikely to have a significant impact on the magnitude of the regression coefficients, it may have a substantial effect on their estimated standard errors. One way to significantly strengthen the current analysis would be to include for 2-3 compounds (say, one species each of PM, VOC, and PAH) a sensitivity analysis in which a mixed effects model is applied to the data sets, to account for random within-"cluster" variation. In SAS, the PROC MIXED procedure would be used for such analysis. Addition of 1st order autocorrelation for data collected simultaneously would also be appropriate here.

It is now reported that typically 1 or 2 homes were sampled on a single day, though some days had 3 or 4 homes samples so clustering should not be an issue. PROC MIXED was run with date as the repeated variable and no autocorrelation was found (page 33-34).

1.5
On a related note, the low partial-Rsqr of most of the regression coefficients should be further explored. One interpretation is that spatial patterns are relatively small contributors to overall variability in ambient concentrations. Another is that given the RIOPA sampling approach, the small number of concurrently-monitored homes resulted in assignment of a larger portion of explained variability to day-to-day/"samling cluster" to "sampling cluster" variation than would be observed given a "balanced" design in which spatial and temporal variability would be more seperable. Recently, the Battelle Memorial Institute conducted an analysis of sources of variability in EPA's pilot project for air toxics monitoring in ambient air within several cities nationally. At a series of fixed sites with simultaneous measurements, within-city spatial variability contributed almost as great a fraction of total variability as temporal variability (Battelle Memorial Institute and Sonoma Technologies, Inc.

(2003) Draft technical report for Phase II air toxics monitoring data: analyses and network design recommendations. Prepared for Lake Michigan Air Directors Consortium, Des Plaines, Illinois 60018). A discussion of the role study design in interpretation of these results is appropriate in the report.

More details of the study design are reported and a copy of a paper in press detailing that information is included. The clustering due to either date or location is not a problem based on the study design, as homes were selected throughout the 18 month study period from all sections of the city without concentrating on any portion of the city during individual time periods.

2.

Appendix A and the results section discuss diagnostic procedures applied to regression outputs to determine multicollinearity. However, neither the Appendix nor main report provide reasoning for decisions to apply corrective measures or not. For instance, it is mentioned that the distance to gas stations is significantly correlated with distance to major roadways. What was the strength of this association? If greater than about 0.85, this could lead to unstable coefficients. The relative significance of the associations provide some assurance that variances are not super-inflated, but when one of the distance terms was removed, did the other remain stable? Such description is necessary for the reader to be able to properly interpret the regression results. Other areas where further rationale is needed include decisions not to correct multicollinearity in cases with failing diagnostics (e.g. condition index).

More details on the reason for the decisions have been included in Appendix A.

3.

Why were "traditional" residual diagnostics not employed? Cook's Distance, etc. provide the standard approach to such diagnosis, but the rationale for not using them is not provided here.

The approach used to look at the residual was a traditional residual diagnostic and is more clearly stated. The Cook's Distance was not use as it was more time consuming and not thought to provide additional information past what was obtain for the objective being considered, to derive a cohesive data base to examine the role of proximity on ambient concentration. The exclusion of outliers, which probably had other variable impacting their concentration, was taken to address this fundamental issue and is a restrictive approach to identify outliers, probably classifying some values as outliers that were not.

4.

Please include a separate reference section, rather than citing the entire source in the text itself.

Provided

COMMENTS OF RICH COOK, EPA OTAQ

Chad --

I only had a chance to skim this before going on AL, but I have few comments:

1) In the section "National Emissions Inventory for 1999," I think a little more discussion of how county level VMT is developed would help. Pechan actually starts with State level VMT reported in HPMS by States (from sampling), which is then allocated to the county level using roadway miles for 12 functional classes and vehicle class splits. This is briefly oulined in the the technical documentation for the NEI. Joe Somers can help with a description if needed.
The VMT analyses was used as a guide to indicate which roadways to group together and in the statistical analyses. The actual emissions were not included in the regression equations. This is now stated in the text. Thus, a more detailed description of how they were derived is not warranted.

2) Table 10 -- residential ambient air concentrations -- I think it would be helpful to compare these data to local ambient monitor data, or maybe even national averages from AIRS, presuming resources permit. Aldehyde concentrations are much higher than typically seen at ambient monitors. I wonder why?
Concentrations in the area measured by NJDEP has been added to the table.

3) When discussing why there is not a roadway proximity relationship for aldehydes, it might be worth presenting some estimates of the secondary contribution. I know that some modeling has estimated 90% of formaldehyde is secondary. Again, this is subject to resource availability.
It is not clear to me how to include more on secondary contribution to aldehydes using the approach taken other than what was done, examining the data by splitting it into days above 10C and below, where different amounts of secondary production should occur. More detailed source emission modeling, which includes secondary production for formaldehyde, is being done by Dr. Panos Georgopoulos with funding from the ACC and may address this issue in the future.

4) I am suprised there is a signal for the two PAH compounds they measured. Nationwide, less than 20% of PAH emissions are from mobile sources. This suggests a pretty strong raodway effect, I think, given all the noise.
The effect does seem strong, but the compounds were selected as ones with major mobile contributions.

5) Cliff says that coronene (which is mispelled in several places) is associated more with gasoline vehicles and benzo(ghi)pyrelene more with diesels, but that their analysis saw no clear difference in source contributions. I checked the emission factors we used in the 1999 NEI and found the following:

The text has been altered to indicate the coronene is predominantly gasoline vehicles derived with the appropriate reference while benzo(ghi)pyrelene is derived from both gasoline and diesel vehicles.

a) Average emission rate for light duty vehicles and trucks (Norbeck, J. M., T. D. Durbin, and T. J. Truex. 1998. Measurement of Primary Particulate Matter Emissions from Light Duty Motor Vehicles. Prepared by College of Engineering, Center for Environmental Research and Technology, University of California, for Coordinating Research Council and South Coast Air Quality Management District. Tables 16 and 17) = 0.017 mg/mi

b) Average emission rate for heavy duty diesels (Watson, J. D., E. Fujita, J. C. Chow, and B. Zielinska. 1998. Northern Front Range Air Quality Study. Desert Research Institute. See Table 4.4-4, page 4-41.) = 0.013 mg/mi

So I am wondering what the source of data is that shows benzo(g,h,i)pyrelene is coming mostly from diesels. There is no reference in the report.

This is a good product. I hope these comments help.

Rich Cook
Environmental Scientist
U.S. EPA
Office of Transportation and Air Quality
2000 Traverwood Drive
Ann Arbor, MI 48105
Phone: 734-214-4827 Fax: 734-214-4939

Stephen Graham
09/13/2004 10:53 AM

To: Chad Bailey/AA/USEPA/US@EPA
cc: Janet Burke/RTP/USEPA/US@EPA
Subject: Re: RIOPA draft report

Hi Chad,
Some brief comments and questions. Overall, the draft needs work on sentence structure in the both the text and descriptions in the tables/figures.

1) Should have more about the sample collection design (what samples collected and when, for those included in this work) in background
More has been included and an in press paper is provided to give greater details.

2) All emission rate estimates in Table 4 are generally correlated (based on the Moblie 6.2 modeling, I assume).
a) there is artificial variability introduced for lesser emitted chemicals (e.g. the aldehydes) due to rounding
b) since they are different roadways, should they not have different distributions of vehicle classes on them resulting in different emission distributions or those chemicals listed?
c) unsure why this was done since not used in regressions
As indicated in the response to a comment by Richard Cook, this was done to facilitate the grouping of the roadways and individual emission rates were not included in the regression model so the effect of rounding is not important. The individual road classes are expected to have different vehicle distributions but each chemical and road class was individually examined so this effect should be accounted for by the analyses.

3) If using statistics for "normal" data, then one should use normal data or at least the most normal data. Several transformations were mentioned on page 33 and then correlations performed on each of the transformed variables. Why do all possible pairwise correlations, other than to 'see what gives the highest R'?
Only the Ln transformation of the concentration data was used in the analyses. As part of the exploratory work to make sure that an association was not missed more extensive correlations were evaluated.

4) In using multiple regression approaches (forward selection, backward elimination, etc) a statement about what each does to the estimate of variance is warranted.
Only the stepwise was used for deriving the final models. The others were run to verify that consistent results were obtained independent of how the regression equations were derived.

5) For influential ("outliers") statistics, why not use something more standard like Cook's D (apparently similar to what was used in this study, cook's uses F distribution rather than t), DFFITS, DFBETAS, COVRATIO?

See explanation provided the 1st reviewer.

6) condition number of 10-30 indicates mild collinearity, 30-100 moderate, >100 severe. Impact of excluding/flagging only severe category should be mentioned, although it looks like even parameters with severe collinearity were indeed included in the "final" models.
We agree that collinearity did exist, but it was predominantly in the meteorology variables, so the models were deemed acceptable for examining, particularly in a semi-quantitative manner, the role of proximity.

7) coronene was misspelled several differing ways.
Fixed

8) number of outliers in table 15 is not consistent with Appendix regression outliers. For example Table 15 lists 13 outliers for m,p-xylene, page A-3 states 17, Table A-2 states 20. QA check should be done here.
Fixed

9) In observing some of the stats for m,p-xylene, it seems that a 5-parameter model was best and used, rather than a 7 parameter mentioned in page A-3
Fixed

10) Table 16
a) lists 5 different significance levels ranging from 0.0001 through 0.105 (and I think in the text it is mentioned on occasion as highly significant, more significant, etc.). Establish a level of significance (e.g., $p<0.05$), and either something is statistically significant or not, rather than varying degrees of significant.
For the final model, which was based on a stepwise procedure, $p<0.15$ was used as the criterion for inclusion of a variable. The other routines were allow to have less stringent significance criteria as part of the exploratory analyses.

b) does not indicate significance level for aldehyde, PM, PAH, and OC/EC parameter estimates
All used $p<0.15$

c) precip units are not mentioned. it is apparently a significant parameter for the PAHs only, but it did not really rain/snow that much over the study period (maximum listed in table 14 is 0.13 mm if units are correct). Would one expect that much washout from so little precipitation? Even if it were inches, the median is 0.01, barely trace-level precipitation. If it is real, why no impact to the PM since essentially these PAH would all be associated with some form of particulate matter? I suspect that the precipitation is acting as a surrogate for some other parameter that has not been measured or possibly systematic error in PAH measurements.
Units now given. The regression suggests and association not an explanation. It could be another variable that both correlate with.

d) for ethyl benzene, inverse squared transformation was used and coefficient estimate is 167.14 in table A-10 and also in Appendix C, however is listed as 0.17 in Table 16. This should be corrected, but I have a comment: It is good to see a general consistency among BTEX coefficient estimates as expected, however, not sure why the inverse squared was used outside of "it made a better model". It is not very significant (r2=0.16) and would rather see it in the same units as the others.
All now use inverse of the transformed variable, not square.

e) in general the distance parameters (FC, GS, DCF, Truck, etc) did not add very much to explaining variation in residential concentrations, even for the true mobile source chemicals. This not surprising since the chemical is more than likely to never travel on a direct vector from highway A to home 1. This tells us immediately that if we want to know the impact a roadway is having on a residence, we need to do a better job of measuring this in the future (i.e, the 'dilution' or mixing with other air not originating from this source as a function of distance and micrometeorolgical conditions (estimated?), the time-of-day, day-of-week, month-of year (i.e., modified AADT)), otherwise we are really just taking stabs at it in the dark.
Agree

f) cannot remember why ambient concentration is not used as a parameter, even for a single central site monitor since it will probably do more for the model than the distance parameters.
No central site data were available.

11) correlation was mentioned among some of the input parameters- I would like to see what the actual correlations between FC14 and GS for the residences are rather than a brief mention. This may be evident in the predictions given in figures 10, 11, and 13 that show no effective difference in using the either the FC or GS distance parameter. What about stability and temperature by season, are there correlations here?
A correlation matrix has now been included in Appendix D.

12) not sure where ridge regression was used (technique mentioned on page 37)
Not used for the data presented, section has been removed.

13) no reference section included
Reference section added.