Appendix G-6: Characterizing Maryland Ozone by Meteorological Regime

CHARACTERIZING MARYLAND OZONE BY METEOROLOGICAL REGIME

FINAL

January 2006

Prepared for:

Monitoring and Planning Program
Air and Radiation Management Administration
Maryland Department of the Environment

Prepared by:

Science Applications International Corporation
615 Oberlin Road, Suite 100
Raleigh, North Carolina 27605

Contract Number MDE-03-6.0-AMA
SAIC Project Number 06-6312-71-2006-000

Executive Summary

Areas of Maryland exceed the 8-hour ozone National Ambient Air Quality Standard on various days during a normal summer. These ozone episodes during the summer season are caused by local emissions and/or emissions transported into Maryland. Previous episodic studies suggest that high ozone concentrations in the Baltimore-Washington area can be attributed to significant transport of ozone and its precursors into the Baltimore-Washington area on hot, humid days.

This study's goal was to quantify the effect that the transported ozone and precursors had on the daily 8-hour ozone maxima for the months from May to September. Instead of a classical modeling approach, statistical techniques using data mining tools were employed. To employ these techniques, daily ozone measurements, surface and aloft meteorology characteristics, indicators of persistent low level jets, and back trajectories were collected from the Environmental Protection Agency, University of Maryland at College Park (UMD), and the National Oceanic and Atmospheric Administration. The data was quality controlled to produce the Maryland Meteorology and Ozone Dataset (MMOD). The MMOD contains 150 fields of information and 2448 records that cover the years from 1989 through 2004.

After the MMOD was created, the data mining tools established five meteorological regimes (clusters) for the **Baltimore** data that showed the following tendencies:

- Cluster 0 (544 records) - Sunny, variable winds, and a higher temperature difference between upper air and surface conditions
- Cluster 1 (464 records) - Cloudy, cool days with winds from east and northeast and the most precipitation
- Cluster 2 (178 records) - Hot and humid with upper air winds from west and moderate precipitation
- Cluster 3 (760 records) - Low wind speeds, limited clouds and little precipitation
- Cluster 4 (497 records) - High wind speeds with little precipitation [surface winds from west, upper winds from northwest]

Similarly the **Washington, DC** data was divided into five clusters with the following tendencies:

- Cluster 0 (606 records) – Sunny, hot days with higher-speed surface and aloft winds from west
- Cluster 1 (484 records) - Cloudy, cool days with winds from east and northeast, most precipitation, high morning wind speeds, and low wind variability
- Cluster 2 (447 records) – Sunny with limited precipitation and high temperature differences between surface and aloft; variable low surface wind speeds with upper winds from the north
- Cluster 3 (695 records) - Low wind speeds from the west with limited clouds and precipitation
- Cluster 4 (216 records) - High temperatures with moderate clouds, low-speed variable winds from the south, upper winds from the west, and moderate precipitation

Each cluster was then subdivided into those days with measured persistent low level jets and those without. By comparison of the subclusters, it was determined that Baltimore and Washington 8-hour ozone concentrations increased by 7 and 5 ppb, respectively, on average days from May through September.

Association rule and classifier models were used to examine the fact that high ozone concentrations in Baltimore and Washington were often tied to high ozone concentrations during the previous night at high-elevation rural monitors. The regional nature of ozone concentrations was predicted by these data mining exercises by examining the nighttime ozone concentrations at Methodist Hill, Pennsylvania and Shenandoah National Park, Virginia. Based on these analyses, 23 to 36 ppb of regional ozone contributed to the Baltimore 8-hour ozone concentrations (which average 57 ppb). Similarly 21 to 32 ppb of regional ozone contributed to the Washington 8-hour ozone concentrations (average 53 ppb). These numbers are subject to uncertainty, and this was expressed by the synoptic correlations between the urban and rural monitors (0.52 to 0.77 for the different clusters).

One-day back trajectories from NOAA's HYSPLIT modeling were also used to distinguish transported and local ozone concentrations within the clusters. However, the HYSPLIT output was tied too closely to the meteorological parameters used for clustering and did not offer significant insight into the contributions of transport. The HYSPLIT results suggested that stagnant conditions in the Baltimore-Washington area affected ozone concentrations more than transport from long distances.

In summary, the low level jets were associated with increased ozone concentrations of 7 and 5 ppb in Baltimore and Washington nonattainment areas. The regional component of ozone was described as that portion that can be attributed to regional effects rather than localized effects in a single nonattainment area and represented by nighttime ozone concentrations at rural sites. The regional component for Baltimore and Washington represented 39 to 64 percent of the measured ozone over the range covering one standard deviation from the average. The HYSPLIT analysis was less successful at distinguishing transported ozone from local ozone concentrations because of the model's relationship to the meteorological parameters used for clustering.

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms*

AIRS – Aerometric Information Retrieval System
CAMx – Comprehensive Air quality Model with Extensions
CART – Classification and Regression Tree
CASTNET – Clean Air Status and Trends Network
CMAQ – Community Multiscale Air Quality
DOW – Day Of Week
EDAS - Eta Data Assimilation System
EM – Expectation Maximization algorithm
EPA – United States Environmental Protection Agency
GMT – Greenwich Mean Time
hr – hour
HYSPLIT – Hybrid Single-Particle Lagrangian Integrated Trajectory
IAD – Washington-Dulles International Airport
LLJ – Low Level Jet
m – meters
mb – millibars
MDE – Maryland Department of the Environment
MH – Methodist Hill, PA
MMOD – Maryland Meteorology and Ozone Dataset
MM5 – Penn State University/NCAR Mesoscale Model
NAA – nonattainment area
NAAQS – National Ambient Air Quality Standard
NCAR – National Center for Atmospheric Research
NCDC – National Climatic Data Center
NOAA – National Oceanic and Atmospheric Administration
NOx – oxides of nitrogen
O_3 – ozone
PM – particulate matter
ppb – parts per billion
ppbv – parts per billion by volume
Shen – Shenandoah National Park, VA
SIP – State Implementation Plan
SO_2 – sulfur dioxide

* MMOD attribute names are covered in Table 3-1.

1. Background

In April 2004 the United States. Environmental Protection Agency (EPA) designated Baltimore and Washington areas as nonattainment for the criteria pollutant ozone under the new 8-hour standard (85 ppb). The designation requires the state of Maryland to develop a State Implementation Plan (SIP) to show how planned emission reductions will reduce the ambient ozone concentrations in future years. However, different modeling studies show that regional transport of ozone across state boundaries is significant in the eastern United States. In response to this evidence, the EPA issued the NOx SIP Call in an effort to curb emissions of oxides of nitrogen (NOx) and consequently ambient ozone concentrations in eastern States[1]. The NOx SIP Call reported that thirty-five, fifteen, and seven percent of the ozone concentrations during the highest 1-hour ozone episode in Baltimore could be attributed to Virginia, West Virginia, and Ohio sources.

To support the findings of regional transport, various modeling studies have previously been done (e.g., Ryan *et al.*[2]), nocturnal low level jets have been measured and shown to carry ozone and its precursors[2,3], and predominantly rural, upwind sites have measured high ozone concentrations at night when prevailing winds blow toward Baltimore and Washington. These studies have been characterized by upper air wind profilers, higher elevation monitors and back trajectory analyses. Studies by other investigators (e.g., Reitebuch *et a* and Corsmeier *et al*)[4,5] show that surface ozone concentrations increase regionally when vertical mixing of the LLJ occurred.

The CMAQ modeling studies have focused on days with recorded high ozone concentrations. Thus, the models are often fit to describe only high ozone concentrations. Preliminary studies from the University of Maryland at College Park (UMD) show ozone reductions of 10 to 19 ppb when SO_2 and NO_x are reduced in Maryland[6]. Significant resources are required to operate the CMAQ models (e.g., temporally resolved gridded emission inventories and Fifth-Generation NCAR / Penn State Mesoscale Model 5 (MM5)), and long-term modeling exercises sometimes exceed computer processing capabilities. However, these models clearly show that ozone episodes in Maryland are often associated with stagnant conditions preceded by transport from the west and northwest[2]. Preliminary studies from UMD suggest that transport can account for a major portion of the ozone in Baltimore and Washington under certain meteorological conditions[7,8]. Although local NO_x emissions reductions may be effective at ameliorating ozone concentrations during the most extreme episodes, Choi *et al*.[8] suggest that additional controls upwind of the Baltimore-Washington area will likely be necessary to comply with the 8-hour ozone NAAQS.

When examining the July 10-15, 1995 ozone episode in Baltimore, a study using the Comprehensive Air quality Model with extensions (CAMx) revealed that less than 26 percent of the ozone could be attributed to Maryland sources[9]. Contributions from other states were strongly dependent on the wind directions for individual days.

Research at UMD has shown that, under particular meteorological conditions, a nocturnal low level jet may develop that travels from the Carolinas and Virginia into the state of Maryland[10].

The jet has been observed at an average height of 470 meters and is sometimes associated with high concentrations of ozone and its precursors. Ryan[10] indicated that low level jets that persist for at least 5 hours were considered "transport relevant" because ozone- and precursor-laden air parcels could travel from Richmond to Baltimore in this time. Studies have shown that ozone concentrations in persistent low level jets are on the order of 60 to 80 ppbv[10]. If the ozone and precursors from the nocturnal low level jet mix downward in the Baltimore-Washington area, they contribute to higher surface ozone concentrations. Verghese *et al.*[11] also found that nocturnal LLJs from the southwest formed at heights between 400 and 800 meters over Philadelphia during high ozone pollution events.

High ambient measurements of ozone concentrations have also been recorded at the fairly rural sites of Shenandoah National Park in Virginia and Methodist Hill in Franklin County, Pennsylvania. These sites also allow monitoring at higher elevation sites that may indicate transported pollutants above the boundary layer. The high ozone concentrations have been measured at night as well and may suggest that ozone concentrations do not decrease at night in rural areas when nighttime NOx reactions are usually expected to destroy ozone. Ryan *et al.* suggest that the most extreme ozone events in the Baltimore-Washington area occur when conditions are conducive to local ozone production and when ozone and its precursors are transported from heavily industrialized areas west and north of the area.[2]

The previous studies relied heavily on standard modeling efforts to assess the contributions from areas outside Maryland. This study instead focuses on statistical modeling efforts that rely predominantly on the meteorological measurements and observed ozone measurements.

2. Goals of Study

To develop meaningful statistical descriptions of Maryland's ozone air quality, it was important to be clear about the goals of the study. The stated goal of determining ozone trends based on meteorological regimes could be accomplished through a variety of approaches and using different techniques.

The standard approach to SIP development would be to use emission estimates, meteorological measurements, chemical constants, and transport/deposition rates in a detailed standard model. The model would be validated based on collected ambient measurements. However, significant uncertainty exists in the chemical constants and deposition rate descriptions for ozone. Significant uncertainty also exists in the emission inventory estimates, both in the areas of emission factors and activity data.

A different approach for SIP demonstrations would be to use statistical methods to choose effective control strategies. The inputs to statistical methods could be based initially strictly on observed variables that might be expected to influence the ozone concentrations. Ambient measurements from a metropolitan area serve as inputs to the statistical models, and so do the meteorological variables such as wind direction and cloud cover. Residual meteorological information (e.g., days since the last precipitation event) may also be included.

Statistical models are currently utilized to forecast Air Quality Action Days during the summer. The current ozone forecast methods often rely on human experts to fine-tune the predictions based on knowledge of events upwind (e.g., high ozone observations) or unusual meteorological conditions (e.g., stalled fronts). The classical statistical approach is through discriminant analysis, and the human expert operates using his/her natural neural network (brain). However, artificial neural networks may replace the human expert in the forecasting arena as more data is collected and interpreted.

An artificial neural network is one data mining tool that connects processing elements (nodes) with inputs, outputs, and processing at each node. The network uses a training set to adjust the strengths of the connections between nodes and weights the connections based on the available data. Other data mining tools include: data visualization which can help characterize gross observations (such as the relationship of ozone concentrations to temperature), decision trees to group variables based on a series of yes-no decisions (such as whether precipitation events of a certain magnitude affect ozone concentrations), or rule association programs. Rule association programs do not rely on hierarchical sets of conditions (such as those employed in the decision-tree method) and have the advantage of being able to infer rules based on overlapping sets of conditions[12].

Neural networks have proven more effective at predicting ozone concentrations than the Classification and Regression Tree (CART) methods. In Baltimore, Maryland, Ryan reported that neural net forecasts of ozone in 1998 were marginally better than the regression models but showed their best correlations on the good and moderate ozone days[13]. The neural network method was observed to have considerable utility for predicting Code Red, Orange, and Yellow

ozone days, even when exact ozone concentrations were not accurately forecast. Neural networks are often good at prediction but offer no description of the model. They cannot describe the effects that various attributes have on the parameter of interest.

The Cubist® software operates using association rules to develop piecewise linear regressions to predict numeric attributes. Cubist was used to determine the effect that Maryland's coal-fired power plant emissions had on the particulate matter measured at nearby IMPROVE sites[14]. Cubist was also used to determine the necessary conditions for mercury emissions from coal-fired power plants to affect local mercury deposition at two national Mercury Deposition Network sites[15].

With this understanding of the available data and tools, this study formed concrete goals that would enable investigators to come to meaningful conclusions as well as help future studies. The goals are listed below:

1) Consider all of the available data and their temporal spans. Many specialty studies have been conducted over the years, but their utility is limited for data mining studies. There have also been numerous ozone monitors in Baltimore and Washington that operated for just a few years, but inclusion of them in the data mining studies would introduce bias to data over the years. Any data set with fewer than five years of available data was not included in this work.

2) Develop a database that contains quality-assured data covering a long time series. Not all of the fields collected for the database were used in this data mining study (e.g., modeled elevations of air parcels) but might prove useful in future work.

3) Choose tools for data evaluation that are readily available to MDE staff. Limited resources prevent the acquisition of expensive niche software systems that might be used by large industries. Open source software is a better option and in line with program development at EPA.

4) Examine the transport indicators to determine the ozone levels associated with transport processes. After examining the distributions of the various attributes and some of the models that are created, it may be possible to quantify the ozone concentrations caused by transport from outside Maryland if the variables are sufficiently independent of the meteorology.

The two following chapters describe the results of these efforts. Chapter 3 discusses the database that was composed for long-term statistical analyses in Maryland. Chapter 4 describes the investigation of the data set to determine the effects of transport on ozone concentrations.

3. Construction of the Maryland Meteorological and Ozone Dataset

In order to conduct statistical analyses of Maryland air quality, it was necessary to compile historic data into a format that allows side-by-side comparison of the data. The Maryland Meteorology and Ozone Dataset (MMOD) represents daily records of Baltimore-Washington ozone concentrations, relevant meteorological parameters, and other relevant measured or modeled parameters. The records cover the months from May through September and the years from 1989 to 2004, for a total of 2448 records. The historic data was gathered from EPA, web resources, and the University of Maryland. In the future, additional daily information (e.g., daily emissions records from facilities) could also be included.

The parameters gathered are provided in Table 3-1. In addition, Table 3-2 presents the time span for the monitor and station data, and the locations are illustrated in Figure 3-1. A value of -999 was recorded for missing values. The sections below describe the data processing in more detail.

Table 3-1. Maryland Meteorology and Ozone Dataset

Field Name	Description	Units
Date	Date	yyyymmdd
METDATE	Date	mm/dd/yyyy
DOW †	*Day of the week (1=Monday through 7=Sunday)*	
EPA Air Quality System		
BltCt	Number of monitors used to calculate day's Baltimore nonattainment area 8-hour ozone (1989-2004)	Unitless
Blt8AvgO3	Average 8-hour daily maximum ozone for Baltimore monitors (1989-2004)	ppb
BltPreDay	*Average 8-hour daily maximum ozone for Baltimore monitors on previous day (1989-2004)*	*ppb*
WashCt	Number of monitors used to calculate Washington nonattainment area 8-hour ozone (1991-2004)	Unitless
Wash8AvgO3	Average 8-hour daily maximum ozone for Washington monitors (1991-2004)	ppb
WashPreDay	*Average 8-hour daily maximum ozone for Washington monitors on previous day (1991-2004)*	*ppb*
MHAvgNightO3	Methodist Hill average ozone from 12am-5am hourly readings (1996-2004)	ppb
DiffBltMH	*Difference between Baltimore average and nighttime Methodist Hill (1996-2004)*	*ppb*
DiffWashMH	*Difference between Washington average and nighttime Methodist Hill (1996-2004)*	*ppb*
ShenAvgNightO3	Shenandoah average ozone from 12am-5am hourly readings (1989-2004)	ppb
DiffBltShen	*Difference between Baltimore average and nighttime Shenandoah (1989-2004)*	*ppb*
DiffWashShen	*Difference between Washington average and nighttime Shenandoah (1989-2004)*	*ppb*

Table 3-1. Maryland Meteorology and Ozone Dataset (continued)

Field Name	Description	Units
Surface Meteorology from Beltsville CASTNET Site		
MAX_TEMP_C	Daily maximum temperature between 6am-10pm hourly readings (1989-2004)	Celsius
TEMP_COUNT	Number of hourly temperature values recorded between 6am-10pm (1989-2004)	Unitless
SOLAR_RAD_AM	Average morning solar radiation from 6am-11am hourly readings (1989-2004)	Watt/m^2
SOLAR_RAD_PM	Average afternoon solar radiation from noon-5pm hourly readings (1989-2004)	Watt/m^2
REL_HUM_AT_MAXT	Relative humidity at the time of maximum temperature (1989-2004)	Percent
PRECIP_MM	Total precipitation (1989-2004)	mm
PRECIP_HRS	Number of hours of when precipitation was recorded (1989-2004)	Hours
DAILY_U_MPS	U component of the daily surface wind speed (1989-2004)	m/s
DAILY_V_MPS	V component of the daily surface wind speed (1989-2004)	m/s
SCALAR_WIN_0_5	Average wind speed from 12am-5am hourly readings (1989-2004)	m/s
SCALAR_WIN_6_11	Average wind speed from 6am-11am hourly readings (1989-2004)	m/s
SCALAR_WIN_12_17	Average wind speed from noon-5pm hourly readings (1989-2004)	m/s
SCALAR_WIN_18_23	Average wind speed from 6pm-11pm hourly readings (1989-2004)	m/s
SIGMA_THET	Standard deviation of wind direction (1989-2004)	degree
MAX_OZONE_BELT_1hr	Maximum ozone one-hour average at Beltsville CASTNET site (1989-2004)	ppb
Profiler Data Provided by Charles Piety of University of Maryland		
Prof5hrLLJTime	Start time of the low level jet measured at Fort Meade (1999-2003)	yyyy mm dd hh:mm
Duration	Duration of the times when there is a low level jet that exists for at least 5 hrs (1999-2003)	hours
NCDC Integrated Global Radiosonde Archive		
Obhr_UA	Upper air observation hour (GMT) at IAD (1989-2004)	hour (11 or 12)
TempUa	Upper air temperature at IAD 850 mb (1989-2004)	Celsius
TempSfc	Surface temperature at IAD (1989-2004)	Celsius
TempDiff	*Temperature difference between upper air and surface at IAD (1989-2004)*	*Celsius*
DewptUa	Dew point depression at IAD 850 mb (1989-2004)	Celsius
DewptSfc	Dew point depression at IAD surface (1989-2004)	Celsius
DewptDiff	*Dew point depression difference between upper air and surface at IAD (1989-2004)*	*Celsius*
WinddirUa	Upper air wind direction at IAD 850 mb (1989-2004)	degrees
WinddirSfc	Surface wind direction at IAD (1989-2004)	degrees
WinddirDiff	*Wind direction difference between upper air and surface at IAD (1989-2004)*	*degrees*
WindspUa	Upper air wind speed at IAD 850 mb (1989-2004)	m/s

Table 3-1. Maryland Meteorology and Ozone Dataset (continued)

Field Name	Description	Units
WindspSfc	Surface wind speed at IAD (1989-2004)	m/s
WindspDiff	*Wind speed difference between upper air and surface at IAD (1989-2004)*	*m/s*
HYSPLIT Back Trajectories from Beltsville at 4 PM		
VLat10, VLon10, VHgt10*	Latitude/longitude/elevation six hours earlier (10 AM) at Beltsville for back trajectory beginning at 500 m (1997-2003)	decimal degrees
VMxDp10	Modeled mixed layer depth for that air parcel at 10 AM	m
*V10loc**	*Regional classifier six hours earlier at Beltsville for back trajectory beginning at 500 m (1997-2003). **N** indicates that air parcel is over Baltimore/Washington nonattainment areas, **S** that the air parcel is over another part of MD, **R** over Region 3, **W** over water, or **O** over another land mass.*	*N, S, R, W, or O*
XLat10, XLon10, XHgt10*	Latitude/longitude/elevation six hours earlier (10 AM) at Beltsville for back trajectory beginning at 1000 m (1997-2003)	decimal degrees
*X10loc**	*Regional classifier six hours earlier at Beltsville for back trajectory beginning at 1000 m (1997-2003). See V10loc* for description of variables.*	*N, S, R, W, or O*
XVLat10, XVLon10, XVHgt10*	Latitude/longitude/elevation six hours earlier (10 AM) at Beltsville for back trajectory beginning at 1500 m (1997-2003)	decimal degrees
*XV10loc**	*Regional classifier six hours earlier at Beltsville for back trajectory beginning at 1500 m (1997-2003). See V10loc* for description of variables.*	*N, S, R, W, or O*

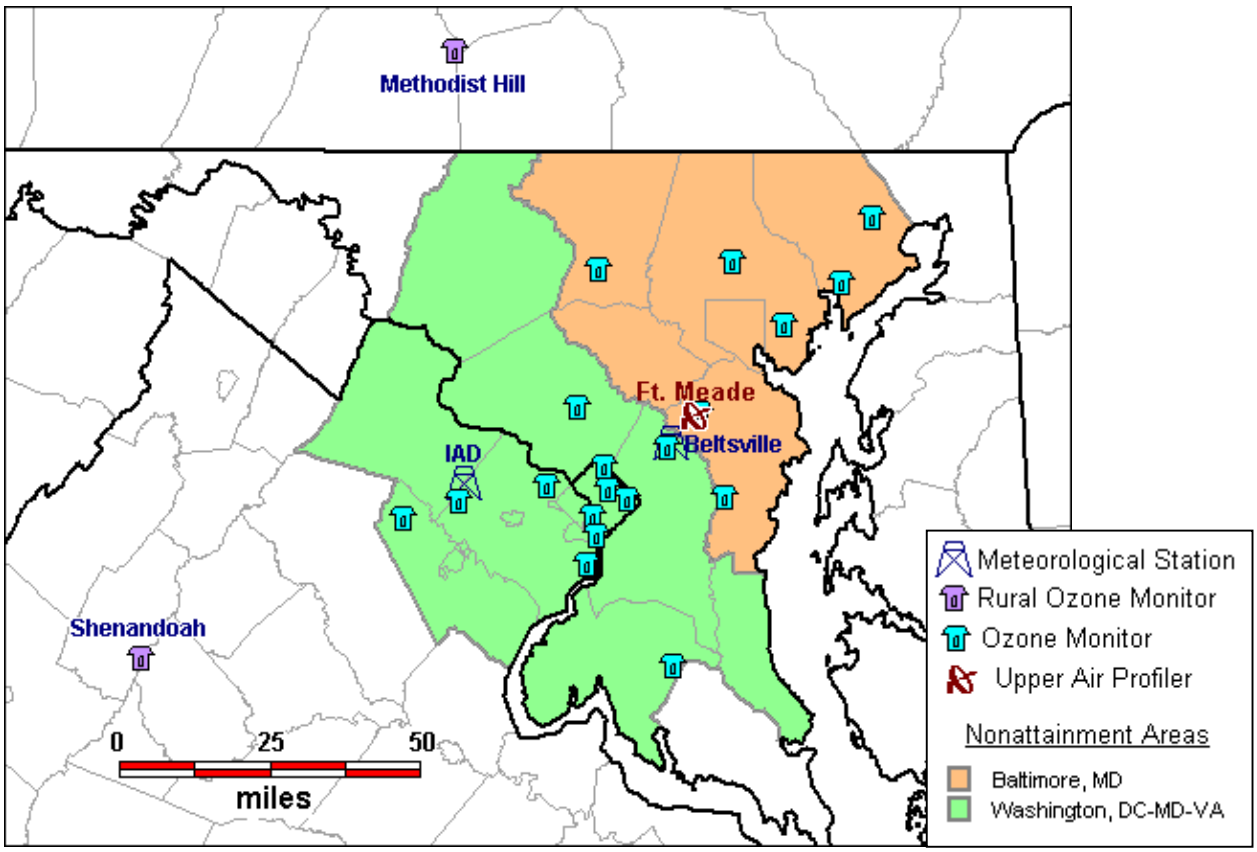
† Italicized fields represent fields calculated directly from other fields.

* The same HYSPLIT parameters are repeated for twelve, eighteen, and twenty-four hours earlier and are labeled xx4xx, xx22xx, and xx16Pxx, respectively, to represent their times before 4 PM.

Table 3-2. Monitor locations

Monitor Number or Name	Time Period	Latitude/Longitude	Elevation (m)
Baltimore Area Ozone Monitors			
240030014 – Queen Anne and Wayson	1989-2004	38.9468°N/ 76.6511°W	44
240030019 – Fort Meade	1989-2004	39.1011°N/ 76.7294°W	46
240051007 – Greenside Drive, Cockeysville	1989-2004	39.4608°N/ 76.6311°W	126
240053001 – Essex	1989-2004	39.3108°N/ 76.4744°W	5
240130001 – Old Liberty Road, Winfield	1989-2004	39.4441°N/ 77.0417°W	226
240251001 - Edgewood	1989-2004	39.41°N/ 76.2967°W	6
240259001 - Aldino	1990-2004	39.5633°N/ 76.2039°W	395
Washington, DC Ozone Monitors			
110010025 – Takoma Sc. Piney Branch Rd. & Dahlia St	1989-2004	38.9752°N/ 77.0228°W	91
110010041 – 34th and Dix Streets	1993-2004	38.8972°N/ 76.9528°W	8
110010043 – McMillian Reservoir	1994-2004	38.9188°N/ 77.0125°W	50
240170010 – Hughesville	1989-2004	38.5041°N/ 76.8119°W	49
240313001 – Rockville	1989-2004	39.1144°N/ 77.1069°W	126
240330002 – Goddard Space Center	1989-2004	39.02°N/ 76.8278°W	49
510130020 – S 18th and Hayes St.	1989-2004	38.8575°N/ 77.0592°W	171
510590005 – Cubrun Treat Plant	1992-2004	38.8938°N/ 77.4653°W	77
510590018 – Mt. Vernon Sherwood Hall Lane	1989-2004	38.7425°N/ 77.0775°W	11
510595001 – Lewinsville Balls Hill Rd.	1989-2004	38.9319°N/ 77.1989°W	106
511530009 – James S Long Park	1991-2004	38.8552°N/ 77.6356°W	111
515100009 – Alexandria Health	1989-2004	38.8108°N/ 77.0447°W	23
Methodist Hill Ozone Monitor			
420550001	1996-2004	39.9611°N/ 77.4756°W	676
Shenandoah Ozone Monitor			
511130003	1989-2004	38.5219°N/ 78.4361°W	1073
Surface Meteorology and Start Point for HYSPLIT Back Trajectories			
Beltsville CASTNET Site	1989-2004	39.06°N / 76.88°W	53
Upper Air Meteorology			
Washington Dulles (IAD)	1989-2004	38.9347°N / 77.4475°W	95
Wind Profiler			
Fort Meade	1999-2003	39.1011°N/ 76.7294°W	46

Figure 3-1. Map showing locations contributing data to the MMOD



3.1 Ozone Monitor Data

3.1.1 Baltimore and Washington Data

EPA provided the raw composite data for ozone for each of the Baltimore and Washington monitors[16]. This set of seven Baltimore monitors included AIRS monitor numbers 240030014, 240030019, 240051007, 240053001, 240130001, 240251001 and 240259001. These monitors were selected because they operated continuously between 1989 and 2004. Other monitors had less than five years of data and were not included in the data set because their limited data would have introduced bias to the ozone averages. No Baltimore City monitors met the criteria for inclusion. The data records contained the daily maximum 8-hour average for ozone between 1989 and 2004. The data were then filtered to include only the months between May and September for each year. If at least five of the seven monitors reported 8-hour averages for the day, those monitors were then averaged to give the average 8-hour daily maximum for ozone in the Baltimore area. This quantity was chosen because data mining techniques are expected to predict averages more readily than maxima, but future studies could consider highest daily reading from the monitor set. If the day did not have at least five monitors with a recording, then that day's average was recorded as -999. The dataset contains the date, that day's ozone value, a

count of how many monitors were included in that day's average, and the ozone value from the previous day.

The understanding was that pollution due to transport would generally affect an entire nonattainment area and not just a single monitor site. Therefore, only area-wide averages are represented in the MMOD to ensure that the influence of local emissions near a single monitor would not affect the data mining exercise. However, the ozone NAAQS is based on violations observed at any single monitor site and the values indicating violations are thus higher than the numbers contained in the MMOD.

The Washington monitors included AIRS monitor numbers 110010025, 110010041, 110010043, 240170010, 240313001, 240330002, 510130020, 510590005, 510590018, 510595001, 511530009 and 515100009. The data contained daily maximum 8-hour average values for ozone between 1991 and 2004. Ten years of data were available for twelve monitors within the Washington, DC area. If at least nine of those monitors had a recording for the day (representing 75% coverage), the ozone values were averaged to give the average 8-hour daily maximum for ozone in the Washington area. To keep a consistent dataset, the values between 1989 and 1990 were recorded as -999.

3.1.2 Methodist Hill and Shenandoah Data

The raw daily ozone data reports from 1996 to 2004 were downloaded for Methodist Hill, Pennsylvania from the EPA's Air Quality System website[17]. This monitor (420550001) was chosen due to its higher elevation, 579 meters, and because it lies in a rural area, 39.5740°N and 77.2832°W. This data included all of the hourly ozone records for each day in the time period. The data were filtered to contain only the six hourly readings between midnight and 5 AM for each day. These hourly values were then averaged to give a nighttime average concentration of ozone. The differences between the Baltimore and Washington area 8-hr ozone daily maxima and the Methodist Hill nighttime value were also computed.

The raw daily ozone data reports from 1989 to 2004 were downloaded for Shenandoah, VA from the EPA's Air Quality System website[17]. This monitor (511130003) was chosen due to its higher elevation, 1074 meters, and because it lies in a rural area, 38.5225°N and 78.4358°W. The data were filtered to contain only the six hourly readings between midnight and 5 AM for each day. These hourly values were then averaged to give a nighttime average concentration of ozone. The differences between the Baltimore and Washington area 8-hr ozone daily maxima and the Shenandoah nighttime value were also computed.

3.2 Surface Meteorology from Beltsville CASTNET Site

Level II validated meteorology data was downloaded from EPA's Clean Air Markets website[18]. The data contained hourly values collected at Beltsville, MD site, 39.06°N and 76.88°W, from 1989 to 2004. This monitor sits at an elevation of 53 meters. Only data that was flagged as being valid was included. This dataset included hourly temperature, solar radiation, relative humidity, precipitation, wind speed, wind direction, sigma theta and a 1-hour ozone maximum.

The maximum surface temperature between 6 AM and 10 PM was recorded for each day as well as a count of temperature values recorded during that time period. The solar radiation was averaged from 6 AM to 11 AM and then from noon to 5 PM, to give calculated values for morning and afternoon solar radiation. The MMOD tracks the relative humidity at the hour of the maximum temperature recorded for the day. The total precipitation for the day (in millimeters) is recorded in the MMOD. The precipitation count parameter tracks the number of hours with measured precipitation during the day. Using the surface wind speed and wind direction, the vector u and v wind components were calculated. The wind speeds between midnight-5 AM, 6 AM-11 AM, noon-5 PM and 6 PM-11 PM were averaged to give average scalar wind speeds for those periods during the day. Hourly sigma theta values were averaged if at least one sigma theta value was reported for that day. The last surface value recorded in this dataset was the maximum 1-hour ozone value recorded for that day.

3.3 Upper Air Data

Upper air data was downloaded from the Integrated Global Radiosonde Archive located on the National Climatic Data Center's website[19]. The dataset downloaded was for location 72403, which is Washington Dulles International Airport (IAD), 38.9347°N and 77.4475°W. This monitor sits at an elevation of 95 meters. Records with observation values of 11 or 12 GMT, which is 7 AM or 8 AM for eastern daylight time, were retrieved. For each day, the upper air temperature, upper air dew point depression, upper air wind direction and upper air wind speed at 850 mb were recorded. The surface temperature, surface dew point depression, surface wind direction and surface wind speed were also recorded for each day. A temperature difference value was calculated by subtracting the upper air temperature from the surface temperature. This same procedure was also done for dew point depression, wind direction and wind speeds. If the wind speed at the surface was reported as "0", then the wind direction at the surface was recorded as -180. The wind direction difference value was restricted to values between -180 and 180 degrees. This gave a dataset that included the upper air value (850 mb), the surface value and the value of the difference between the two for each meteorological parameter.

3.4 Low Level Jet Data

Wind profiler data from 1999 to 2003 was provided by Charles Piety from the University of Maryland – College Park[20]. The wind profiler is located at Fort Meade, but the included ozone data came from monitors located from within the Baltimore Air Quality Forecast Area. This data gave a 1-hour ozone maximum and an 8-hour ozone maximum. The data also recorded whether or not each day within this time period had a low level jet (LLJ) that persisted for at least five hours. If that particular day had a persistent LLJ, then the start time and duration of the LLJ was recorded. For days without LLJs, then starting times and durations of "0" were recorded. In order to keep the dataset consistent, a value of -999 was recorded for LLJ parameters in the years before 1999 and after 2003.

3.5 HYSPLIT Data

Back trajectory information was downloaded using the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) web interface located on the National Oceanic and Atmospheric Administration's (NOAA) Air Resources Laboratory web site[21]. The web interface program was run to gather information for the years 1997 through 2003. The settings used were model vertical velocity for vertical motion, plot projection was set to the default, and vertical plot height units was set to meters AGL. The other variables deal with the colors of the diagram. The back trajectories started at the Beltsville site, 39.06°N and 76.88°W, in Maryland and used the EDAS 80-kilometer meteorological data set. This site operates at an elevation of 53 meters. The back trajectories began at 4 PM and at 500, 1000, and 1500 meters above ground level. The trajectories gave the locations of the three air parcels at 4 AM and 10 AM of that same day and at 4 PM and 10 PM the previous day. The locations generated by the HYSPLIT model were recorded in degrees latitude and longitude. From these locations, the air parcel beginning at 4 PM and its back trajectories for that particular day could be categorized. If the air parcel was in the Baltimore or Washington nonattainment area the location field was labeled "N," in the State of Maryland labeled "S," in Region 3 labeled "R," outside the region labeled "O," and over water labeled "W.

4. Transport Assessment

The first section of this chapter discusses some of the early data assessment that was done to understand the parameter inter-relationships that would be observed later. The data assessment also served as a high-level quality assurance to determine which parameters were most reliably reported.

After the presentation of some important features of the data assessment, the approach for data mining will be discussed. The approach was determined from both the data assessment and trial-and-error with various algorithms. The approach involved the following broad steps:

1. Clustering the data to group instances with common meteorological conditions
2. Evaluating the effect of persistent low level jets on the clustered data
3. Determining the correlations between rural and urban data within the clusters
4. Analyzing the HYSPLIT data to determine if the parameters indicate additional transport
5. Presenting additional uncertainties in the results

4.1 Data Assessment

Before data mining tools determined transport characteristics, it was important to understand the distributions and relationships of the various attributes. In data mining nomenclature, attributes refer to the different fields in a database, and instances refer to the individual records. The Maryland Meteorology and Ozone Dataset (MMOD) contains many attributes but also includes many blank values where no observations were made. Many of the attributes (e.g., temperature and solar radiation) are closely related and cannot be considered independent.

Table 4-1 shows selected correlation coefficients when the attributes are compared directly among one another. Note that the absolute values of the correlation coefficients are high among some of the meteorological attributes, not just the ozone attributes. When conducting clustering analyses, the strong correlations may bias the clusters to weight more heavily toward the attributes that are similar. Therefore, it is important to use only as many of these interdependent attributes as necessary. Similarly, it is necessary to remove extraneous strongly interdependent attributes before generating association rules and classifiers because they may cancel one another out and have a zero net effect on the attribute of interest. During the data mining exercises, many interdependent attributes were included and then excluded to determine their effects on the results and data fitting ability. They remained in the analyses only if they significantly improved the data fitting capabilities.

One such example was the attribute of evening scalar wind speed from 6 PM to 11 PM (SCALAR_WIN_18_23). The wind patterns were generally well described by the earlier scalar wind speeds and the daily vector wind speed attributes, so inclusion of the evening scalar wind speed did not change the data mining results. It was also not intuitively obvious that the evening scalar wind speed would have an effect on the 8-hour ozone maximum for the day, so this attribute was not included in the data mining exercises.

Table 4-1. Correlation among numeric attributes in the MMOD
(Correlation coefficients over 0.50 and under –0.50)

Attribute 1	Attribute 2	Correlation Coefficient
Blt8AvgO3	Wash8AvgO3	0.95
BltPreDay	WashPreDay	0.95
Blt8AvgO3	MAX_OZONE__BELT_1hr	0.93
Wash8AvgO3	MAX_OZONE__BELT_1hr	0.92
TempUa	MAX_TEMP_C	0.79
SOLAR_RAD_AM	SOLAR_RAD_PM	0.78
ShenAvgNightO3	MHAvgNightO3	0.76
PRECIP_HRS	PRECIP_MM	0.75
MAX_TEMP_C	MAX_OZONE__BELT_1hr	0.75
SCALAR_WIN_6_11	SCALAR_WIN_12_17	0.73
Blt8AvgO3	MAX_TEMP_C	0.73
Blt8AvgO3	MHAvgNightO3	0.73
Wash8AvgO3	MAX_TEMP_C	0.72
MHAvgNightO3	WashPreDay	0.71
Wash8AvgO3	MHAvgNightO3	0.71
MHAvgNightO3	BltPreDay	0.70
MHAvgNightO3	MAX_OZONE__BELT_1hr	0.67
ShenAvgNightO3	WashPreDay	0.66
Blt8AvgO3	WashPreDay	0.62
Wash8AvgO3	WashPreDay	0.62
Wash8AvgO3	ShenAvgNightO3	0.62
ShenAvgNightO3	BltPreDay	0.61
Blt8AvgO3	BltPreDay	0.61
Wash8AvgO3	BltPreDay	0.60
WashPreDay	MAX_OZONE__BELT_1hr	0.59
Wash8AvgO3	SOLAR_RAD_AM	0.59
Blt8AvgO3	ShenAvgNightO3	0.58
BltPreDay	MAX_OZONE__BELT_1hr	0.58
Wash8AvgO3	SOLAR_RAD_PM	0.58
SCALAR_WIN_0_5	SCALAR_WIN_6_11	0.57
BltPreDay	MAX_TEMP_C	0.57
WashPreDay	MAX_TEMP_C	0.56
Blt8AvgO3	SOLAR_RAD_AM	0.56
ShenAvgNightO3	MAX_OZONE__BELT_1hr	0.54
Blt8AvgO3	SOLAR_RAD_PM	0.54
WindspUa	SCALAR_WIN_6_11	0.54
MHAvgNightO3	MAX_TEMP_C	0.51
MAX_OZONE__BELT_1hr	SOLAR_RAD_AM	0.51
SCALAR_WIN_12_17	DAILY_U_MPS	0.51
SCALAR_WIN_12_17	WindspUa	0.51
BltPreDay	TempUa	0.50
SOLAR_RAD_AM	PRECIP_HRS	-0.55
SIGMA_THET	SCALAR_WIN_0_5	-0.55
SOLAR_RAD_PM	PRECIP_HRS	-0.56
SIGMA_THET	SCALAR_WIN_6_11	-0.61

None of the correlations observed in Table 4-1 were unexpected, and this served as an additional check that the data had been correctly processed and cleaned.

The understanding was that pollution due to transport would generally affect an entire nonattainment area and not just a single monitor site. Therefore, only area-wide averages are represented in the MMOD to ensure that the influence of local emissions near a single monitor would not affect the data mining exercise. However, the ozone NAAQS is based on violations observed at any single monitor site and the values indicating violations are thus higher than the numbers contained in the MMOD.

Figure 4-1 shows a box-and-whisker plot of the ozone concentrations reported in the MMOD. The 8-hour averages for the Beltsville CASTNET site were not computed because MDE and EPA would not have quality assured the data for designation purposes. Figure 4-1 shows that the nighttime ozone concentration percentiles at the rural Methodist Hill and Shenandoah monitor sites were not much lower than the peak 8-hour ozone concentration percentiles in the Baltimore and Washington nonattainment areas. For each of the displayed percentiles, the rural nighttime measurements were no more than 20 ppb lower than the urban daytime measurements.

Figure 4-1. Distribution of Ozone Concentrations in the MMOD

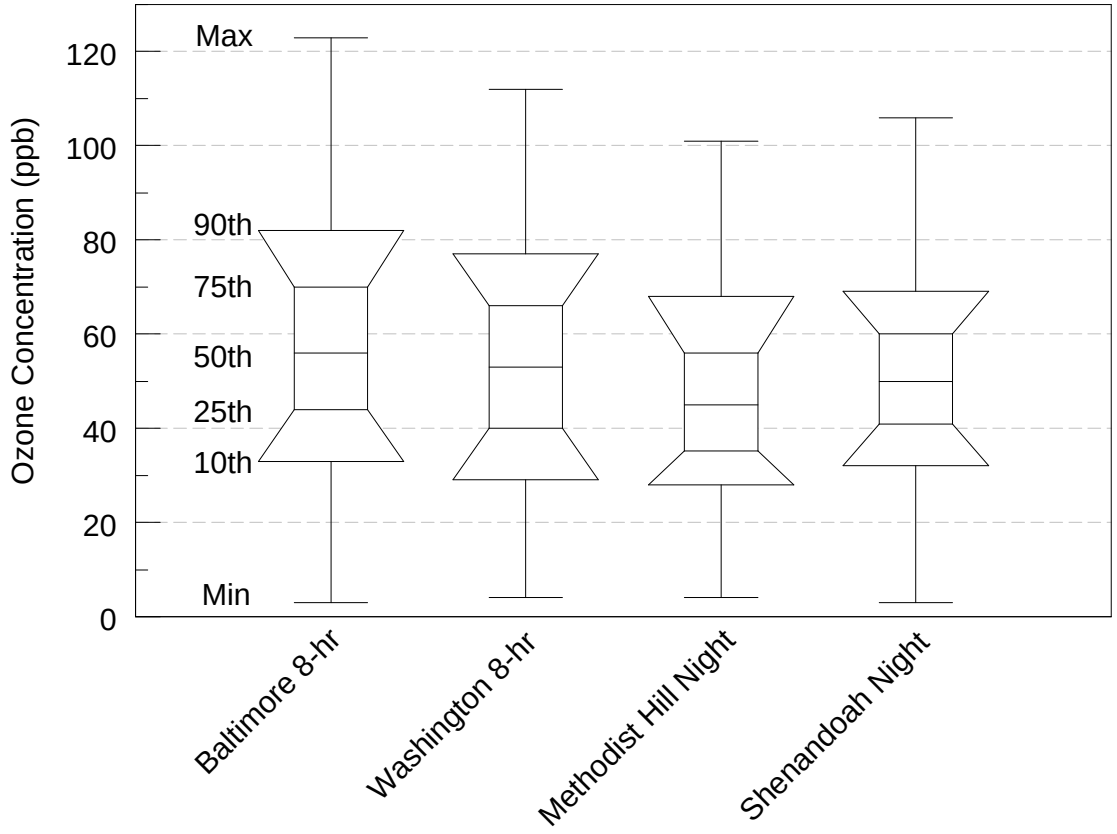
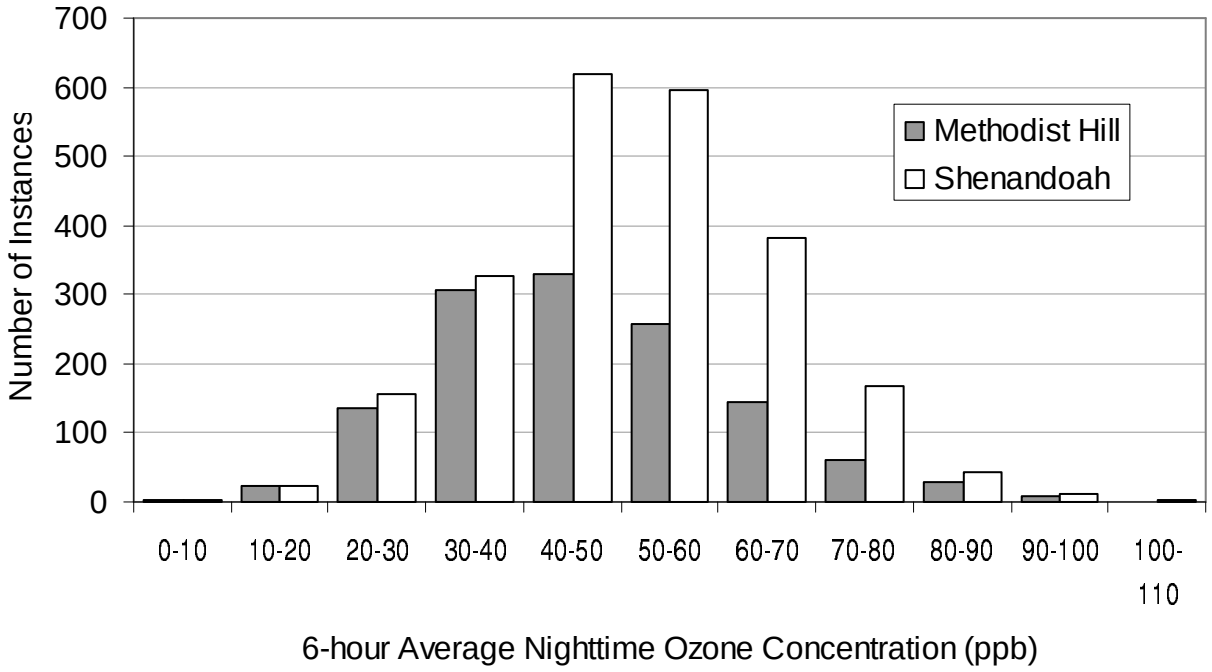


Figure 4-2 illustrates another feature of the Methodist Hill data; the data distribution is fairly smooth with just a single peak. Therefore, it is less likely that a certain set of infrequent meteorological conditions is responsible for the higher observed ozone concentrations (e.g.,

plumes from a nearby source). The same conditions were observed for the distribution of the Shenandoah data. Figure 4-2 also shows that the data are not always normally distributed and may show tails in the distribution. Future studies with the MMOD could focus on expressing each attribute with a well-defined distribution (e.g. log normal) to improve the fits associated with data mining.

Figure 4-2. Distribution of Rural Nighttime Ozone Measurements



4.2 Approach for Data Mining

Ozone concentrations are driven by many different parameters (e.g., solar radiation, precursor concentrations, and the mixing between surface and aloft air parcels). Unlike some hazardous air pollutants, the ambient chemistry associated with ozone is very complex and the emission sources are numerous. The goal of this approach is to determine if the observed ambient concentrations show any trends that can be linked to any measured transport indicators.

As discussed in previous sections, the transport indicators include observations of low level jet activity, nighttime ozone concentrations at two rural sites north and west of the nonattainment areas, and modeled back trajectories as indicators of where air parcels had previously been located.

To remove the influence of the local meteorological conditions (e.g., surface temperature) on ozone concentrations, five natural clusters were first determined for the entire data set. The desired number of clusters (five) was chosen to ensure an adequate number of instances in each cluster and a reasonable description of the number of summertime meteorological conditions. A similar approach has been undertaken by UMD using aircraft measurements to cluster daily data;

the eight directional clusters indicate the greatest ozone transport from the northern Ohio River Valley during the previous 48 hours[7]. The clustering relied only on meteorological conditions, previous days' ozone concentrations, days of the week, and years. The temporal attributes were included originally because they might be able to distinguish emissions patterns. However, it was later discovered that they were good indicators of inefficient clustering (see discussion in the next section).

To determine the effects of low level jets, the change in ozone concentrations within each cluster was determined by subdividing the cluster into cases with and those without persistent low level jets.

Because the nighttime concentrations at Methodist Hill and Shenandoah were generally high, determining their influences on Baltimore-Washington ozone required a different approach. The clustering and rural site concentrations were used to predict the concentrations within the nonattainment areas using association rules. The slopes of the piecewise linear regressions determined the contributions associated with a regional effect.

Using the HYSPLIT back trajectories, an approach similar to the one for low level jets was attempted. Within each cluster, dates with trajectories that had been within nonattainment areas and the state of Maryland twelve hours earlier were compared to those days with trajectories that were over land outside Region 3 twelve hours earlier.

The final section of this chapter discusses additional measures of uncertainty in the method that should be considered when evaluating the data mining estimates.

The Weka collection of machine learning algorithms was used for the clustering and many of the other data mining tasks under this study (http://www.cs.waikato.ac.nz/%7Eml/weka/index.html). Weka is open source software issued under the GNU General Public License. Because Weka is open source software, it is readily available to MDE staff for further analyses they may wish to perform.

4.3 Clustering

Weka offers four clustering algorithms[22]:

1. SimpleKMeans clusters the data about *k* initially chosen cluster centers by assigning each data instance to the nearest cluster and then iterating to find optimal cluster centers.
2. The Classit algorithm within the Cobweb clusterer employs an incremental clustering procedure based on merging and splitting operations.
3. The FarthestFirst algorithm uses the farthest-first traversal algorithm based on a fast, simple, approximate clusterer modeled on *k*-means.
4. The EM (expectation-maximization) algorithm assumes that a mixture of up to two normal distributions can describe each data set. It calculates the cluster probabilities and then the likely distribution parameters through a series of iterations.

All four clusterers were tested with the MMOD, and their arguments were varied to conclude at approximately five clusters that aimed to meet two conditions. The first condition was that singular or low numbers of instances did not form independent clusters. Small clusters would not offer sufficient data to compare the statistics once the clusters were further subdivided into transport and non-transport data sets. The second condition was that the clusterer should subdivide the data primarily based on meteorological and previous day ozone attributes and secondarily on temporal attributes (day of the week and year). The complete data set showed no clear dependence on the temporal attributes, so divisions of the clusters based on temporal attributes indicated that a meteorological characteristic was not being sufficiently considered.

The Simple K Means algorithm yielded five clusters that were difficult to differentiate. One cluster represented winds from the east, another westerly winds, and a third cloudy, cool days. However, the two other clusters showed no particular attributes that indicated they belonged to a category other than Remaining Instances.

In eighteen exercises with Classit/Cobweb and the Baltimore dataset, the acuity and cutoff were changed in efforts to report approximately five clusters with distributions that did not use single instances to describe clusters. The Cobweb algorithm (Classit) could form the clusters, but there were generally a few singular clusters. The only meteorological variable that this algorithm separated the data into was based on daily precipitation levels around 5 mm. Temperature, solar radiation, wind directions and speeds were not important classifiers, so this algorithm was dropped from further consideration.

The FarthestFirst algorithm performed better than the Cobweb algorithm with the number of clusters set to five. Again using the Baltimore data set, this clusterer did create a class for low solar radiation/temperatures, a class for high wind speeds, and a class for upper air winds from the east. However, the cluster distribution was poor (6, 129, 610, 35, and 1668 instances in the five clusters), and only seventeen clustered instances showed low level jet activity outside of the last cluster.

The best cluster distribution occurred using the EM algorithm. Several attributes (e.g., day of week) were added and subtracted to determine the effects on the clustering, but the EM algorithm altered the clustering very little. The Baltimore data was clustered into five data sets that can roughly be described as follows:

- **Cluster 0** (544 records) - Sunny, variable winds, and a higher temperature difference between upper air and surface conditions
- **Cluster 1** (464 records) - Cloudy, cool days with winds from east and northeast and the most precipitation
- **Cluster 2** (178 records) - Hot and humid with upper air winds from west and moderate precipitation
- **Cluster 3** (760 records) - Low wind speeds, limited clouds and little precipitation
- **Cluster 4** (497 records) - High wind speeds with little precipitation [surface winds from west, upper winds from northwest]

The figures below show some of the features of the clusters when examining the histogram distributions. Figures 4-3a and 4-3b show that Cluster 1 (cloudy, rainy days) have the lowest ozone concentrations and afternoon solar radiation. Figure 4-3c indicates the predominance of winds from the west in Cluster 4, and Figure 4-3d shows that Clusters 1 and 4 had fewer variable winds than other clusters (high wind speeds are more likely to be associated with lower variability than more stagnant conditions). Figure 4-3e illustrates that the morning upper air winds (at 850 millibars) in Clusters 2, 3, and 4 were almost never from the east.

The drop near 70 ppb in Figure 4-3a was traced back to the operation of the fewer than seven monitors within the Baltimore network. A smoother bell curve was produced using only the data when all seven monitors operated. The drop occurred in histograms with just five or six monitors in operation. This effect on the average illustrates how the operation of an individual monitor can affect a distribution. A similar discontinuity appears in the Washington ozone data (Figure 4-4a) and also seems to be related to the operation of all twelve monitors.

Since the only attribute difference between the Baltimore and Washington clustering exercises was the ozone concentration from the previous day and Table 4-1 reports a 0.95 correlation between the two attributes, it was expected that the Washington data clusters would appear very similar to those from the Baltimore data. Figure 4-4 shows the distribution of selected attributes among the clusters, but some of the Washington clusters have changed:

- **Cluster 0** (606 records) – Sunny, hot days with higher-speed surface and aloft winds from west
- **Cluster 1** (484 records) - Cloudy, cool days with winds from east and northeast, most precipitation, high morning wind speeds, and low wind variability
- **Cluster 2** (447 records) – Sunny with limited precipitation and high temperature differences between surface and aloft; highly variable low surface wind speeds with upper winds from the north
- **Cluster 3** (695 records) – Highly variable low wind speeds from the west with limited clouds and precipitation
- **Cluster 4** (216 records) - High temperatures with moderate cloudiness, low-speed variable winds from the south, upper winds from the west, and moderate precipitation

The fact that the cluster groupings varied from Baltimore to Washington illustrates that the data mining exercises may not always yield the same results. Table 4-2 shows how closely the Baltimore and Washington clusters compare. Optimization can be defined many different ways, and the algorithms iterate until the error reduction with each successive iteration becomes negligible. This approach means that a minimum error is reached, but it may not be the global minimum for the dataset. To investigate the variability in clustering, Weka users may change the random number seed (29 was used for the exercises in this study), and this will alter the order in which the data points are considered. For this investigation, the random number seed was not altered because the clustered data readily described familiar meteorological conditions.

Figure 4-3. Data distributions for Baltimore (blue=Cluster 0, red=Cluster 1, cyan=Cluster 2, blue-gray=Cluster 3, pink=Cluster 4)

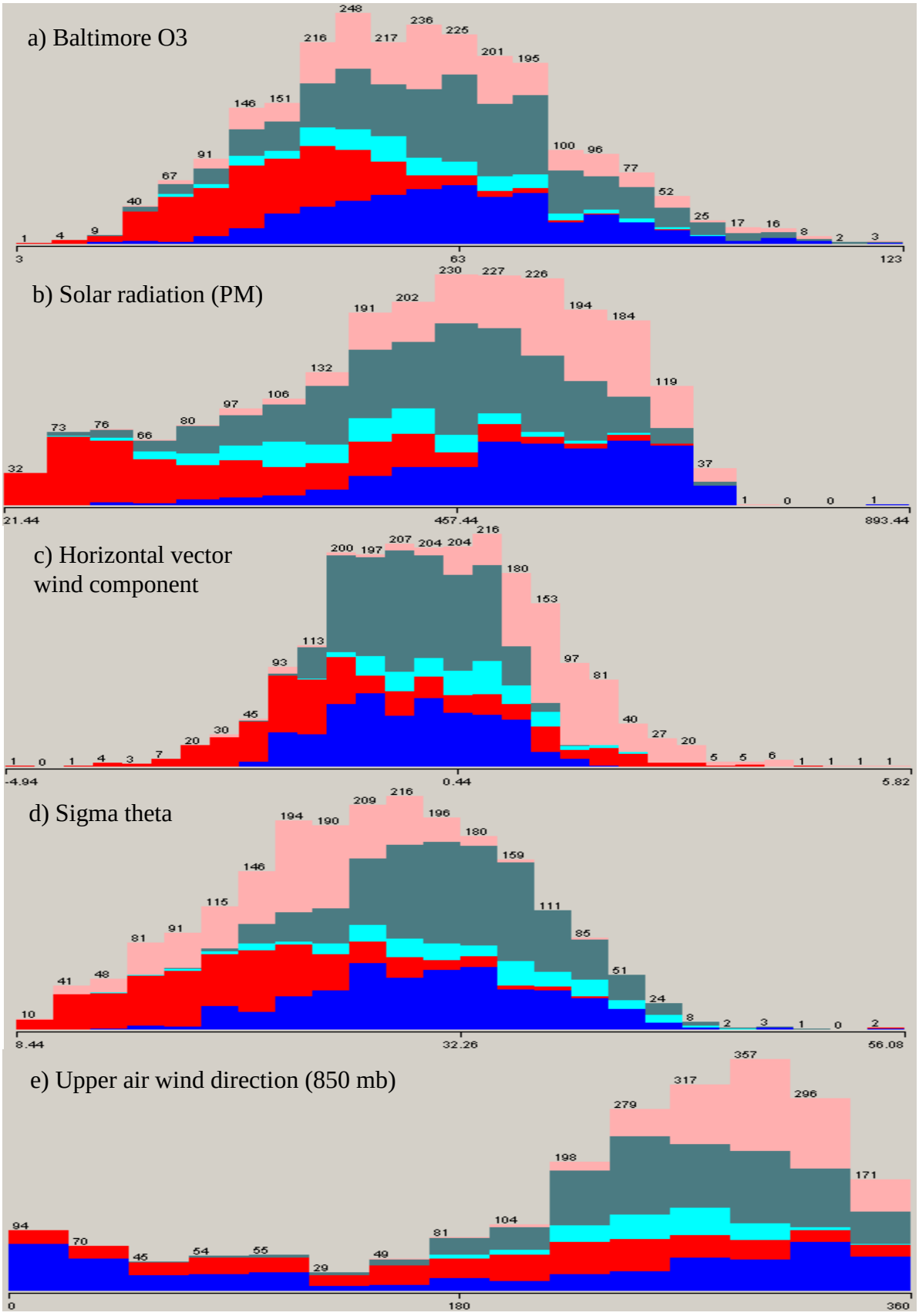


Figure 4-4. Data distributions for Washington (blue=Cluster 0, red=Cluster 1, cyan=Cluster 2, blue-gray=Cluster 3, pink=Cluster 4)

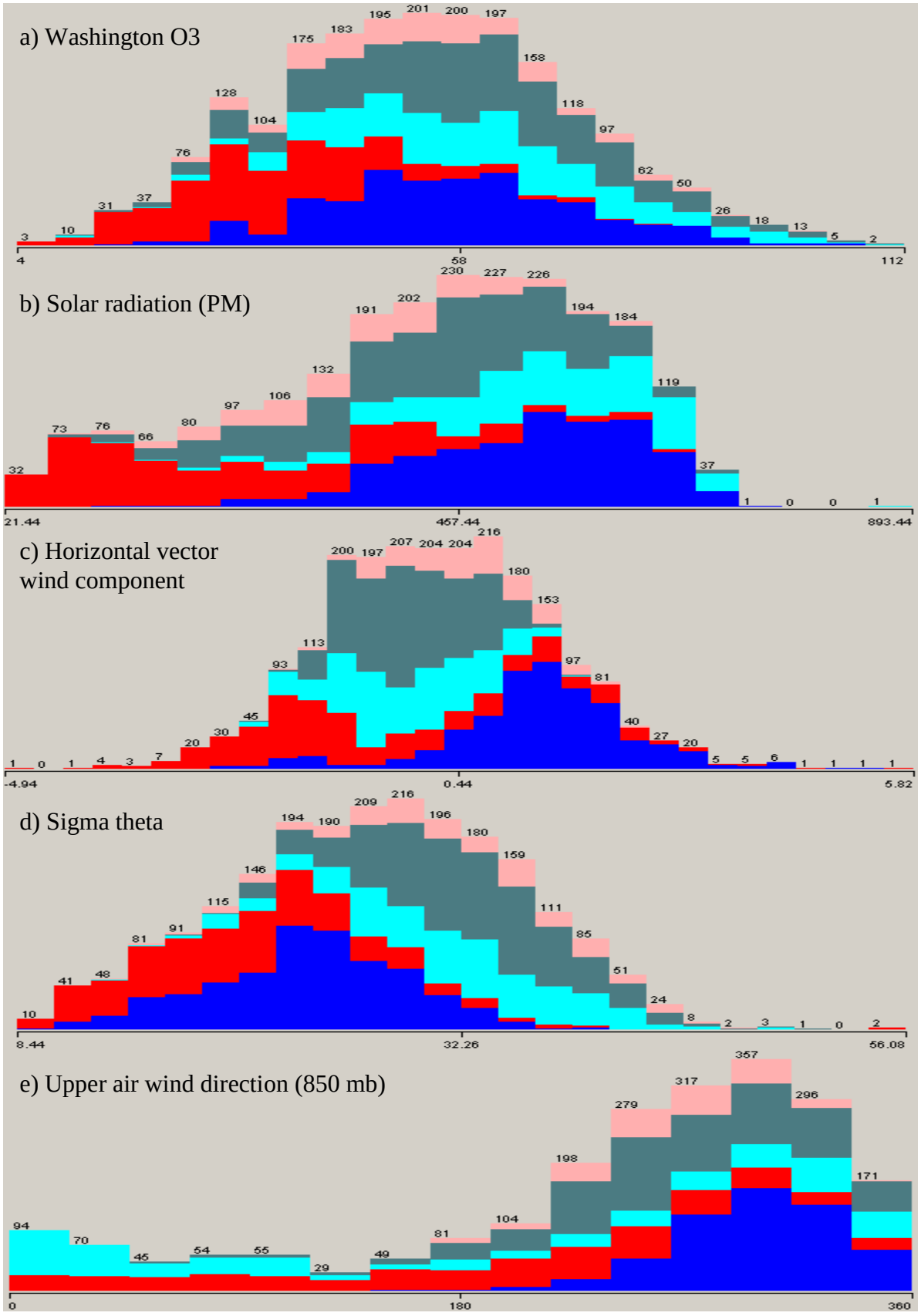


Table 4-2. Number of shared instances in Baltimore and Washington clusters

		Washington				
		Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Baltimore	Cluster 0	72	12	417	10	33
	Cluster 1	4	460	0	1	0
	Cluster 2	3	5	1	1	168
	Cluster 3	34	4	29	683	11
	Cluster 4	493	3	0	0	4

Table 4-3 presents the ranges and medians for some sample meteorological attributes within the clusters. Note that considerable overlap occurs among the attributes, so the bulleted descriptions above only indicate cluster tendencies and not specific meteorological conditions.

Table 4-3. Sample ranges and medians for Baltimore and Washington clusters

	Maximum Daily Surface Temperature (Celsius)	Temperature Difference Between Surface and 850 mb (Celsius)	Precipitation (mm)	Surface Wind Speed 6am-11am (m/s)	Wind Speed at 850 mb (m/s)
Baltimore					
Cluster 0	12 – 36 range (27 median)	-12 – 11 range (-2.0 median)	0.0 – 24 (0.0)	0.45 – 3.8 (2.0)	0.50 – 17 (6.0)
Cluster 1	8.1 – 35 (22)	-9.6 – 4.3 (-5.3)	0.0 – 83 (4.3)	0.77 – 7.5 (2.6)	0.50 – 28 (8.0)
Cluster 2	21 – 36 (29)	-7.8 – 1.0 (-4.3)	0.0 – 35 (7.4)	0.79 – 4.3 (1.8)	1.0 – 18 (7.9)
Cluster 3	14 – 37 (28)	-10 – 3.1 (-3.9)	0.0 – 2.0 (0.0)	0.38 – 3.0 (1.7)	0.50 – 16 (5.1)
Cluster 4	10 – 37 (28)	-12 – 3.0 (-4.6)	0.0 – 0.76 (0.0)	1.7 – 7.2 (3.1)	1.0 – 37 (11)
Washington					
Cluster 0	10 – 37 (28)	-12 – 4.4 (-4.2)	0.0 – 0.50 (0.0)	1.6 – 7.2 (3.0)	1.0 – 37 (10)
Cluster 1	8.1 – 35 (22)	-9.6 – 4.3 (-5.3)	0.0 – 83 (4.1)	0.45 – 7.5 (2.6)	0.50 – 28 (8.0)
Cluster 2	16 – 36 (27)	-10 – 11 (-1.5)	0.0 – 24 (0.0)	0.82 – 3.8 (1.9)	0.50 – 17 (5.7)
Cluster 3	15 – 37 (28)	-8.4 – 2.6 (-3.9)	0.0 – 1.8 (0.0)	0.38 – 3.0 (1.6)	0.50 – 14 (5.1)
Cluster 4	19 – 36 (29)	-12 – 3.5 (-4.3)	0.0 – 35 (5.8)	0.79 – 4.3 (1.9)	1.0 – 23 (7.7)

4.4 Influence of Low Level Jets

The University of Maryland provided analyses of persistent low level jets (LLJs) for this study. The Duration attribute in the MMOD reports the number of hours that LLJs were recorded if the measurements suggested that the jet persisted for five or more hours. Measurements were collected from 1999 through 2003 and did not distinguish direction or start time. Figures 4-5 and 4-6 suggest that the longest recorded jet persisted for seventeen hours.

Figure 4-5. Effect of LLJ duration on Baltimore ozone concentrations

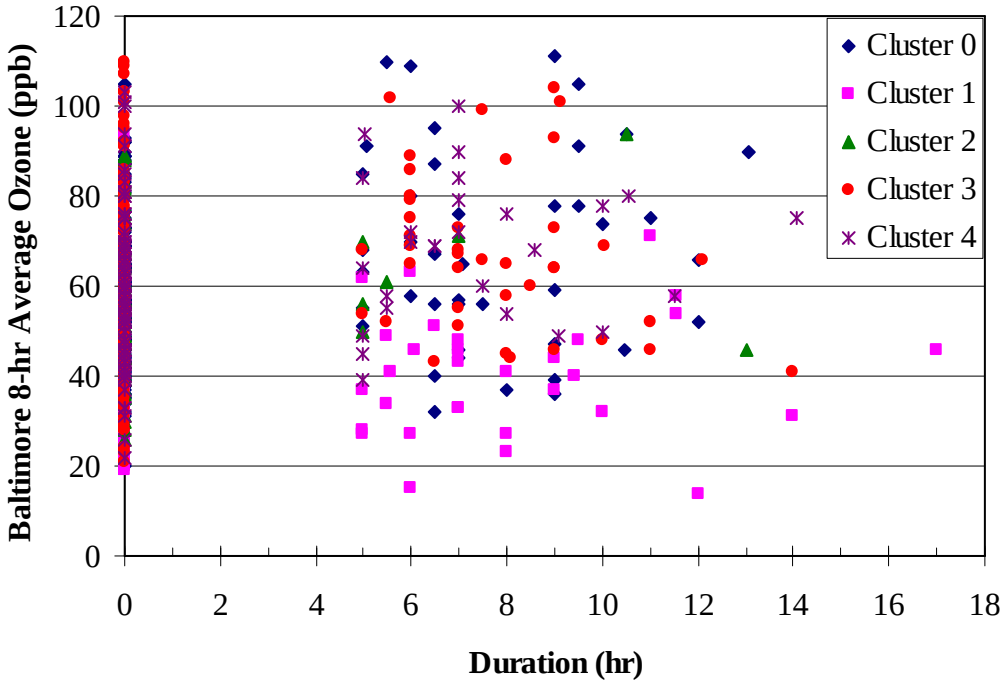
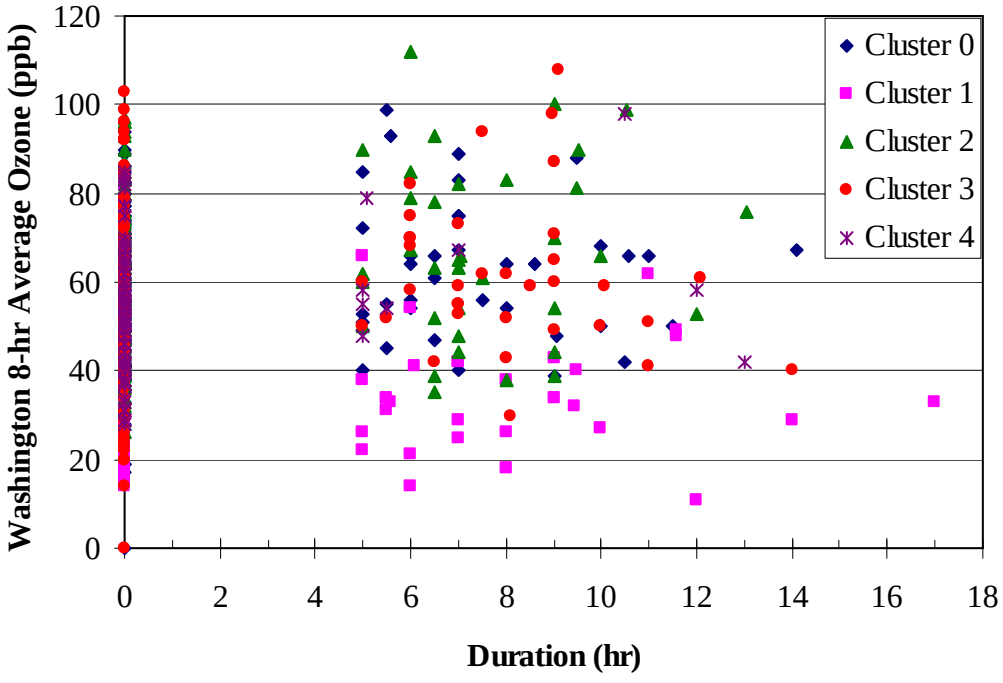


Figure 4-6. Effect of LLJ duration on Washington ozone concentrations



Figures 4-5 and 4-6 also show that LLJs were observed in all of the clusters and that they were present under many ozone concentrations. For the years 1999 through 2003 (the years when the University of Maryland recorded either persistent LLJ presence or absence), each cluster was

subdivided based on the presence or absence of a measured persistent LLJ. Tables 4-4 and 4-5 present the comparisons of cluster data with and without the influence of LLJs.

The *Average LLJ effect* for the individual clusters was calculated by subtracting the *Average without LLJ* from the *Average with LLJ*. The values for the LLJ effect range from –0.3 to +10.4 ppb in the individual clusters. To calculate the Average LLJ effect for all records, the cluster values were weighted based on the total count within each cluster. Note that the Cluster 1 data for both Baltimore and Washington showed a small effect from the LLJ. The Cluster 1 data for both data sets represents the cloudy, cool days with winds from the east and northeast as well as the most precipitation. The observed low level jets seemed to have the smallest effects on the ozone concentrations on these days.

Table 4-4. Effect of LLJ on Baltimore ozone data

Calculation	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	All Records
Average with LLJ (ppb)	68.2	40.3	64.0	67.6	68.2	
Std Deviation with LLJ (ppb)	22	14	16	18	16	
Number of records with LLJ	*41*	*31*	*7*	*40*	*27*	
Average without LLJ (ppb)	58.2	38.2	55.1	60.5	59.6	
Std Deviation without LLJ (ppb)	15	13	14	18	16	
Number without LLJ	*146*	*132*	*70*	*175*	*95*	
Average LLJ effect (ppb)	**+10**	**+2.1**	**+8.9**	**+7.0**	**+8.6**	**+7**
T-test probability	0.0044	0.44	0.20	0.026	0.016	
Total Count	*544*	*464*	*178*	*760*	*497*	*2443*

Table 4-5. Effect of LLJ on Washington ozone data

Calculation	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	All Records
Average with LLJ (ppb)	62.4	34.8	66.8	61.8	62.1	
Std Deviation with LLJ (ppb)	16	13	20	17	17	
Number of records with LLJ	*37*	*29*	*36*	*35*	*9*	
Average without LLJ (ppb)	58.4	35.0	56.4	56.0	52.7	
Std Deviation without LLJ (ppb)	15	12	16	18	14	
Number without LLJ	*119*	*138*	*129*	*152*	*78*	
Average LLJ effect (ppb)	**+4.0**	**-0.3**	**+10**	**+5.8**	**+9.4**	**+5**
T-test probability	0.18	0.92	0.0051	0.073	0.15	
Total Count	*505*	*409*	*408*	*575*	*192*	*2089*

Further proof that the LLJ had little effect on the Cluster 1 data is contained in the T-test probabilities of 0.44 and 0.92. These probabilities indicate that the LLJ data are 44% and 92% likely to be indistinguishable from the non-LLJ data. The T-test probabilities for the remaining clusters indicate that the measured ozone concentrations with and without LLJs are not likely from the same data distributions. The T-test works well when data is normally distributed; although the cluster subsets were small, the assumption of normal distributions appeared more reasonable than other common distributions. Across all clusters the presence of LLJs in Baltimore results in a 7 ppb increase in ozone concentrations, and a 5 ppb increase in Washington.

4.5 Correlation Between Rural and Urban Data

Analysis of the LLJ data was straightforward because the records could clearly be divided into sets indicating either the presence or absence of a persistent LLJ. However, Figure 4-2 shows that the Methodist Hill data cannot be clearly delineated into subsets, and the same is true for the Shenandoah site. To employ data mining tools in this situation, the cluster designations were used as an independent attribute, and the meteorological conditions were removed from the data mining exercise. The software was only permitted to use the following attributes to predict the ozone concentrations:

- Year
- Day of the week
- Previous day's ozone concentration in that nonattainment area
- Cluster classification
- LLJ duration
- Methodist Hill ozone concentration (12 AM-5 AM average)
- Shenandoah ozone concentration (12 AM-5 AM average)

By taking this simplified approach, the cluster designation represented the reactive conditions for the day, and the previous day's and rural ozone concentrations and LLJ duration indicated the mixture of available pollutants and precursors. The models were originally run including the maximum temperature as an attribute because five cluster designations are likely too few to generate accurate numeric predictions. This inclusion decreased the error of the generated models but also decreased the observed dependence of the ozone concentration on the nighttime concentrations at the rural monitors. Since it affected the model dependence on the mixture of pollutants and precursors, temperature was excluded from the model as an attribute for reactive conditions.

One appropriate and often successful model technique for numeric predictions are made for each subset. The Cubist® and M5 Rules algorithms use multivariate linear regression within the subsets to predict the attributes as a series of rules. One advantage to the association rule approach is that multiple rules may apply to overlapping subsets of training data. An association rule may take the form:

If

Methodist Hill nighttime ozone > 43 ppb and
Cluster = 0, 2, 3, or 4

Then

Baltimore ozone = 16.2 + 0.47 × (Methodist Hill nighttime ozone)
+ 0.23 × (Baltimore ozone on previous day)
+ 0.15 × (Shenandoah nighttime ozone)
+ 0.3 × (LLJ duration)

The predictive ability of models is determined by comparison with test data sets that are independent of the data used for training. A leave-one-out approach for data with n instances maximizes the size of training and test data sets by creating n models using n-1 instances; the remaining instance is used as the test data and the process is repeated with each instance taking a turn as the test data. Since the MMOD instances with available rural data were limited to 763, the leave-one-out approach took less than a minute to run.

Using a leave-one-out approach to determine the accuracy of the model, the Cubist® software was used to generate association rules and yield numerical predictions of the ozone concentrations (with the model set to use rules without nearest neighbors, 4% minimum rule cover, 763 cross folds, a maximum of ten rules, and 10% extrapolation):

```
Rule 1: [162 cases, mean 38.7, range 14 to 94, est err 8.5]

  if
    Cluster = cluster1                (Cool, cloudy days)
  then
    Blt8AvgO3 = 13.5 + 0.4 MHAvgNightO3 + 0.2 ShenAvgNightO3 + 0.3 Duration
                + 0.03 BltPreDay

Rule 2: [51 cases, mean 46.6, range 20 to 94, est err 10.1]

  if
    MHAvgNightO3 > 43
    Cluster = cluster1                (Cool, cloudy days)
  then
    Blt8AvgO3 = 33.8 + 0.14 MHAvgNightO3 + 0.06 BltPreDay
                + 0.03 ShenAvgNightO3

Rule 3: [235 cases, mean 48.4, range 20 to 76, est err 7.6]

  if
    MHAvgNightO3 <= 43                (low nighttime rural ozone)
    Cluster in {cluster0, cluster2, cluster3, cluster4}
  then
    Blt8AvgO3 = 9.9 + 0.57 MHAvgNightO3 + 0.21 BltPreDay
                + 0.21 ShenAvgNightO3

Rule 4: [367 cases, mean 68.4, range 34 to 111, est err 9.7]

  if
    MHAvgNightO3 > 43                (higher nighttime rural ozone)
    Cluster in {cluster0, cluster2, cluster3, cluster4}
```

```
    then
      Blt8AvgO3 = 16.2 + 0.47 MHAvgNightO3 + 0.23 BltPreDay
                  + 0.15 ShenAvgNightO3 + 0.3 Duration

Summary:

    Average  |error|                8.9
    Relative |error|               0.60
    Correlation coefficient        0.79
```

The cases covered only those years when the LLJ data were available (1999-2003). Note that the instances covered by Rules 1 and 2 overlap; Cubist averages the predictions from the two rules to estimate the Baltimore ozone concentration. The four association rules show that the Baltimore ozone concentration is affected by the Methodist Hill concentration by slopes ranging from 0.14 (Rule 2) to 0.57 (Rule 3). In order to calculate the overall effect of Methodist Hill on Baltimore ozone as expressed in the rules, the averages and standard deviations of the Methodist Hill and Shenandoah concentrations in each of the subsets were considered.

Table 4-6 presents the calculation of the effect of the rules. Although the data subsets are not normally distributed, the standard deviations were used as a measure of uncertainty in order to bound the estimates of the rule effects. For each rule, the lower estimate of the rule effect is calculated by the expression:

(MH avg – MH std dev) × MH Rule Effect + (Shen avg – Shen std dev) × Shen Rule Effect

and the upper estimate of the rule effect by the expression:

(MH avg + MH std dev) × MH Rule Effect + (Shen avg + Shen std dev) × Shen Rule Effect

Table 4-6. Calculation of Cubist® rule effects on Baltimore ozone concentrations

	Rule 1	Rule 2	Rule 3	Rule 4
Baltimore Ozone Average (ppb)	39.2	47.2	48.9	68.9
Standard Deviation (ppb)	13	13	12	16
Count	464	83	429	615
Methodist Hill Ozone Average (ppb)	37.8	52.9	34.5	58.4
Standard Deviation (ppb)	13	7	7	11
Rule Effect (coefficient)	0.4	0.14	0.57	0.47
Shenandoah Ozone Average (ppb)	43.7	56.1	43.8	61.1
Standard Deviation (ppb)	15	14	10	12
Rule Effect (coefficient)	0.2	0.03	0.21	0.15
Lower Estimate of Rule Effect (ppb)	16	8	23	29
Upper Estimate of Rule Effect (ppb)	32	10	35	44
Overall Estimate of Rule Effects	23 – 36 ppb			
Average Baltimore Ozone	57 ppb			

In Table 4-6 Rule 2 (cool, cloudy days with higher rural ozone concentrations) shows the smallest estimates of the rule effects (8-10 ppb), and Rule 4 (higher rural ozone concentrations on clearer days) shows the largest estimates (29-44 ppb). The overall estimate of rule effects (23-36 ppb) was calculated from the averages of the lower and upper estimates of the rule effects, weighted based on the count of days that fall within the rule. The overall estimate suggests that 40 to 64 percent of the 8-hour ozone concentrations at Baltimore can be attributed to regional effects rather than localized effects that influence only the Baltimore area.

Table 4-7. Calculation of Cubist® rule effects on Washington ozone concentrations

	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Rule 7
Washington Ozone Average (ppb)	34.9	42.6	46.4	55.0	63.1	71.8	83.9
Standard Deviation (ppb)	13	13	12	12	13	14	14
Count	409	84	420	70	425	104	58
Methodist Hill Ozone Average (ppb)	37.6	53.1	34.7	54.8	56.0	65.9	71.7
Standard Deviation (ppb)	13	7	6	8	10	12	13
Rule Effect (coefficient)	0.22	0.77	0.52	0.12	0.28	0.52	0.56
Shenandoah Ozone Average (ppb)	43.8	56.2	43.9	59.5	58.3	67.2	73.1
Standard Deviation (ppb)	15	14	10	11	11	14	12
Rule Effect (coefficient)	0.3	0	0.31	0.08	0.09	0.05	0.05
Lower Estimate of Rule Effect (ppb)	14	36	25	10	17	31	36
Upper Estimate of Rule Effect (ppb)	29	46	38	13	25	44	52
Overall Estimate of Rule Effects	21-32 ppb						
Average Washington Ozone	53 ppb						

Table 4-7 shows the Cubist rule effects calculated for the Washington ozone data. Seven rules were generated, and Rule 6 was concerned with high Washington ozone concentrations on the previous day (over 76 ppb) from Clusters 0 and 4 (upper air winds from the west). Rule 7 dealt with Clusters 2 and 3 (low wind speeds), high nighttime ozone concentrations at Methodist Hill (over 43 ppb), and high ozone concentrations on the previous day (over 76 ppb). Rules 6 and 7 both give estimates of the rule effect in the range of 43 to 62 percent, just slightly higher than the overall estimate for Washington of 39 to 60 percent.

It was surprising that Washington rules were more dependent on Methodist Hill concentrations than on Shenandoah concentrations, but Table 4-1 shows that both Washington and Baltimore sites were better correlated with Methodist Hill than Shenandoah. The Methodist Hill site is located 500 meters lower than Shenandoah and may therefore measure an air mass more similar to ones in Baltimore and Washington.

These estimates of rule effects are indicators of a regional component of ozone concentrations in Baltimore and Washington, but it cannot be assumed that the regional component from Methodist Hill and Shenandoah is necessarily transported from other areas.

A classifier model in the Weka software (M5 Rules) yielded similar results:

=== Run information ===

Scheme: weka.classifiers.rules.M5Rules -M 4.0
Relation: mastercluster092605-weka.filters.unsupervised.attribute.Remove-R1,33_clustered-weka.filters.unsupervised.attribute.Remove-R1-2,5-6,10-11,13-25,27-33,35-37,39-41,43-45-weka.filters.unsupervised.attribute.Remove-R8-10
Instances: 2448
Attributes: 8
Yearo
DOW
Wash8AvgO3
WashPreDay
MHAvgNightO3
ShenAvgNightO3
Duration
Cluster
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

M5 pruned model rules
(using smoothed linear models) :
Number of Rules : 3

Rule: 1
IF
WashPreDay > 48.5
THEN
Wash8AvgO3 =
-0.3525 * Yearo
+ 0.0023 * DOW
+ 0.3005 * WashPreDay
+ 0.2693 * MHAvgNightO3
+ 0.3049 * ShenAvgNightO3
+ 12.6154 * Cluster=cluster4,cluster0,cluster3,cluster2
+ 5.2328 * Cluster=cluster0,cluster3,cluster2
+ 2.8277 * Cluster=cluster3,cluster2
+ 1.6478 * Cluster=cluster2
+ 697.4301 [1264/80.113%]

Rule: 2
IF
Cluster=cluster4,cluster0,cluster3,cluster2 > 0.5
THEN
Wash8AvgO3 =
-0.2022 * Yearo
+ 0.784 * DOW
+ 0.2622 * WashPreDay
+ 0.3247 * MHAvgNightO3
+ 0.3429 * ShenAvgNightO3
+ 0.2776 * Cluster=cluster4,cluster0,cluster3,cluster2
+ 1.747 * Cluster=cluster3,cluster2
+ 408.5626 [573/94.016%]

Rule: 3
Wash8AvgO3 =

0.2654 * WashPreDay
+ 0.3341 * ShenAvgNightO3
+ 9.5777 [252/109.257%]

Time taken to build model: 3.89 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.7734
Mean absolute error	9.0259
Root mean squared error	11.58
Relative absolute error	60.9238 %
Root relative squared error	63.345 %
Total Number of Instances	2089
Ignored Class Unknown Instances	359

The output above assumes that true conditions (e.g., "Cluster=cluster3,cluster2") have a value of 1 and false conditions a value of zero. Table 4-8 shows a table representing the rule effects of the above model. The overall estimate of rule effects for the M5 Rules algorithm at Washington (21-35 ppb) showed similar effects as those from the Cubist software (21-32 ppb), despite the fact that classifier and association rules arrive at models using different algorithms. Tests on the Baltimore data also found similarity between the overall estimates for the M5 Rules and Cubist algorithms.

Table 4-8. Calculation of M5 rule effects on Washington ozone concentrations

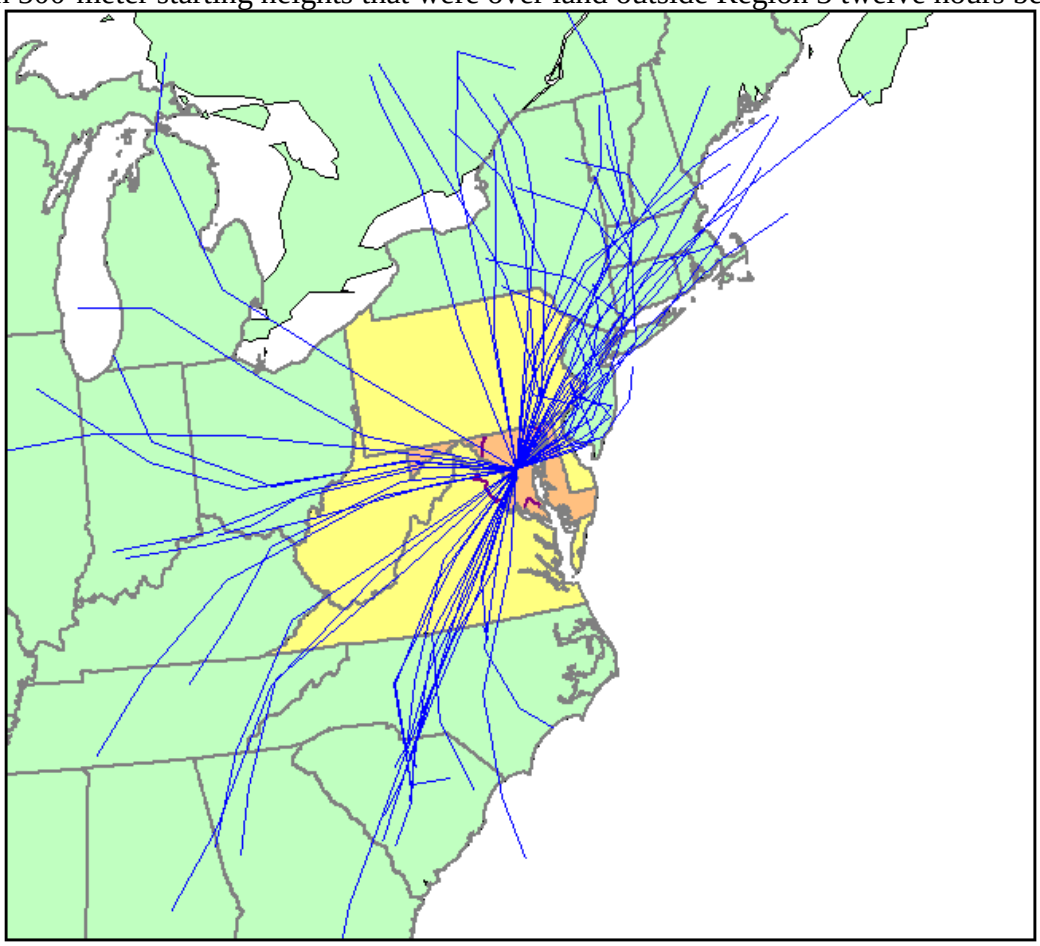
	Rule 1	Rule 2	Rule 3
Washington Ozone Average (ppb)	60.8	47.1	31.1
Standard Deviation (ppb)	17	13	11
Count	1257	573	252
Methodist Hill Ozone Average (ppb)	53.5	38.2	31.8
Standard Deviation (ppb)	14	11	10.5
Rule Effect (coefficient)	0.2693	0.3247	0
Shenandoah Ozone Average (ppb)	57.7	43.3	37.7
Standard Deviation (ppb)	12.8	10	12
Rule Effect (coefficient)	0.3049	0.3429	0.3341
Lower Estimate of Rule Effect (ppb)	24	20	9
Upper Estimate of Rule Effect (ppb)	40	34	17
Overall Estimate of Rule Effects	21-35 ppb		
Average Washington Ozone	53 ppb		

4.6 HYSPLIT Analysis

The HYSPLIT data offers geospatial information about air parcel locations at earlier times. The MMOD tracks the air parcel locations and elevation above ground level at times 6, 12, 18, and 24 hours prior to 4 PM on a given day. The exact geospatial information (latitude and longitude)

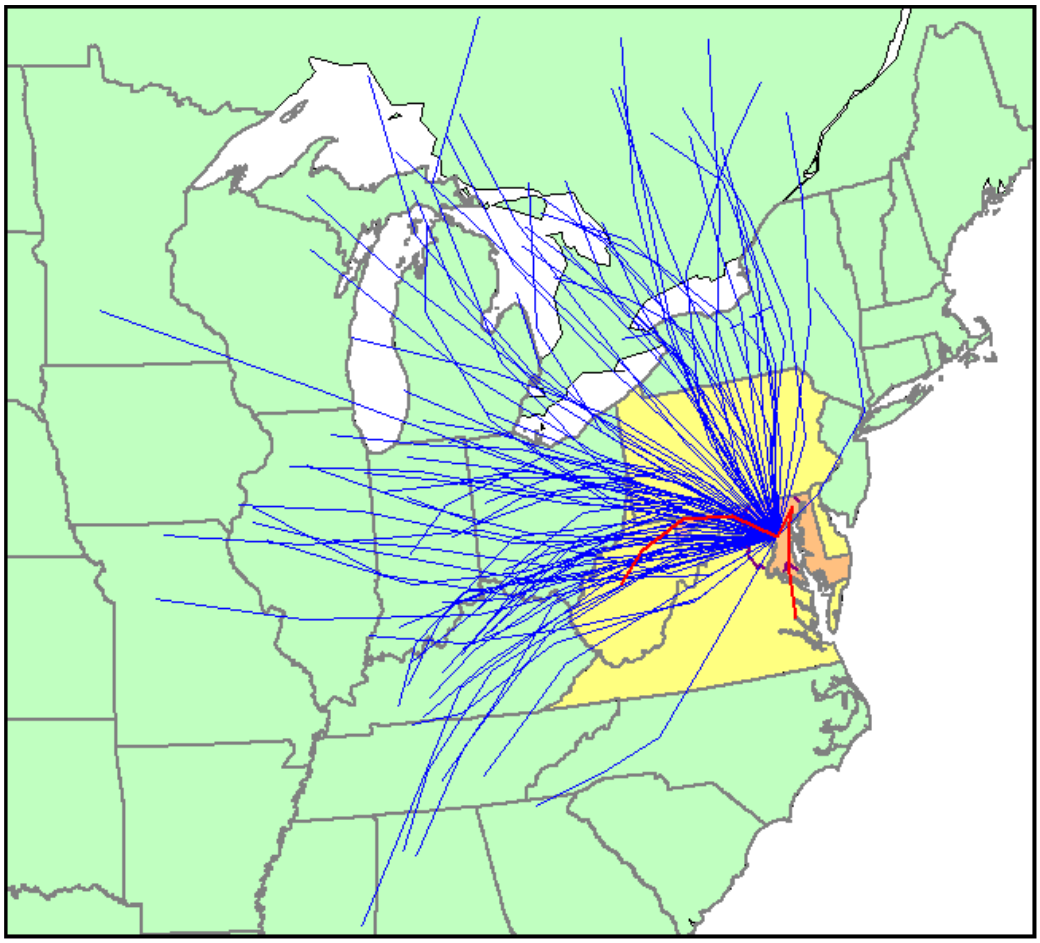
would be difficult for existing data mining software to interpret; the field of geospatial data mining is still in its infancy. Instead, each of the HYSPLIT points was assigned a classifier to indicate if it fell within the Nonattainment areas, within Maryland, within Region 3, over Other land outside Region 3, or over Water. Figure 4-7 illustrates the trajectories from the Baltimore Cluster 1 data that had air parcels with 500-meter starting heights that were over other land at 4 AM (twelve hours before the trajectory began). Note that the Cluster 1 data points over other land often were located in New York, New Jersey, or the Carolinas twelve hours earlier.

Figure 4-7. Back trajectories (24-hour) from the Baltimore Cluster 1 data that had air parcels with 500-meter starting heights that were over land outside Region 3 twelve hours before 4 PM



In blue Figure 4-8 shows Baltimore Cluster 4 data that HYSPLIT indicates the parcel beginning at 500 meters will be over land outside Region 3 twelve hours previous to the start, and in red are shown the two parcels that are over Maryland and/or the Baltimore and Washington nonattainment areas. Figures 4-7 and 4-8 illustrate the utility of the clustering for describing the data but also show that the clusters (calculated based on surface and upper air meteorology) are already sorted partially based on the wind directions. A quick comparison of the trajectories in Figure 4-8 shows that Cluster 4 contains air parcels that travel anywhere from 100 to 1000 miles in just one day.

Figure 4-8. Back trajectories (24-hour) with 500-meter starting heights from the Baltimore Cluster 4 data. Trajectories shown in blue had air parcels that were over land outside Region 3 twelve hours before 4 PM, and the two red trajectories had air parcels that were over Maryland and/or the Baltimore and Washington nonattainment areas twelve hours before 4 PM

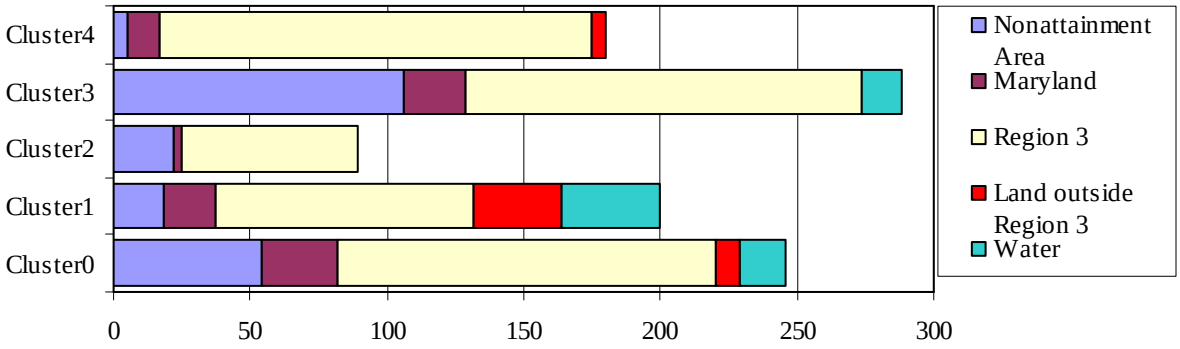


Note in Figure 4-8 that only two trajectories remain within the state over twelve hours. To perform a worthwhile statistical comparison between transported parcels versus local ones with the HYSPLIT data, it was important to find HYSPLIT data fields that included a reasonable number of instances with both near and far transport. The fact that only two instances were available in Cluster 4 indicates significant uncertainty in the estimates. Figures 4-9 and 4-10 illustrate the number of instances found for each cluster that were characterized by HYSPLIT as being located over the different areas (with 500-meter and 1000-meter starting heights).

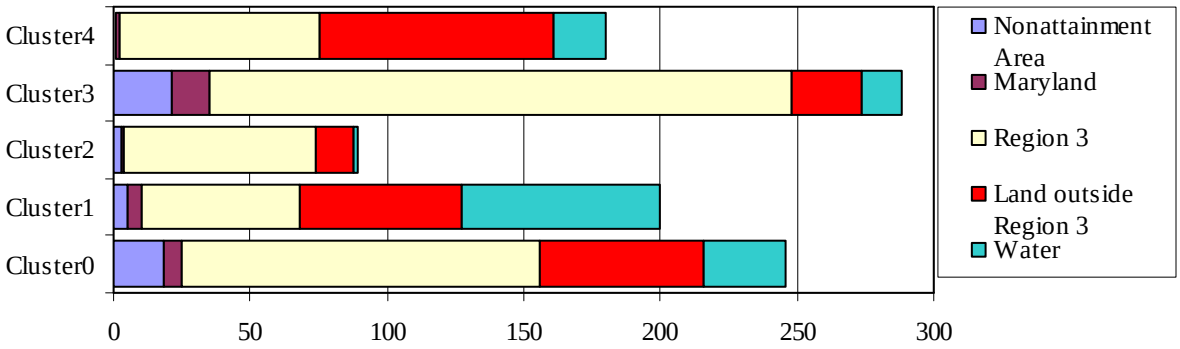
Figures 4-9 and 4-10 (as well as data at the 1500-meter starting height) were used to evaluate which times before start and which starting heights could be used in an effort to divide each cluster into both local and transport subclusters. This evaluation focussed on the number of instances within the nonattainment area or Maryland for the local subcluster and the land outside Region 3 for the transported cluster. The areas over water were not considered areas from where pollutants would be transported since most trajectories were over the Atlantic Ocean. The best data sets to work with appeared to be those for 4 AM, twelve hours before the back trajectory

Figure 4-9. Locations of HYSPLIT back trajectories from 500-meter starting heights for the Baltimore clusters

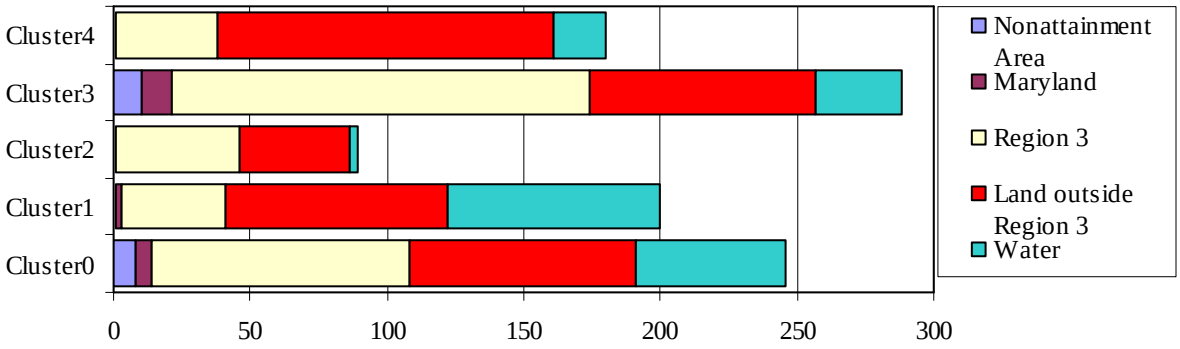
a) Six hours before the 4 PM start.



b) 12 hours before start.



c) 18 hours before start.



d) 24 hours before start.

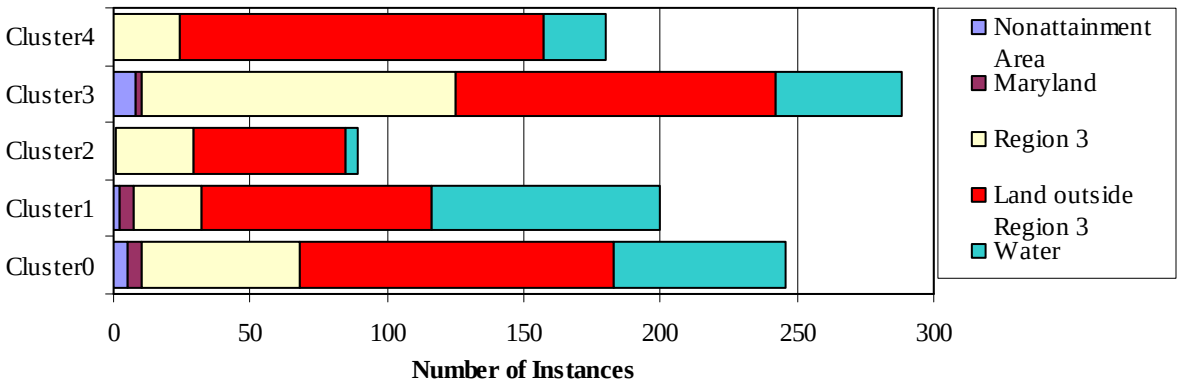
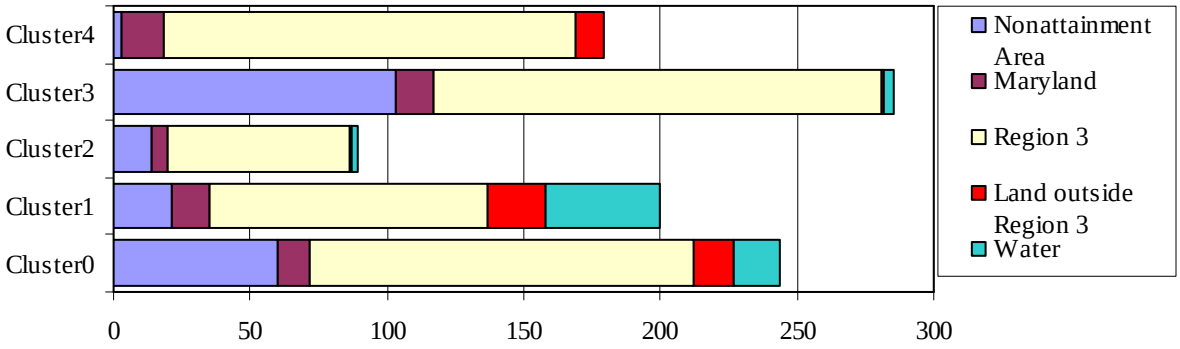
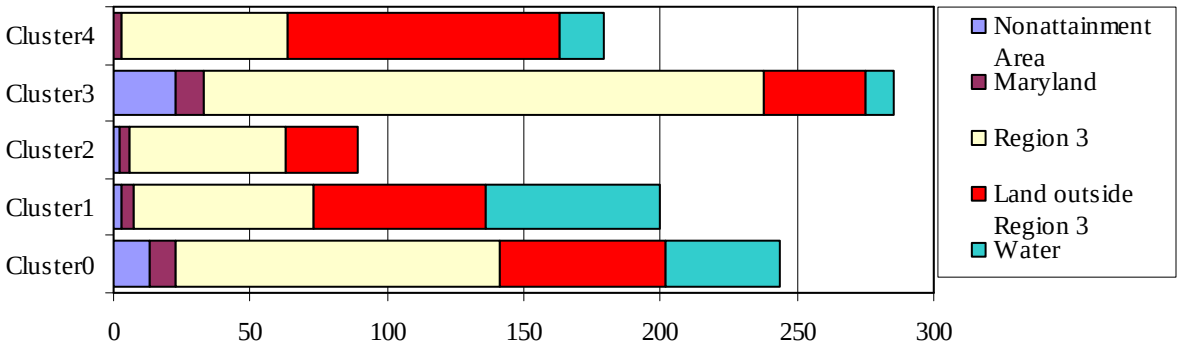


Figure 4-10. Locations of HYSPLIT back trajectories from 1000-meter starting heights for the Baltimore clusters

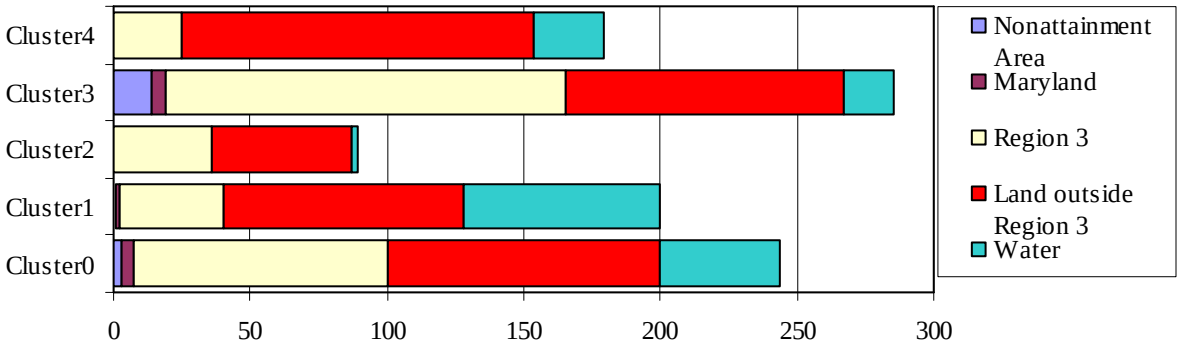
a) Six hours before the 4 PM start.



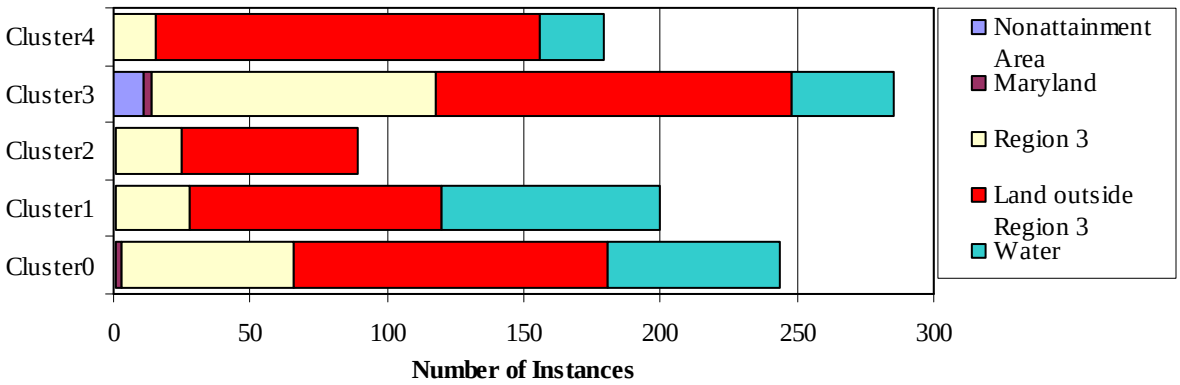
b) 12 hours before start.



c) 18 hours before start.



d) 24 hours before start.



began, corresponding to Figures 4-9b and 4-10b. The nonattainment area/Maryland instances contained 72 to 82 values at the three heights, and the land outside Region 3 instances contained 245 to 313 values.

Table 4-9 shows a comparison of the Baltimore average 8-hour ozone data for the V4loc attribute (location at 4 AM for the 500-meter air parcel) in the MMOD. All five clusters showed a negative effect on the ozone concentration if the air parcel had come from outside Region 3 twelve hours earlier. The HYSPLIT back trajectories that were located over the nonattainment areas and/or Maryland twelve hours earlier were represented in all five clusters and are shown in Figure 4-11. These back trajectories represent fairly stagnant conditions with the 500-meter winds traveling at a maximum values under 7 m/s averaged for the twenty-four hour period. These trajectories were still over the Baltimore and Washington nonattainment areas at 4 AM, so their pollutant concentrations would likely have been affected by the morning rush hours in Baltimore-Washington on weekdays. Therefore, the finding that air parcels that had traveled longer distances had lower ozone concentrations was not unexpected. The higher wind speeds are also associated with greater dispersal of pollutants.

In particular the Cluster 4 data in Table 4-10 suggest that days with transport have very high ozone data. However, only two instances are represented, and one of those days (July 19, 1999) measured 100 ppb ozone. Although some meteorological attributes (e.g., temperature) suggest that this instance might belong in a different cluster, the instance was likely included in Cluster 4 because of the high daily wind speed u component and the high afternoon wind speeds.

Table 4-9. Effect of transport on Baltimore ozone data based on HYSPLIT back trajectories (500-meter start, 12 hours previous)

Calculation	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	All Records
Average for trajectories over land outside Region 3 (ppb)	54.4	40.0	52.4	58.5	56.7	
Std Deviation (ppb)	14	15	9	18	16	
Number of records	*60*	*59*	*14*	*26*	*86*	
Average for trajectories over Region 3 (ppb)	64.4	39.5	57.3	63.4	66.3	
Std Deviation (ppb)	18	12	16	19	17	
Number of records	*131*	*58*	*70*	*213*	*73*	
Average for trajectories over Maryland/nonattainment areas (ppb)	67.5	40.1	59.8	68.7	80.0	
Std Deviation (ppb)	17	14	16	19	28	
Number of records	*25*	*10*	*4*	*35*	*2*	
Average transport effect comparing land outside Region 3 with Maryland/nonattainment (ppb)	**-13**	**-0.1**	**-7.4**	**-10**	**-23**	**-11**
T-test probability	0.007	0.86	0.34	0.03	0.42	
Total Count	*544*	*464*	*178*	*760*	*497*	*2443*

Figure 4-11. Back trajectories (24-hour) from the Baltimore cluster data that had air parcels with 500-meter starting heights over Maryland or the nonattainment areas twelve hours before 4 PM

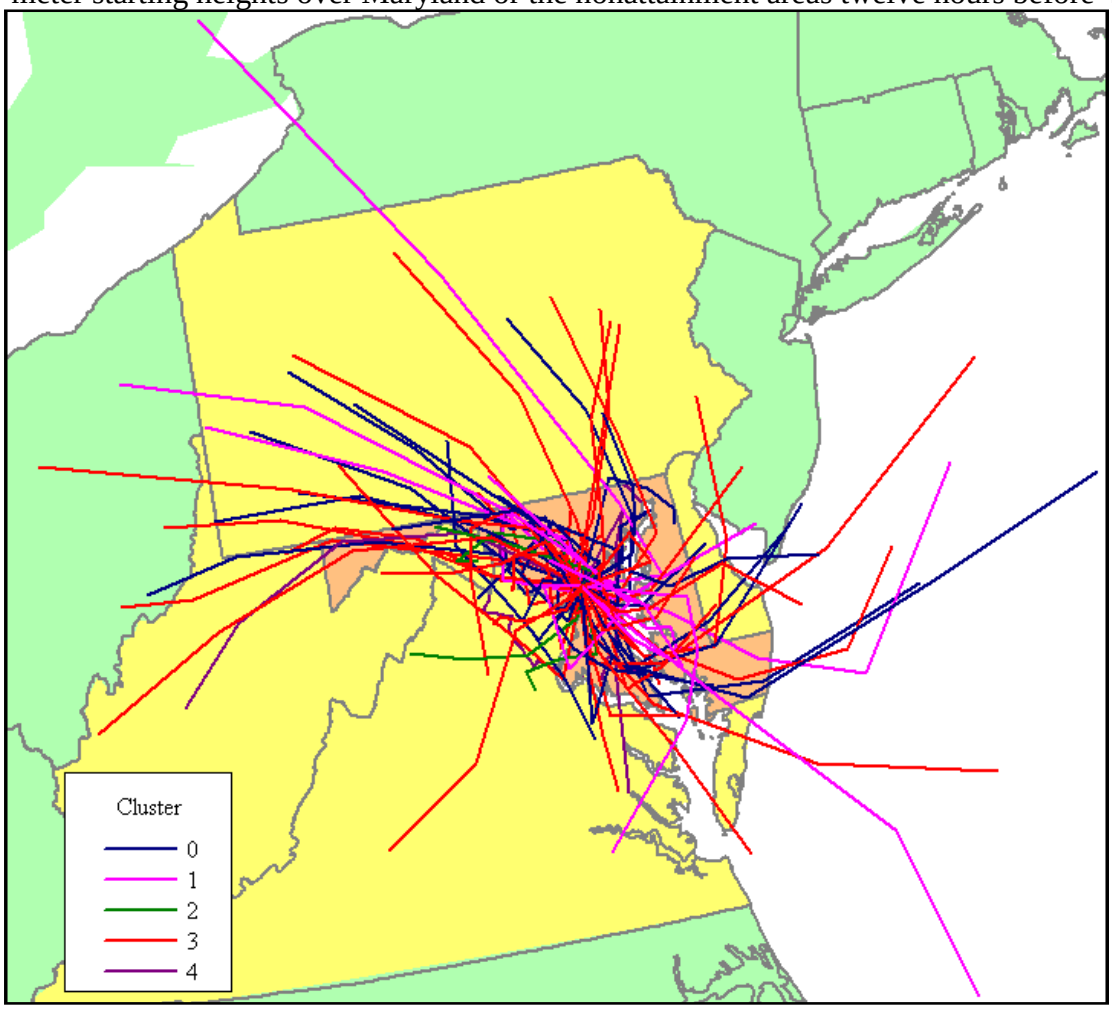


Table 4-9 also shows that back trajectories that were over Region 3 (but not the nonattainment areas or Maryland) were associated with lower ozone concentrations than those transported short distances and higher ozone concentrations than those transported from outside the region. Table 4-10 presents the HYSPLIT differences that were calculated for the other trajectory heights and times. Table 4-11 shows that many of the data subclusters for Washington, DC also point to lower ozone concentrations for those air parcels that are transported from outside Region 3. Table 4-10 has 38 negative values, 2 zero values, and 12 positive values, and Table 4-11 includes 44 negative values, 2 zero values, and 8 positive values. The presence of the positive values may be due to uncertainty in the modeling but might also indicate cases where long range transport of pollutants is more important than the stagnation processes in the Baltimore-Washington corridor.

Table 4-10. Average difference in Baltimore 8-hr ozone concentrations between air parcels over land outside Region 3 and those within the Baltimore/Washington nonattainment areas and Maryland

Hours prior to start time	Starting height (m)	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
6	500	-22 ppb	+2 ppb	- *	- *	-27 ppb
6	1000	-15 ppb	+2 ppb	-2 ppb	+14 ppb	-21 ppb
6	1500	-11 ppb	+7 ppb	-14 ppb	-16 ppb	-32 ppb
12	500	-13 ppb	0 ppb	-7 ppb	-10 ppb	-23 ppb
12	1000	-8 ppb	+10 ppb	-9 ppb	-9 ppb	-36 ppb
12	1500	-15 ppb	+11 ppb	-10 ppb	-11 ppb	-5 ppb
18	500	-12 ppb	-1 ppb	+8 ppb	-8 ppb	-1 ppb
18	1000	-9 ppb	+8 ppb	- *	-5 ppb	- *
18	1500	-16 ppb	+7 ppb	-10 ppb	-3 ppb	- *
24	500	-19 ppb	+2 ppb	+7 ppb	-10 ppb	- *
24	1000	-1 ppb	+17 ppb	-27 ppb	-6 ppb	- *
24	1500	-10 ppb	0 ppb	-19 ppb	-6 ppb	- *

* Not enough instances available for calculation

Table 4-11. Average difference in Washington 8-hr ozone concentrations between air parcels over land outside Region 3 and those within the Baltimore/Washington nonattainment areas and Maryland

Hours prior to start time	Starting height (m)	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
6	500	-19 ppb	-1 ppb	-25 ppb	- *	- *
6	1000	-13 ppb	0 ppb	-19 ppb	+20 ppb	-5 ppb
6	1500	-25 ppb	+6 ppb	-13 ppb	-14 ppb	-12 ppb
12	500	-20 ppb	0 ppb	-13 ppb	-11 ppb	-9 ppb
12	1000	-25 ppb	+3 ppb	-6 ppb	-11 ppb	-13 ppb
12	1500	-24 ppb	+10 ppb	-13 ppb	-10 ppb	-9 ppb
18	500	-6 ppb	-3 ppb	-10 ppb	-8 ppb	-1 ppb
18	1000	- *	+10 ppb	-10 ppb	-4 ppb	- *
18	1500	-17 ppb	+4 ppb	-12 ppb	-2 ppb	-6 ppb
24	500	-13 ppb	+2 ppb	-15 ppb	-12 ppb	-3 ppb
24	1000	- *	+12 ppb	-10 ppb	-3 ppb	-26 ppb
24	1500	-9 ppb	-2 ppb	-9 ppb	-2 ppb	-17 ppb

* Not enough instances available for calculation

The behavior of these individual subclusters may be an area for future work because unexpected behavior can be investigated. For example, Baltimore Cluster 1 shows the most positive values in Tables 4-10 and 4-11. Its subcluster remaining in the nonattainment areas/Maryland has 8 out of 10 readings on Wednesdays and Thursdays even though its remaining subclusters (outside Maryland and the nonattainment areas) are evenly distributed among the days of the week.

The HYSPLIT attributes and the ones used to form the clusters (surface and upper air meteorology) are both based primarily on meteorological conditions. The average geopotential height for the upper air soundings at IAD was 1500 meters, so the same air mass is being tracked in both situations. Plots of the upper air data versus the HYSPLIT parameters showed the close correlations. Because the attributes were not independent of meteorology, it could be expected that the HYSPLIT modeling would not yield the clean findings that were found when examining the aberrant phenomenon of the persistent LLJ. Ryan[8] observed that the LLJs had maximum heights below 500 meters.

Runs were also performed using the HYSPLIT data with the Cubist and M5 Rules algorithms. However, these models continued to base their predictions primarily on the meteorological attributes and not those associated with HYSPLIT. Further conclusions about the utility of HYSPLIT for source attribution might be made if the MMOD data is divided into subsets and examined with geographic or statistical plots in future studies.

4.7 Uncertainties

Section 4.5 suggests a strong regional relationship between the nighttime ozone averages at Methodist Hill and Shenandoah sites with the 8-hour maximum averages for Baltimore and Washington. A strong relationship is expected both because the sites experience similar meteorological conditions on any given day and because the air parcel containing the pollutants is shared on a regional basis. However, this regionality should be quantified before it is assumed. After all, a data mining exercise such as the one described in Section 4.5 may have shown strong relationships between Baltimore and Chicago, but one would not assume that the same air parcel is experienced on the same days.

Vukovich *et al.*[23] presented an analysis that showed that Baltimore and Washington ozone data can be broken into separate time scales that include a long-term mean and interannual, intra-annual, and synoptic perturbations. The synoptic perturbation component is calculated by finding the difference between the monthly mean value and the sub-weekly sample at time *t* as follows:

$$C'(t) = C(t) - C_{month}$$

where C'(t) represents the monthly-average synoptic perturbation value for time *t*, C(t) represents the raw data sample at time *t*, and C_{month} represents the average monthly concentrations for the same year of the sub-weekly data sample. The synoptic perturbations may be viewed as the "high frequency" variability around the monthly values. Because the sum of the synoptic perturbations $\Sigma C'(t)$ is equal to zero for each month, the synoptic component from daily variations is excluded from the monthly averages and the seasonal trends in pollutant concentrations. Table 4-12 shows the monthly-average synoptic correlations of the 8-hour maxima among the individual Baltimore ozone monitors and also how they correlated with the nighttime ozone concentrations at Methodist Hill and Shenandoah (monitors pictured in Figure 4-12). The 8-hour maxima show only moderate synoptic correlations among the monitors, but

the moderate correlation coefficients among the synoptic perturbations suggest that the regionality is not strong, even among monitors located in the same counties.

Table 4-12. Correlation coefficients for synoptic perturbations among the Baltimore nonattainment area and rural ozone monitors

Site Location	Monitor ID	240030014	240030019	240051007	240053001	240130001	240251001	240259001	Methodist Hill	Shenandoah
Queen Anne and Wayson	240030014	1	0.89	0.81	0.88	0.77	0.87	0.85	0.63	0.53
Fort Meade	240030019	0.89	1	0.88	0.89	0.84	0.89	0.89	0.63	0.53
Greenside Drive, Cockeysville	240051007	0.81	0.88	1	0.86	0.89	0.87	0.89	0.66	0.51
Essex	240053001	0.88	0.89	0.86	1	0.79	0.91	0.89	0.65	0.51
Old Liberty Road, Winfield	240130001	0.77	0.84	0.89	0.79	1	0.82	0.84	0.67	0.53
Edgewood	240251001	0.87	0.89	0.87	0.91	0.82	1	0.94	0.66	0.52
Aldino	240259001	0.85	0.89	0.89	0.89	0.84	0.94	1	0.67	0.54
Methodist Hill		0.63	0.63	0.66	0.65	0.67	0.66	0.67	1	0.75
Shenandoah		0.53	0.53	0.51	0.51	0.53	0.52	0.54	0.75	1

For this study, the high frequency variability was also characterized as the perturbation from the cluster averages:

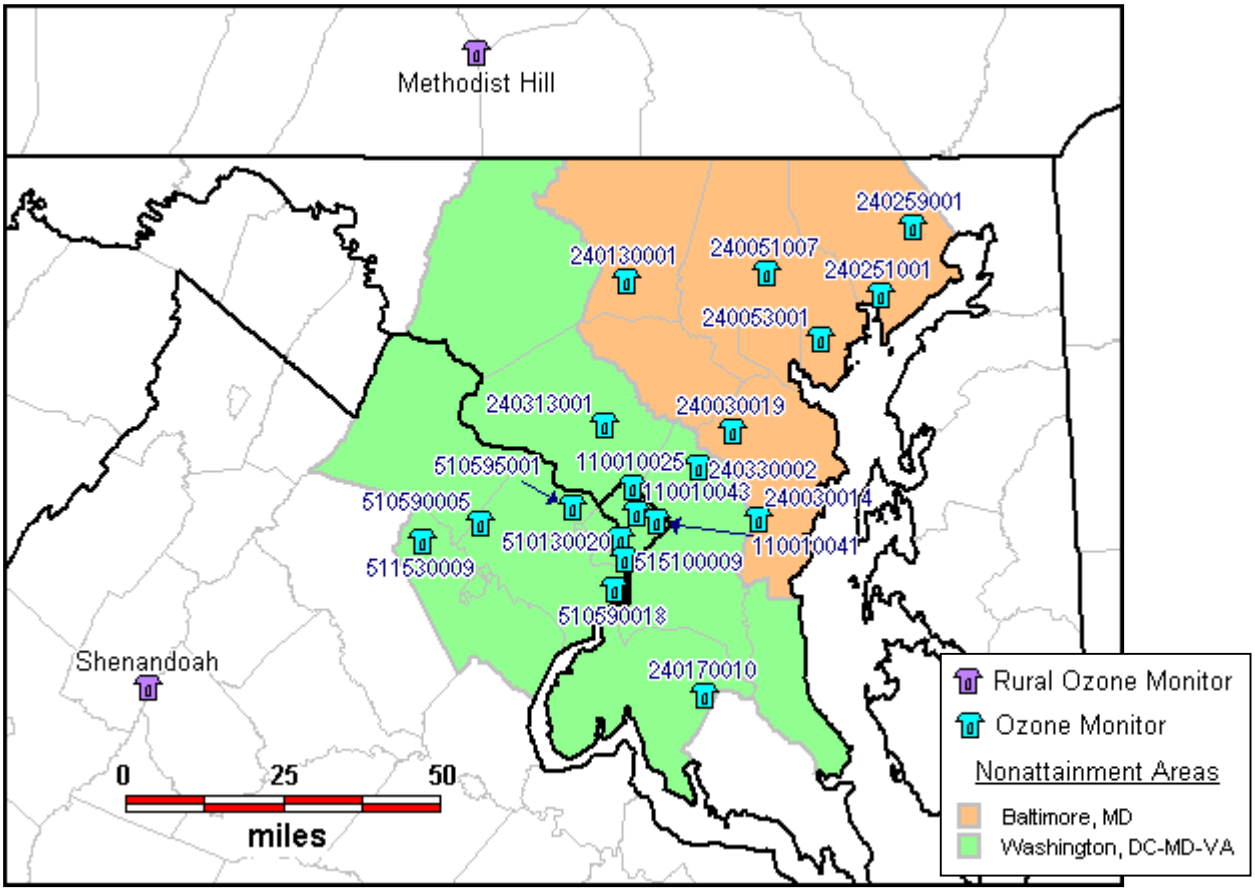
$$C''(t) = C(t) - C_{cluster}$$

where C”(t) represents the cluster-average synoptic perturbation, and $C_{cluster}$ the average concentration for one of the five clusters defined for Baltimore or for Washington.

Comparing the synoptic perturbations among the different ozone data fields can be useful for quantifying the regional structure and variations in ozone on short time scales. Ozone shows strong meteorological and seasonal patterns, which the clustering aims to capture. Thus, similarities in daily effects on the ozone data from different locations will often be due to strong meteorological/seasonal/clustering factors that affect ozone formation and destruction. To avoid these meteorological/seasonal/clustering effects, synoptic perturbations were compared to determine if significant homogeneity exists in ozone levels among the different ozone fields on synoptic scales.

The ranges in the clustered ozone values were very high for all four data sets: Baltimore, Washington, Shenandoah, and Methodist Hill. The cluster ranges were between 63 and 107 ppb, and so some cluster perturbations were over 60 ppb. This variance might be reduced in future work if more clusters are chosen or if more than a single ozone attribute is introduced to the clusterer (only the previous day’s ozone concentrations were used in the clustering).

Figure 4-12. Locations of Baltimore and Washington, DC area ozone monitors



The cluster-average synoptic correlations between the nonattainment area and rural monitors are shown in Table 4-13. These correlation coefficients can be considered a measure of how well the rural data can be used to predict the nonattainment area concentrations. The correlation coefficients ranged from 0.52 for Washington-Methodist Hill's Cluster 1 (cloudy and cool with winds from east and northeast) to 0.77 for Washington-Methodist Hill's Cluster 0 (sunny and hot with fast, steady winds from west). Figure 4-13 shows that Cluster 0 trajectories are predominantly from the west and northwest, so it is not surprising that the Methodist Hill nighttime values would correlate well with the observed ozone later in the day.

Cluster 1 for the Washington data was the only case where Shenandoah was better correlated than Methodist Hill, and the other cases may correlate better with Methodist Hill than Shenandoah because many of the winds come from the northwest and also because the site is closer in elevation to the nonattainment areas.

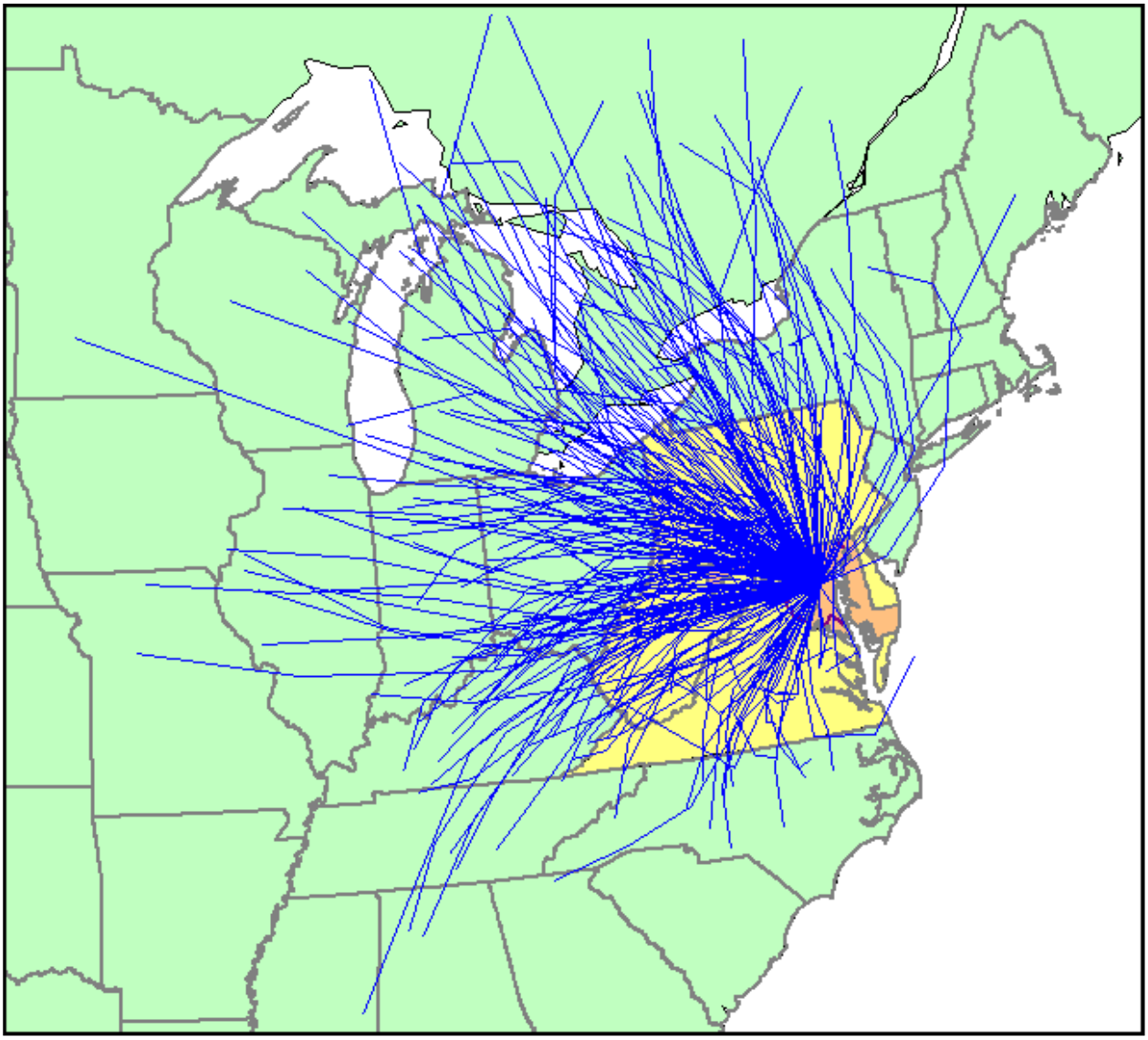
These cluster perturbation correlation coefficients are quite similar to those for the monthly perturbations in Table 4-12. Therefore, the clusters seem to adequately reflect the uncertainty associated with the spatial differences in the monitor readings.

Table 4-13. Synoptic correlations between nonattainment area and rural monitor data

Site Pair	Parameter	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Baltimore – Methodist Hill	Correl. coef.	0.74	0.57	0.54	0.76	0.73
	Slope	*0.79*	*0.53*	*0.55*	*0.91*	*0.81*
	Intercept (ppb)	*-1.5*	*0.05*	*0.33*	*-0.15*	*-1.1*
Baltimore – Shenandoah	Correl. coef.	0.55	0.53	0.54	0.54	0.66
	Slope	*0.64*	*0.48*	*0.61*	*0.72*	*0.82*
	Intercept (ppb)	*0.16*	*0.09*	*0.19*	*0.05*	*0.17*
Washington – Methodist Hill	Correl. coef.	0.77	0.52	0.71	0.69	0.60
	Slope	*0.78*	*0.47*	*0.74*	*0.82*	*0.64*
	Intercept (ppb)	*-10*	*-23*	*-9.5*	*-10*	*-13*
Washington - Shenandoah	Correl. coef.	0.71	0.57	0.58	0.54	0.57
	Slope	*0.78*	*0.51*	*0.67*	*0.73*	*0.60*
	Intercept (ppb)	*5.6*	*-14*	*7.9*	*7.1*	*-0.3*

A previous study[24] examined the correlations between Washington and Shenandoah IMPROVE monitor sites that speciated particulate matter concentrations for particles with aerodynamic diameters less than 2.5 microns in diameter. The synoptic correlation coefficients for sulfate, organic carbon, and elemental carbon particulate matter were respectively 0.74, 0.47, and 0.43 in that study for the monitors eighty miles apart. Sulfate particulate is widely regarded as a regional pollutant whose effects spread across several states. Since the ozone synoptic correlations in Table 4-13 are comparable in the best clusters, this suggests a certain regional influence (when those cluster conditions are met) between the nighttime concentrations at the rural sites and the daily 8-hour maximum averages in the Baltimore and Washington nonattainment areas. Note that this correlation is for the average of the 8-hour maxima among the monitors in the area and may not reflect the behavior at the site measuring the peak value.

Figure 4-13. Back trajectories (24-hour) from the Washington Cluster 1 data that had air parcels beginning at 500 meters and 4 PM.



5. Conclusions

The initial work in this study entailed combining long-term data sets to produce the Maryland Meteorology and Ozone Dataset (MMOD). The daily database covers the years 1989 through 2004 during the months of May through September. The MMOD contains more than 150 measured and derived attributes that consider surface and 850-mb meteorology, detection of LLJs, back trajectory analyses, and ozone concentrations in and around the Baltimore-Washington area.

Data mining tools were used to investigate the MMOD to determine the ozone concentrations that could be associated with transport events. The EM clusterer served as the best tool for clustering the dataset based on surface meteorology, upper air soundings, and the previous day's ozone concentrations. The tool clustered the Baltimore data into five meteorological data sets that can roughly be described as follows:

- **Cluster 0** (544 records) - Sunny, variable winds, and a higher temperature difference between upper air and surface conditions
- **Cluster 1** (464 records) - Cloudy, cool days with winds from east and northeast and the most precipitation
- **Cluster 2** (178 records) - Hot and humid with upper air winds from west and moderate precipitation
- **Cluster 3** (760 records) - Low wind speeds, limited clouds and little precipitation
- **Cluster 4** (497 records) - High wind speeds with little precipitation [surface winds from west, upper winds from northwest]

Similarly the Washington data was clustered into five meteorological data sets:

- **Cluster 0** (606 records) – Sunny, hot days with higher-speed surface and aloft winds from west
- **Cluster 1** (484 records) - Cloudy, cool days with winds from east and northeast, most precipitation, high morning wind speeds, and low wind variability
- **Cluster 2** (447 records) – Sunny with limited precipitation and high temperature differences between surface and aloft; highly variable low surface wind speeds with upper winds from the north
- **Cluster 3** (695 records) - Low wind speeds from the west with limited clouds and precipitation
- **Cluster 4** (216 records) - High temperatures with moderate clouds, low-speed variables winds from the south, upper winds from the west, and moderate precipitation

The clusters were subdivided into those with and those without measured persistent LLJs. The ozone concentrations in Baltimore and Washington nonattainment areas were statistically different for Clusters 0, 2, 3, and 4 depending on whether or not a persistent LLJ was observed. On average, the presence of a LLJ increased the Baltimore 8-hour maximum ozone concentration by 7 ppb and the Washington concentration by 5 ppb.

To measure the regional nature of ozone pollution, the Baltimore and Washington 8-hour maximum averages were compared to nighttime average ozone concentrations at the Methodist Hill and Shenandoah elevated/rural sites. Using the clusters described above, association rule and classifier tools generated predictive models to evaluate the relationship between rural and nonattainment area concentrations. The models from the Cubist software predicted that 40-64 percent in the Baltimore ozone concentrations could be considered regional and 39-60 percent in Washington ozone concentrations. The M5 Rules classifier algorithm predicted roughly the same percentages. The assumption of regionality between the nighttime ozone concentrations at rural sites with the 8-hour ozone maxima the next day (R^2 of 0.52-0.77 for clusters) was not as good as the regionality assumption for sulfate particulate matter in a previous study (R^2 of 0.74) but better than the correlations for organic and elemental carbonaceous particulate (R^2 of 0.47 and 0.43).

Attributes based on the back trajectories from the HYSPLIT model were also used in the data mining exercises to determine the differences when air parcels came from longer distances. However, these attributes were not independent of the basic measured wind speed and direction parameters. The models were unable to use the back trajectory information to distinguish ozone created based on stagnant conditions from that where significant transport of pollutants occurred.

Table 5-1 presents possible future studies that could build from the MMOD and the findings from this study. Several tasks could be combined within a single coherent study.

Table 5-1. Possible future studies

Goal	Proposed Work
Expand the MMOD to improve models	Re-evaluate with new parameters (e.g., PAMS data) or additional years of data
Develop the MMOD to account for sequential events	Choose an earlier start time for HYSPLIT back trajectories that does not overlap meteorological parameters (and possibly account for stagnation that occurs after significant westerly transport)
Better quantify the effects of LLJs on ozone	Evaluate other LLJ parameters (e.g., volumetric flow, speed, start time, or direction)
Better define and identify transport-relevant or nocturnal LLJs	If some instances can be categorized as clearly important LLJs and others as unimportant jets by experts, data mining tools can segregate the remaining instances into their most likely category based on parameters within the MMOD
Understand effects of pollution controls or source operation	Add fields that could account for operational changes such as outages at nearby facilities and investigate the effects on the ozone concentrations
Establish a method for source attribution	Conduct HYSPLIT runs at lower elevations or evaluate locations differently (e.g., by distance or proximity to large sources)
Improve model performance and descriptive abilities	Create more accurate models by using appropriate distribution functions for data

Table 5-1. Possible future studies (continued)

Goal	Proposed Work
Identify range of conditions that lead to NAAQS violations	Examine highest monitor values instead of nonattainment area averages
Assess transported ozone from HYSPLIT data	Use MMOD data with non-clustered data sets
Investigate other pollutants	Expand the MMOD to include ambient measurements of other pollutants (e.g., particulate matter) or visibility impairment

References

1. Technology Transfer Network Ozone Implementation. http://www.epa.gov/ttn/naaqs/ozone/rto/rto.html (accessed October 2005).

2. Ryan, W.F *et al.* "Pollution Transport During a Regional O_3 Episode in the Mid-Atlantic States." *J. Air & Waste Manage. Assoc*. 1998, **48**, 786-797.

3. Corsmeier, U *et al.* "Ozone Concentration Jump in the Stable Nocturnal Boundary Layer during a LLJ-Event." *Atmospheric Env.* 1997, 31, 1977-1989.

4. Reitebuch, O *et al.* "Nocturnal Secondary Ozone Concentration Maxima Analysed by Sodar Observation and Surface Measurements." *Atmospheric Env.* 2000, 34, 4315-4329.

5. Weaver, S. Diurnal Variations of Low-Level Jets over the Mid-Atlantic States as Diagnosed from Wind Profiler and Model Data. Masters Thesis, University of Maryland, College Park, Maryland.

6. Email correspondence among Regional Atmospheric Measurement Modeling and Prediction Program (RAMMPP) investigators. Castellanos, P; Stehr, J. Received September 23, 2005. russ@atmos.umd.edu.

7. Email correspondence among RAMMPP investigators. Taubman, B.F *et al.* Maryland Dept. of Environment. Received, April 17, 2005. russ@atmos.umd.edu.

8. Email correspondence among RAMMPP investigators. Choi, Y.J *et al.* Maryland Dept. of Environment. Received, August 24, 2004. russ@atmos.umd.edu.

9. Science Applications International Corporation. Discussion of CAMx and OSAT Model Runs for July 1995 NO_x SIP Call Episode using Emissions from the 2007 Base Year. Report for Maryland Department of the Environment, Air and Radiation Administration, 2001 under Contract Number MDE-99-6.0-AMA.

10. Ryan, W.F. The Low Level Jet in Maryland: Profiler Observations and Preliminary Climatology. Report for Maryland Department of the Environment, Air and Radiation Administration. 2004.

11. Verghese, S. *et al.* "Characterization of Nocturnal Jets over Philadelphia during Air-Pollution Episodes." Penn State University.

12. Help menu from WizWhy Version 2.01 Demo software, 1998. Demo data mining software available at http://www.wizsoft.com/ (last accessed October 2005).

13. Ryan, W.F. "Ozone Forecast Technical Discussion: Summary for 1998 Season," October 1998, downloaded from http://www.meto.umd.edu/~ryan/summary.htm on January 21, 2000.

14. Walsh and Sherwell. "Examining Ambient Monitoring Data with Rule Learning Techniques." Presented at Air and Waste Management Association Annual Meeting, San Diego, California, 2003.

15. Walsh, Milligan, and Sherwell. "Data Mining to Determine Local Effects of Mercury Emissions." Presented at Air and Waste Management Association Annual Meeting, Minneapolis, Minnesota, 2005.

16. Personal communication between EPA and SAIC. Data received between August 29, 2005 and September 1, 2005.

17. EPA – TTN AQS Download data files. http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm (last accessed September 2005).

18. Clean Air Markets – Data and Maps. http://cfpub.epa.gov/gdm/index.cfm?fuseaction=aciddeposition.prepackageddatasets (last accessed September 2005).

19. Integrated Global Radiosonde Archive. http://www.ncdc.noaa.gov/oa/climate/igra/index.php (last accessed September 2005).

20. Personal communication between Charles Piety and SAIC. Data received on August 8, 2005. Work is documented in Radar Wind Profiler Observations in Maryland: A Preliminary Climatology of the Low Level Jet. Report for Maryland Department of the Environment, Air and Radiation Administration. 2005.

21. NOAA ARL Real-time Environmental Applications and Display System – Dispersion Models. http://www.arl.noaa.gov/ready/hysplit4.html (last accessed September 2005).

22. Witten, I.H.; Frank, E. Data Mining Practical Machine Learning Tools and Techniques. Morgan Kaufmann: San Francisco, 2005.

23. Vukovich, F.M; Wayland, R; Sherwell, J. "Characteristics of Ozone in the Baltimore-Washington Area as Established from One-Hour Average Concentrations." *J. Air & Waste Manage. Assoc*. 1999, **49**, 794-803.

24. Walsh, K.J; Gilliland, A.B. "Regional Variations in Sulfate and Nitrate on Annual, Seasonal, Synoptic Time Scales." *J. Air & Waste Manage. Assoc*. 2001, **51**, 1339-1345.