APPENDIX AVAILABLE ON THE HEI WEB SITE

Research Report 183

Development of Statistical Methods for Multipollutant Research

Part 2. Development of Enhanced Statistical Methods for Assessing Health Effects Associated with an Unknown Number of Major Sources of Multiple Air Pollutants

E.S. Park et al.

Appendix E. Database Development and Summary Statistics for Harris County Data

Note: Appendices available only on the Web have been reviewed solely for spelling, grammar, and cross-references to the main text. They have not been formatted or fully edited by HEI.

Correspondence may be addressed to Dr. Eun Sug Park, Texas A&M Transportation Institute, The Texas A&M University System, 3135 TAMU, College Station, TX 77843-3135; e-park@tamu.edu.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83467701 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

This document was reviewed by the HEI Health Review Committee.

Appendix E: Database Development and Summary Statistics for Harris County Data

MORTALITY DATA FOR THE AREA AROUND THE CLINTON DRIVE MONITORING SITE AND FOR HARRIS COUNTY

Data Acquisition

We obtained electronic files of the mortality data for Harris County residents for the period 2000 to 2005 from the Texas Department of State Health Services Center for Health Statistics (TX DSHS). Table E.1 provides a summary of death counts by year; a total of 122,744 all-cause deaths (ICD-10 codes A00–Y89) occurred during the six-year period with approximately 20,000 deaths per year.

Table E.1. Summary of death counts from all causes (ICD-10 codes A00-Y89) for Harris County, TX (FIPS code: 44101)

Year	2000	2001	2002	2003	2004	2005	Total
Number	19,615	20,652	20,837	20,646	20,195	20,799	122,744

Source: Texas Department of State Health Services Center for Health Statistics

The electronic files included all items on the death certificates (except social security numbers), including key variables such as age, gender, race, ethnicity, education, year of death, and primary cause of death. In addition, a few more variables not on the death certificates were requested, especially geocoding-related variables, including latitude and longitude, Geographical Information System (GIS) match code, GIS location code, and geocoding accuracy, to ensure the quality of geocoding.

Data Cleaning

Geocoding (or address matching) is the process of assigning coordinates, for example commonly used latitude and longitude, to a street address (in our application, the address of the decedent’s residence at the time of death), which is the first step in using locational data for further spatial analyses. The mortality data were already geocoded by GIS staff at the TX DSHS. We received coordinates along with key geocoding-related variables (GIS match and location codes), which were used to evaluate the accuracy of assigned geocodes. We conducted additional quality checks of these geocoded records, and below is the brief summary of major steps that were followed for quality checking and subsequent data cleaning.

1) Overall address matching rate for all nonaccidental death records during the 6-year period (*n* = 111,319) was 97.1%; not-matched records (2.9%, 3,132) were excluded from subsequent spatial analyses because they require locational data (for example, the buffering-based analyses). The most common reasons for not-matched records included, “no matching streets found in directory” and “no matching segments”, which are common problems associated with reference theme (road network) data.

2) An additional 1,415 records (1.3%) were excluded because they were located outside of the study area. We found this discrepancy when we mapped the geocoded coordinates provided by the TX DSHS. We confirmed that mortality data were tabulated based on 'self-reported' county and state of residence, following the guidelines provided by the CDC-National Center for Health Statistics; thus, there may be small discrepancies between self-reported information and their actual geocoded home addresses.

3) Among those matched records for decedents residing in Harris County at the time of their death (*n* = 106,772), more than 99% were matched based on street-level accuracy. A breakdown

of the records by type of match (Manual, ZIP+4, and Street match) are presented in Table E.2. “Street” match is the most accurate geocode available, often based on house range address geocode or center of the street segment (99.4% of death records). For the remaining records, the other two methods of less accurate geocodes, “Manual” and “ZIP+4” match, were used (0.3% and 0.3% death records, respectively). The “ZIP+4” match indicates addresses matched to the centroid of the ZIP+4, while “Manual” match often involves several trials to find the best matching candidate in an interactive manner.

Table E.2. Geocoding accuracy summary

	Frequency	Percent
Manual Match	286	0.3
Street Match	106,179	99.4
ZIP+4 Match	307	0.3
Total matched	106,772	100.0

Table E.3 presents the total and mean daily number of deaths for specific cardiovascular and respiratory causes — Ischemic Heart Disease (IHD) (ICD-10 I20–I25), Acute Myocardial Infarction (MI) (ICD-10 I21), Heart Failure (ICD-10 I50), Stroke (ICD-10 I60–69), Chronic Obstructive Pulmonary Disease (COPD) (ICD-10 J40–44), and Pneumonia (ICD-10 J12–18). Also, descriptive statistics of selected potential effect modifiers — age, gender, education, race, ethnicity — for all-cause nonaccidental deaths, as well as deaths due to cardiovascular and respiratory causes are included later in Appendix E.

Table E.3. Summary statistics for all nonaccidental mortality and specific causes of cardiovascular and respiratory mortality, Harris County, Texas, 2000–2005

Mortality cause	Number	Mean daily number of deaths
Cardiovascular (I00–I99)	41,708	19.0
IHD (I20–I25)	20,370	9.3
Acute MI (I21)	7,786	3.6
Heart Failure (I50)	2,884	1.3
Stroke (I60–69)	8,129	3.7
Respiratory (J00–J98)	8,478	3.9
COPD (J40–44)	4,124	1.9
Pneumonia (J12–18)	2,335	1.1
Nonaccidental causes (A00–R99)	106,772	48.7

Subset data generation

We also created subsets of the data for decedents whose residences at the time of death were near one specific monitoring site (the Clinton Drive monitoring site [U.S. EPA monitoring site #: 48-201-1035]) located near the Houston Ship Channel, a dense petrochemical complex in Harris County. A map of geocoded residences of decedents in Harris County showing multiple buffers (5-, 10-, 20-, and 30-mile radii) around the monitoring site, are presented in Figure E.1 (the red triangle depicts the Clinton Drive monitoring site). Because the 20- and 30-mile buffers extend to regions outside of Harris County, we decided to restrict our analyses around the Clinton Monitoring station to 10-mile buffer in addition to looking at Harris County as a whole.

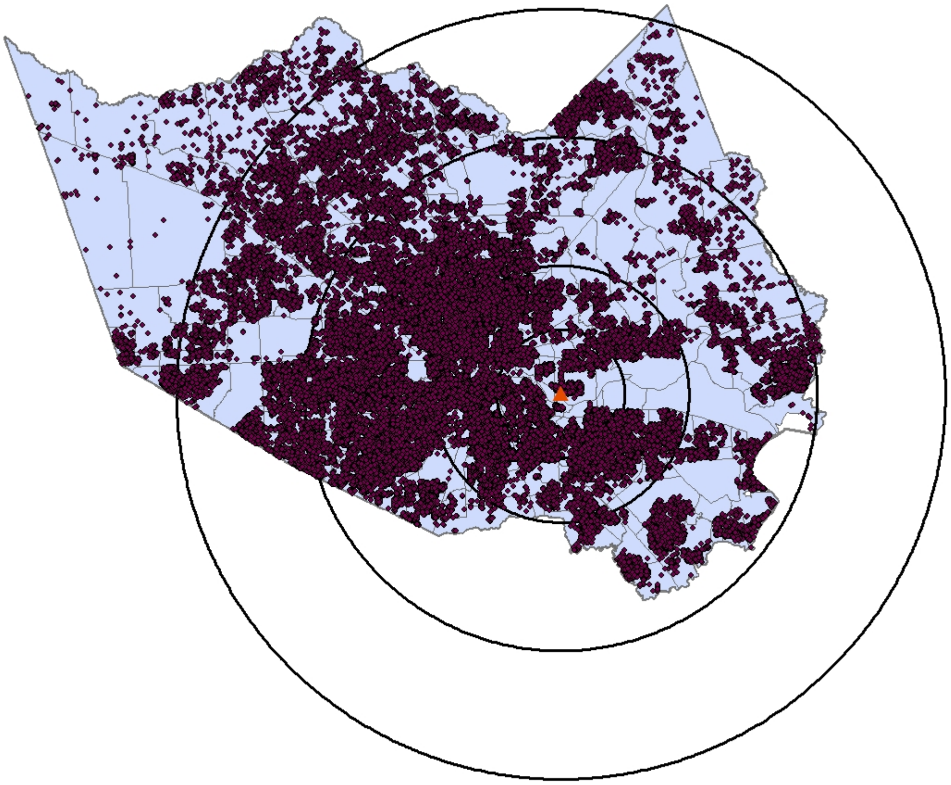


Figure E.1: Circular buffers created near the Clinton Drive monitoring site (red triangle); 5-, 10-, 20-, and 30-mile buffers from smallest to largest. Geocoded residences of decedents in Harris County, TX, 2000–2005, are also shown (purple dots).

Figure E.2 depicts the map of residences for the 10-mile buffer surrounding the Clinton Drive monitoring site. Similar to Figure E.1, the red triangle depicts the Clinton Drive monitoring site and the geocoded residences of decedents are identified by purple dots.

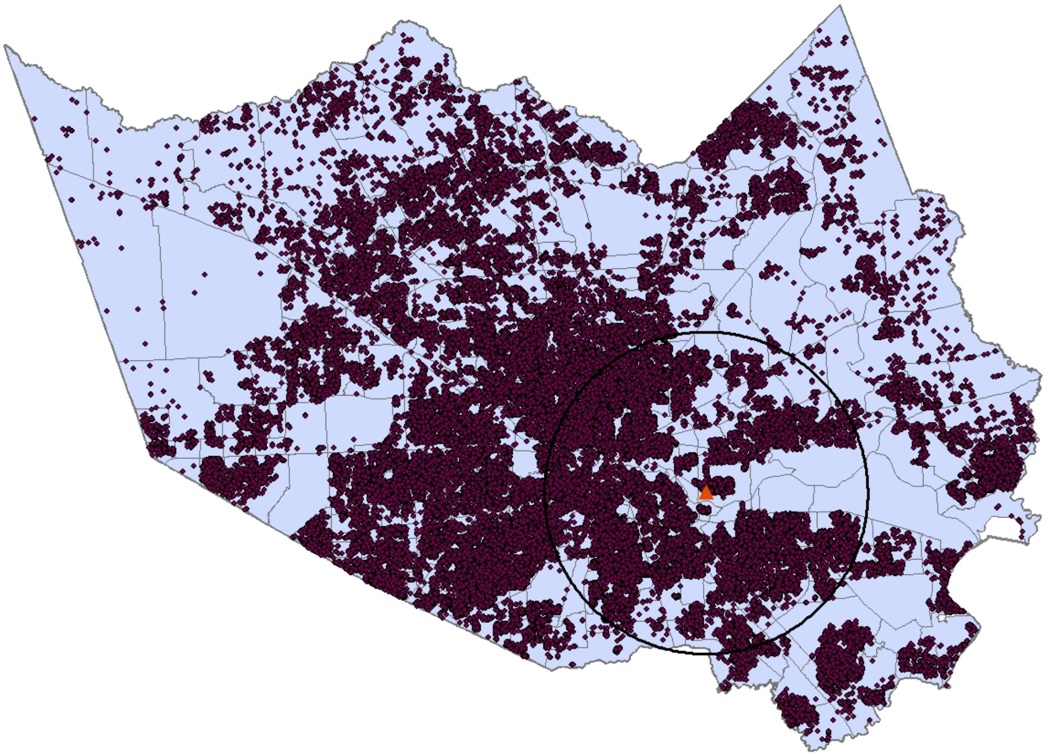


Figure E.2. 10-mile buffer surrounding the Clinton Drive monitoring site (red triangle) with geocoded residences of decedents (purple dots), Harris County, TX, 2000–2005.

Table E.4 summarizes the total and mean daily number of deaths for all-cause nonaccidental deaths as well as specific cardiovascular- and respiratory-cause deaths within the 10-mile buffer region; there were a total of 38,610 all-cause nonaccidental deaths with 17.6 deaths per day, and 14,794 (6.7 deaths per day) and 2,818 (1.3 deaths per day) deaths from cardiovascular and respiratory causes, respectively.

Table E.4. Summary statistics for all nonaccidental mortality and specific causes of cardiovascular and respiratory mortality, 2000–2005 for the 10-mile buffer region surrounding the Clinton Drive monitoring site (U.S. EPA monitoring site #: 48-201-1035)

Mortality cause	Number	Mean daily number of deaths
Cardiovascular (I00–I99)	15,316	6.7
Ischemic Heart Disease (IHD) (I20–I25)	7,384	3.4
Acute MI (I21)	2,910	1.3
Heart Failure (I50)	1,088	0.5
Stroke (I60–69)	2,806	1.3
Respiratory (J00–J98)	2,818	1.3
COPD (J40–44)	1,317	0.6
Pneumonia (J12–18)	825	0.4
Nonaccidental causes (A00–R99)	38,610	17.6

We also summarized the mean daily deaths of the mortality data. Tables E.5–E.10 provide the summaries for Harris County, as well as for the subpopulations defined in each of 5-mile and 10-mile buffers. Among deaths due to cardiovascular diseases for Harris County over the six-year period (see Table E.5), the highest daily mean was for IHD (9.29 deaths/day) followed by stroke (3.71 deaths/day) and acute MI (3.55 deaths/day). For the respiratory diseases under investigation in our study, COPD averaged 1.88 deaths per day and pneumonia 1.07 deaths per day. As expected, there were little differences by sex (23.83 and 24.88 deaths from nonaccidental causes for males and females, respectively) and the daily mean values increased by age group. Similar patterns were evident for deaths occurring within the 5- and 10-mile buffers surrounding the Clinton Drive monitoring station (see Tables E.6 and E.7). Tables E.8–E.10 provide a further demographic breakdown of the mortality data for Harris County from 2005–2007, as well as for the 5- and 10-mile buffers surrounding the Clinton Drive Monitoring Station for the following variables: race (White, Black, Other (includes Native American, Chinese, Japanese, Hawaiian, Filipino, Other Asian, Central-South American Indian, Asian Indian, Korean, Samoan, Vietnamese, Guamanian), Hispanic origin (yes, no, unknown), race/ethnicity (Non-Hispanic White, Non-Hispanic Black, Hispanic and Other), and years of completed education (< 12, 12, 13–16, >16, and unknown).

Table E.5. Average daily mortality by cause of death, Harris County, Texas, 2000–2005

							Selected Percentiles				
Underlying Cause of Death* (ICD-10 codes)	*N*	No. of days in which there were no deaths	Mean	SD	Min	Max	5th	25th	50th	75th	95th
Cardiovascular diseases (I00–I99)	41,708	0	19.03	4.70	5	40	12	16	19	22	27
IHD (I20–I25)	20,370	0	9.29	3.21	2	23	4	7	9	11	15
Acute MI (I21)	7,786	71	3.55	1.95	0	12	1	2	3	5	7
Heart Failure (I50)	2,884	571	1.32	1.13	0	6	0	0	1	2	3
Stroke (I60–I69)	8,129	49	3.71	1.98	0	12	1	2	4	5	7
Respiratory diseases (J00–J98)	8,478	68	3.87	2.13	0	13	1	2	4	5	8
COPD (J40–J44)	4,124	365	1.88	1.44	0	9	0	1	2	3	5
Pneumonia (J12–J18)	2,335	781	1.07	1.08	0	6	0	0	1	2	3
Nonaccidental (A00–R99)	106,772	0	48.71	7.73	28	80	37	43	48	54	62
Male	52,234	0	23.83	5.09	7	43	16	20	24	27	33
Female	54,538	0	24.88	5.42	10	48	17	21	25	28	34
< 20 yr	2,965	571	1.83	1.00	1	7	1	1	2	2	4
20–49 yr	11,254	11	5.16	2.23	1	16	2	4	5	7	9
50–64 yr	20,567	0	9.38	3.11	1	22	5	7	9	11	15
65–74 yr	20,175	1	9.21	3.19	1	21	4	7	9	11	15
75+ yr	51,811	0	23.64	5.33	7	46	16	20	23	27	33

*IHD = Ischemic Heart Disease; MI = Myocardial Infarction; COPD = Chronic Obstructive Pulmonary Disease

Table E.6. Average daily mortality by cause of death, 10-mile buffer region surrounding the Clinton Drive monitoring site in Houston, TX, 2000–2005

							Selected Percentiles				
Underlying Cause of Death* (ICD-10 codes)	*N*	No. of days in which there were no deaths	Mean	SD	Min	Max	5th	25th	50th	75th	95th
Cardiovascular diseases (I00–I99)	15,316	3	6.99	2.77	0	21	3	5	7	9	12
IHD(I20–I25)	7,384	83	3.37	1.86	0	11	1	2	3	4	7
Acute MI (I21)	2,190	590	1.33	1.17	0	6	0	0	1	2	4
Heart Failure (I50)	1,088	1,335	0.50	0.70	0	4	0	0	0	1	2
Stroke (I60–I69)	2,806	623	1.28	1.14	0	7	0	0	1	2	3
Respiratory diseases (J00–J98)	2,818	618	1.29	1.16	0	7	0	0	1	2	3
COPD (J40–J44)	1,317	1,205	0.60	0.78	0	5	0	0	0	1	2
Pneumonia (J12–J18)	825	1,524	0.38	0.63	0	4	0	0	0	1	2
Nonaccidental (A00–R99)	38,610	0	17.61	4.47	5	35	11	14	17	21	25
Male	19,364	2	8.84	3.05	1	23	4	7	9	11	14
Female	19,246	0	8.78	3.08	1	20	4	7	9	11	14
< 20 yr	953	1,406	1.21	0.49	1	5	1	1	1	1	2
20–49 yr	4,383	295	2.31	1.28	1	8	1	1	2	3	5
50–64 yr	7,780	75	3.68	1.81	1	13	1	2	3	5	7
65–74 yr	7,952	70	3.75	1.90	1	13	1	2	4	5	7
75+ yr	17,542	0	8.00	2.93	1	20	4	6	8	10	13

*IHD = Ischemic Heart Disease; MI = Myocardial Infarction; COPD = Chronic Obstructive Pulmonary Disease

Table E.7. Average daily mortality by cause of death, 5-mile buffer region surrounding the Clinton Drive monitoring site in Houston, TX, 2000–2005

							Selected Percentiles				
Underlying Cause of Death* (ICD-10 codes)	*N*	No. of days in which there were no deaths	Mean	SD	Min	Max	5th	25th	50th	75th	95th
Cardiovascular diseases (I00–I99)	3,673	436	1.68	1.33	0	8	0	1	1	2	4
IHD(I20–I25)	1,857	939	0.85	0.90	0	5	0	0	1	1	3
Acute MI (I21)	792	1,524	0.36	0.59	0	3	0	0	0	1	2
Heart Failure (I50)	284	1,932	0.13	0.37	0	3	0	0	0	0	1
Stroke (I60–I69)	719	1,609	0.30	0.55	0	3	0	0	0	1	1
espiratory diseases (J00–J98)	709	1,601	0.32	0.58	0	3	0	0	0	1	1
COPD (J40–J44)	324	1,897	0.15	0.39	0	3	0	0	0	0	1
Pneumonia (J12–J18)	197	2,007	0.09	0.30	0	2	0	0	0	0	1
Nonaccidental (A00–R99)	9,303	32	4.24	2.15	0	14	1	3	4	6	8
Male	4,654	270	2.42	1.37	1	9	1	1	2	3	5
Female	4,649	281	2.43	1.36	1	11	1	1	2	3	5
< 20 yr	260	1,946	1.06	0.25	1	3	1	1	1	1	2
20–49 yr	957	1,416	1.23	0.50	1	4	1	1	1	1	2
50–64 yr	1,703	1,010	1.44	0.69	1	5	1	1	1	2	3
65–74 yr	1,814	976	1.49	0.75	1	5	1	1	1	2	3
75+ yr	4,569	279	2.39	1.32	1	10	1	1	2	3	5

*IHD = Ischemic Heart Disease; MI = Myocardial Infarction; COPD = Chronic Obstructive Pulmonary Disease

Table E.8. Demographic breakdown for all nonaccidental cause mortality (ICD 10 A00–R99) and cause-specific mortality due to respiratory (ICD 10J00–J98) and cardiovascular diseases (ICD 10 I00–I99); Harris County, Texas, 2000–2005.

	All nonaccidental causes (*n* = 106,772)		Cardiovascular causes (*n* = 41,708)		Respiratory causes (*n* = 8,478)	
	N	%	*N*	%	*N*	%
Sex						
Male	52,234	48.9	20,155	48.3	4,015	47.4
Female	54,538	51.1	21,553	51.7	4,463	52.6
Age (years)						
< 20	2,965	2.8	144	.3	125	1.5
20–49	11,254	10.5	3,020	7.2	388	4.6
50–64	20,567	19.3	7,119	17.1	1,035	12.2
65–74	20,175	18.9	7,462	17.9	1,789	21.1
≥ 75	51,811	48.5	23,963	57.5	5,141	60.6
Race						
White	77,701	72.8	30,203	72.4	6,774	79.9
Black	26,398	24.7	10,521	25.2	1,514	17.9
Other*	2,605	2.4	966	2.3	187	2.2
Unknown	68	.1	18	.0	3	.0
Hispanic origin						
Yes	13,788	12.9	4,561	10.9	767	9.0
No	92,744	86.9	37,062	88.9	7,695	90.8
Unknown	240	.2	85	.2	16	.2
Race/ethnicity						
Non-Hispanic White	64,017	60.0	25,682	61.6	6,017	71.0
Non-Hispanic Black	26,385	24.7	10,515	25.2	1,513	17.8
Hispanic	13,717	12.8	4,536	10.9	759	9.0
Other	2,653	2.5	975	2.3	189	2.2
Education (years)						
< 12	34,265	32.1	13,507	32.4	2,799	33.0
12	35,183	33.0	13,852	33.2	2,949	34.8
13–16	27,287	25.6	10,500	25.2	2,026	23.9
> 16	6,764	6.3	2,649	6.4	425	5.0
Unknown	3,273	3.1	1,200	2.9	279	3.3
Year of death						
2000	17,386	16.3	7,185	17.2	1,362	16.1
2001	17,958	16.8	7,251	17.4	1,461	17.2
2002	18,096	16.9	7,187	17.2	1,389	16.4
2003	17,873	16.7	7,058	16.9	1,427	16.8
2004	17,422	16.3	6,612	15.9	1,373	16.2
2005	18,037	16.9	6,415	15.4	1,466	17.3

*Other includes the following: Native American, Chinese, Japanese, Hawaiian, Filipino, Other Asian, Central-South American Indian, Asian Indian, Korean, Samoan, Vietnamese, Guamanian

Table E.9. Demographic breakdown for all nonaccidental cause mortality (ICD 10 A00–R99) and cause-specific mortality due to respiratory (ICD 10J00–J98) and cardiovascular diseases (ICD 10 I00–I99); 10-mile buffer region surrounding the Clinton Drive monitoring site, 2000–2005.

	All nonaccidental causes (*n* = 38,610)		Cardiovascular causes (*n* = 15,316)		Respiratory causes (*n* = 2,818)	
	N	%	*N*	%	*N*	%
Sex						
Male	19,364	50.2	7,487	48.9	1,394	49.5
Female	19,246	49.8	7,829	51.1	1,424	50.5
Age (years)						
< 20	953	2.5	41	.3	40	1.4
20–49	4,383	11.4	1,163	7.6	156	5.5
50–64	7,780	20.2	2,825	18.4	372	13.2
65–74	7,952	20.6	3,149	20.6	662	23.5
≥75	17,542	45.4	8,138	53.1	1,588	56.4
Race						
White	22,510	58.3	8,770	57.3	1,873	66.5
Black	15,776	40.9	6,415	41.9	922	32.7
Other*	304	.8	124	.8	20	.7
Unknown	20	.1	7	.0	3	.1
Hispanic origin						
Yes	6,954	18.0	2,405	15.7	390	13.8
No	31,554	81.7	12,873	84.0	2,421	85.9
Unknown	102	.3	38	.2	7	.2
Race/ethnicity						
Non-Hispanic White	15,595	40.4	6,379	41.6	1,488	52.8
Non-Hispanic Black	15,770	40.8	6,414	41.9	922	32.7
Hispanic	6,927	17.9	2,394	15.6	386	13.7
Other	318	.8	129	.8	22	.8
Education (years)						
< 12	15,923	41.2	6,363	41.5	1,199	42.5
12	12,742	33.0	5,035	32.9	961	34.1
13–16	6,811	17.6	2,698	17.6	452	16.0
> 16	1,525	3.9	625	4.1	88	3.1
Unknown	1,609	4.2	595	3.9	118	4.2
Year of death						
2000	6,552	17.0	2,721	17.8	485	17.2
2001	6,690	17.3	2,772	18.1	492	17.5
2002	6,548	17.0	2,631	17.2	444	15.8
2003	6,432	16.7	2,628	17.2	468	16.6
2004	6,127	15.9	2,321	15.2	450	16.0
2005	6,261	16.2	2,243	14.6	479	17.0

*Other includes the following: Native American, Chinese, Japanese, Hawaiian, Filipino, Other Asian, Central-South American Indian, Asian Indian, Korean, Samoan, Vietnamese, Guamanian

Table E.10. Demographic breakdown for all nonaccidental cause mortality (ICD 10 A00–R99) and cause-specific mortality due to respiratory (ICD 10J00–J98) and cardiovascular diseases (ICD 10 I00–I99); 5-mile buffer region surrounding the Clinton Drive monitoring site, 2000–2005.

	All nonaccidental causes (*n* = 9,303)		Cardiovascular causes (*n* = 3,673)		Respiratory causes (*n* = 709)	
	N	%	*N*	%	*N*	%
Sex						
Male	4,654	50.0	1,789	48.7	357	50.4
Female	4,649	50.0	1,884	51.3	352	49.6
Age (years)						
< 20	260	2.8	10	.3	12	1.7
20–49	957	10.3	231	6.3	35	4.9
50–64	1,703	18.3	602	16.4	62	8.7
65–74	1,814	19.5	709	19.3	160	22.6
≥ 75	4,569	49.1	2,121	57.7	440	62.1
Race						
White	7,505	80.7	2,922	79.6	599	84.5
Black	1,715	18.4	722	19.7	105	14.8
Other*	79	.8	29	.8	5	.7
Unknown	4	.0	0	0	0	0
Hispanic origin						
Yes	3,295	35.4	1,127	30.7	178	25.1
No	5,992	64.4	2,542	69.2	531	74.9
Unknown	16	.2	4	.1	0	0
Race/ethnicity						
Non-Hispanic White	4,218	45.3	1,799	49.0	421	59.4
Non-Hispanic Black	1,714	18.4	722	19.7	105	14.8
Hispanic	3,290	35.4	1,123	30.6	178	25.1
Other	81	.9	29	.8	5	.7
Education (years)						
< 12	4,475	48.1	1,715	46.7	327	46.1
12	2,762	29.7	1,143	31.1	211	29.8
13–16	1,384	14.9	561	15.3	114	16.1
> 16	214	2.3	88	2.4	10	1.4
Unknown	468	5.0	166	4.5	47	6.6
Year of death						
2000	1,643	17.7	684	18.6	121	17.1
2001	1,649	17.7	673	18.3	127	17.9
2002	1,584	17.0	658	17.9	96	13.5
2003	1,550	16.7	615	16.7	125	17.6
2004	1,409	15.1	525	14.3	109	15.4
2005	1,468	15.8	518	14.1	131	18.5

*Other includes the following: Native American, Chinese, Japanese, Hawaiian, Filipino, Other Asian, Central-South American Indian, Asian Indian, Korean, Samoan, Vietnamese, Guamanian

AIR POLLUTION DATA FROM CLINTON DRIVE AND HARRIS COUNTY

We obtained the following data from the Texas Commission on Environmental Quality (TCEQ):

1. Validated hourly AutoGC VOC data for Harris County for January 2000 – June 2010
2. Validated 24-hr canister Volatile Organic Compounds (VOC) data for Harris County measured every 6 days for January 2000 – June 2010.
3. Validated 24-hr PM_{10} speciation data for Harris County for January 2000 – June 2010
4. Validated 24-hr $PM_{2.5}$ speciation data for Harris County for January 2000 – December 2005.

All but the AutoGC VOC data were in a readily useable format. There are 8 AutoGC monitoring sites in Harris County (see Figure E.3). Table E.11 shows the list of AutoGC monitoring sites and years for which data were collected at each site. For some years, especially for the beginning and ending years, data were available only for part of the year. The AutoGC data were provided in more than 700 pipe-delimited text files, each containing approximately 20,000–35,000 records, which are contained in multiple subfolders. (Each first-tier subfolder represents a site, and within each site folder there are subfolders for each year.) The number of data files for each site is also shown in Table E.11. These files were combined to make 8 files. One combined file for each site was made.

Table E.11. AutoGC Monitoring sites in Harris County

Site Number	Site Name	Number of files	Years covered
482010026	Channelview	102	2001–2010
482010069	Milby Park	60	2005–2010
482010617	Wallisville	81	2003–2010
482010803	HRM3	84	2003–2010
482011015	Lynchburg Ferry	85	2003–2010
482011035	Clinton	125	2000–2010
482011039	Deer Park	121	2000–2010
482016000	Cesar Chavez	70	2004–2010

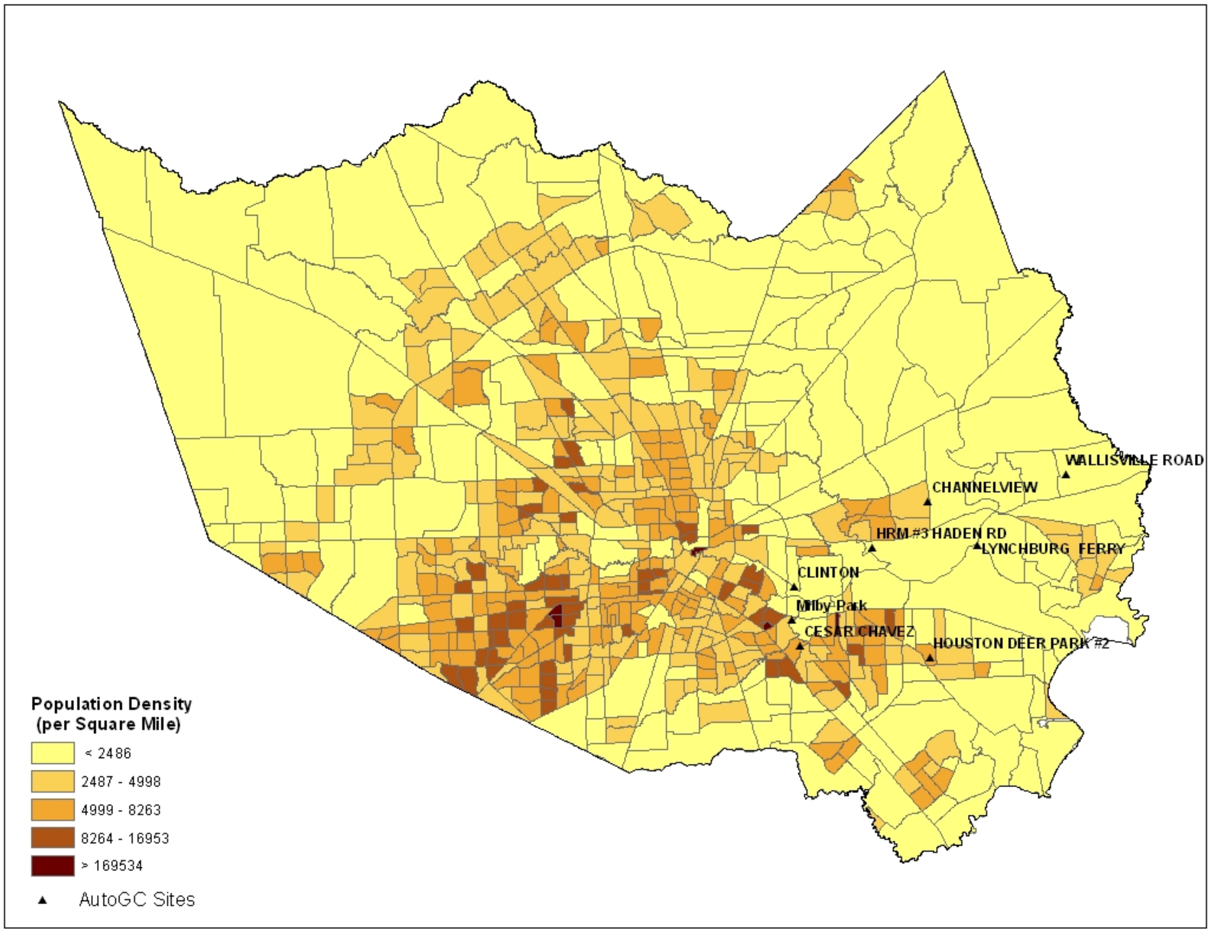


Figure E.3. Map of eight AutoGC monitoring sites with population density in Harris County, TX

The canister VOC data provided by the TCEQ consist of 24-hour measurements on 107 species collected every 6 days from 17 monitoring sites in Harris County during January 2000–August 2009. Table E.12 contains the list of canister VOC monitoring sites and years for which data were collected at each site.

Table E.12. Canister VOC Monitoring sites in Harris County

Site Number	Site Name	Years covered
482011039	Houston Deer Park #2	2000–2009
482011035	Clinton	2000–2006
482010803	HRM #3 Haden Rd	2000–2009
482010069	Milby Park	2000–2005
482010061	Shores Acres	2000–2009
482010058	Baytown	2000–2009
482010055	Houston Bayland Park	2000–2009
482010029	Northwest Harris County	2000–2009
482010026	Channelview	2000–2006
482010024	Houston Aldine	2000–2009
482010057	Galena Park	2000–2009
482011041	San Jacinto Monument	2000–2003
482011015	Lynchburg Ferry	2003–2009
482010307	Manchester/Central	2005–2009
482010030	Channelview North	2005–2007
482010036	Jacinto Port	2006–2009
482011049	Pasadena North	2008–2009

The 24-hour PM_{10} speciation data provided by TCEQ contain the speciated PM_{10} measurements collected every 3rd or 6th day from two monitoring sites in Table E.13.

Table E.13. Sites with available PM_{10} speciation data in Harris County

Site Number	Site Name	Years covered
482011035	Clinton	2000–2010
482011039	Houston Deer Park #2	2000–2009

Table E.14 shows the site names and the periods of availability for the 24-hour $PM_{2.5}$ speciation data (collected every 3rd or 6th day) during 2000–2005 in Harris County.

Table E.14. Sites with available $PM_{2.5}$ speciation data during 2000–2005 in Harris County

Site Number	Site Name	Periods of data
482010024	Aldine	Aug. 2000–Dec. 2005
482010026	Channelview	Aug. 2000–Aug. 2005
482010055	Bayland Park	Aug. 2000–Aug. 2005
482010803	HRM 3	Aug. 2000–Nov. 2001
482011034	Houston East	Jan. 2002–Aug. 2005
482011039	Deer Park	Feb. 2000–Dec. 2005

For the assessment of source-specific health effects in the Clinton Drive region, we decided to use the $PM_{2.5}$ speciation data for which some information on potential source types around the area were available from previous studies. Because there are no $PM_{2.5}$ speciation data available at the Clinton Drive monitoring site, the $PM_{2.5}$ data measured every 6th day for January 2002–August 2005 from Houston East were used for the analysis. This site is closest to Clinton Drive (it is 3 miles northeast of Clinton), and the data from this site has been used in other source-apportionment analysis of Clinton Drive (see Sullivan 2006). Summary statistics for the original 77 $PM_{2.5}$ species measured at the Houston East monitoring station are provided in Table E.15.

Table E.15. Summary Statistics for $PM_{2.5}$ Chemical Species Measured at Houston East

$PM_{2.5}$ species	Number of nonmissing values	Average	SD	Minimum	Maximum
Aluminum	217	0.073	0.191	0	1.410
Ammonium Ion	217	1.251	0.884	0	5.690
Antimony	217	0.004	0.006	0	0.028
Arsenic	217	0.001	0.002	0	0.018
Barium	217	0.010	0.010	0	0.061
Bromine	217	0.003	0.002	0	0.013
Cadmium	217	0.001	0.002	0	0.010
Calcium	217	0.099	0.066	0	0.395
Carbonate Carbon Csn Tot	95	0.000	0.000	0	0.000
Cerium	217	0.002	0.006	0	0.037
Cesium	217	0.003	0.004	0	0.023
Chlorine	217	0.070	0.184	0	1.440
Chromium	217	0.001	0.003	0	0.035
Cobalt	217	0.000	0.001	0	0.012
Copper	217	0.005	0.003	0	0.017
EC Csn Tot	224	0.676	0.348	0	2.050
EC Tor	0	.	.	.	.
EC1	0	.	.	.	.
EC2	0	.	.	.	.
EC3	0	.	.	.	.
Europium	217	0.007	0.015	0	0.129
Gallium	217	0.000	0.001	0	0.008
Gold	217	0.001	0.001	0	0.005
Hafnium	217	0.003	0.005	0	0.046
Indium	217	0.002	0.002	0	0.011
Iridium	217	0.000	0.001	0	0.004
Iron	217	0.111	0.114	0	0.877
Lanthanum	217	0.003	0.006	0	0.037
Lead	217	0.003	0.006	0	0.086
Magnesium	217	0.029	0.056	0	0.329
Manganese	217	0.002	0.002	0	0.010
Mercury	217	0.000	0.001	0	0.007
Molybdenum	217	0.001	0.002	0	0.007
Nickel	217	0.002	0.002	0	0.008
Niobium	217	0.000	0.001	0	0.006
Non-Volatile Nitrate	217	0.345	0.438	0.01685	3.630
Oc Csn Unadjusted Tot	224	3.464	1.690	0.416	9.610
Oc Tor	0	.	.	.	.
Oc1 Csn Unadjusted Tot	129	0.842	0.423	0.075	2.085
Oc1	0	.	.	.	.
Oc2 Csn Unadjusted Tot	129	0.863	0.344	0.199	1.851
Oc2	0	.	.	.	.
Oc3 Csn Unadjusted Tot	129	0.583	0.251	0.0896	1.550
Oc3	0	.	.	.	.

Oc4 Csn Unadjusted Tot	129	1.160	0.697	0.0185	3.692
Oc4	0	.	.	.	.
Och Tot	0	.	.	.	.
Ocx Carbon	0	.	.	.	.
Ocx2 Carbon	95	1.319	0.588	0.211	3.410
Op Csn Tot	81	0.225	0.560	0	3.510
Op Tor	0	.	.	.	.
Phosphorus	217	0.015	0.034	0	0.178
$PM_{2.5}$ – Local Conditions	216	12.944	5.825	0.3	35.000
Potassium Ion	217	0.042	0.048	0	0.346
Potassium	217	0.075	0.059	0	0.359
Rubidium	217	0.000	0.000	0	0.004
Samarium	217	0.003	0.011	0	0.129
Scandium	217	0.000	0.001	0	0.007
Selenium	217	0.001	0.001	0	0.003
Silicon	217	0.283	0.426	0	3.220
Silver	217	0.001	0.002	0	0.013
Sodium Ion	217	0.223	0.267	0	1.420
Sodium	217	0.158	0.236	0	1.490
Strontium	217	0.001	0.001	0	0.006
Sulfate	217	3.761	2.279	0.0204	14.800
Sulfur	217	1.291	0.781	0	5.100
Tantalum	217	0.002	0.003	0	0.018
Terbium	217	0.003	0.011	0	0.087
Tin	217	0.003	0.005	0	0.029
Titanium	217	0.008	0.013	0	0.113
Total Nitrate	0	.	.	.	.
Tungsten	217	0.002	0.002	0	0.014
Vanadium	217	0.005	0.005	0	0.029
Volatile Nitrate	0	.	.	.	.
Yttrium	217	0.000	0.001	0	0.003
Zinc	217	0.016	0.021	0	0.201
Zirconium	217	0.001	0.001	0	0.008

Note: All units are in µg/m^3.

Based on optimal spatial coverage and years of availability (2000–2005), the Canister VOC data from the following 9 monitoring sites shown in Figure E.4 were selected for multisite analysis within Harris County.

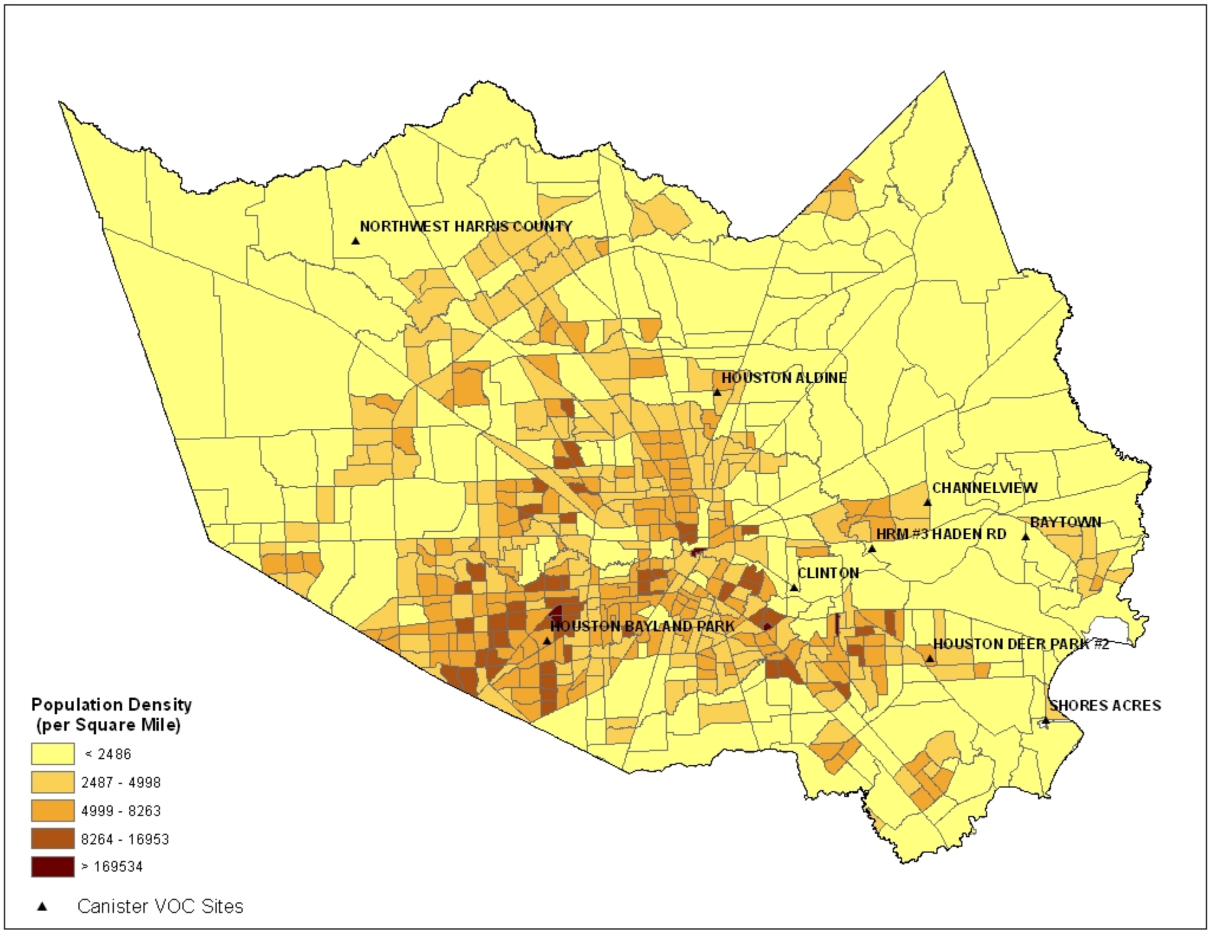


Figure E.4. Map of nine Canister VOC monitoring sites with population density in Harris County, TX

Tables E.16 and E.17 shows the information on the summary of missing values for the Canister VOC data from 9 monitoring stations in Figure E.4.

Table E.16: Proportion of missing values for Canister VOC data by species

Species name	Missing	Total	Percent Missing
1,2,4-Trimethylbenzene	788	3,789	20.80
1,3-Butadiene	788	3,789	20.80
2,2,4-Trimethylpentane	788	3,789	20.80
Acetylene	788	3,789	20.80
Benzene	788	3,789	20.80
Ethane	788	3,789	20.80
Ethylbenzene	788	3,789	20.80
Ethylene	788	3,789	20.80
Isobutane	788	3,789	20.80
Isopentane	788	3,789	20.80
Propane	788	3,789	20.80
Propylene	788	3,789	20.80
Toluene	788	3,789	20.80
n-Butane	788	3,789	20.80
n-Hexane	788	3,789	20.80
n-Pentane	788	3,789	20.80
p-Xylene+m-Xylene	788	3,789	20.80

Table E.17: Proportion of missing values for Canister VOC data by Monitoring site

Monitoring site	**Missing**	**Total**	**Percent Missing**
1	91	421	21.0
2	128	421	30.0
3	116	421	27.0
4	83	421	19.0
5	92	421	21.0
6	76	421	18.0
7	76	421	18.0
8	34	421	8.0
9	92	421	21.0

Tables E.18–E.21 contain the sample correlations over monitoring stations for some of the VOC species. It can be seen that for some species such as Acetylene and Ethane, spatial correlations are more significant compared to Benzene or Ethylene.

Table E.18: Multivariate correlations of Acetylene concentrations between 9 monitoring stations

	Acetylene 1	**Acetylene 2**	**Acetylene 3**	**Acetylene 4**	**Acetylene 5**	**Acetylene 6**	**Acetylene 7**	**Acetylene 8**	**Acetylene 9**
Acetylene 1	1.0000	0.5545	0.6111	0.8103	0.3814	0.3328	0.8119	0.7532	0.6868
Acetylene 2	0.5545	1.0000	0.1507	0.5576	0.2315	0.1870	0.4367	0.5297	0.4865
Acetylene 3	0.6111	0.1507	1.0000	0.3833	0.2301	0.3215	0.6733	0.3615	0.2930
Acetylene 4	0.8103	0.5576	0.3833	1.0000	0.2809	0.2587	0.6599	0.8322	0.8147
Acetylene 5	0.3814	0.2315	0.2301	0.2809	1.0000	0.2115	0.3178	0.2821	0.2391
Acetylene 6	0.3328	0.1870	0.3215	0.2587	0.2115	1.0000	0.3956	0.4737	0.2813
Acetylene 7	0.8119	0.4367	0.6733	0.6599	0.3178	0.3956	1.0000	0.6786	0.5363
Acetylene 8	0.7532	0.5297	0.3615	0.8322	0.2821	0.4737	0.6786	1.0000	0.6920
Acetylene 9	0.6868	0.4865	0.2930	0.8147	0.2391	0.2813	0.5363	0.6920	1.0000

There are 279 missing values. The correlations are estimated by REML method.

Table E.19: Multivariate correlations of Benzene concentrations between 9 monitoring stations

	Benzene 1	**Benzene 2**	**Benzene 3**	**Benzene 4**	**Benzene 5**	**Benzene 6**	**Benzene 7**	**Benzene 8**	**Benzene 9**
Benzene 1	1.0000	0.3964	0.6152	0.4980	0.3635	0.1236	0.6639	0.2487	0.3898
Benzene 2	0.3964	1.0000	0.3265	0.4212	0.1868	0.0231	0.4176	0.2858	0.2172
Benzene 3	0.6152	0.3265	1.0000	0.3549	0.1899	0.1395	0.4390	0.1492	0.3419
Benzene 4	0.4980	0.4212	0.3549	1.0000	0.1398	0.0665	0.5129	0.1711	0.3276
Benzene 5	0.3635	0.1868	0.1899	0.1398	1.0000	0.0831	0.3345	0.1031	0.0226
Benzene 6	0.1236	0.0231	0.1395	0.0665	0.0831	1.0000	0.1263	0.0217	0.0932
Benzene 7	0.6639	0.4176	0.4390	0.5129	0.3345	0.1263	1.0000	0.3140	0.1632
Benzene 8	0.2487	0.2858	0.1492	0.1711	0.1031	0.0217	0.3140	1.0000	–0.0462
Benzene 9	0.3898	0.2172	0.3419	0.3276	0.0226	0.0932	0.1632	–0.0462	1.0000

There are 279 missing values. The correlations are estimated by REML method

Table E.20: Multivariate correlations of Ethane concentrations between 9 monitoring stations

	Ethane 1	Ethane 2	Ethane 3	Ethane 4	Ethane 5	Ethane 6	Ethane 7	Ethane 8	Ethane 9
Ethane 1	1.0000	0.7249	0.6679	0.8565	0.7150	0.4328	0.8384	0.8494	0.7395
Ethane 2	0.7249	1.0000	0.6138	0.7280	0.6423	0.4379	0.7384	0.7261	0.6756
Ethane 3	0.6679	0.6138	1.0000	0.6583	0.5205	0.2338	0.5810	0.5860	0.6016
Ethane 4	0.8565	0.7280	0.6583	1.0000	0.7609	0.4399	0.7748	0.8732	0.8163
Ethane 5	0.7150	0.6423	0.5205	0.7609	1.0000	0.5901	0.6721	0.7696	0.7978
Ethane 6	0.4328	0.4379	0.2338	0.4399	0.5901	1.0000	0.3562	0.4712	0.4921
Ethane 7	0.8384	0.7384	0.5810	0.7748	0.6721	0.3562	1.0000	0.8523	0.6707
Ethane 8	0.8494	0.7261	0.5860	0.8732	0.7696	0.4712	0.8523	1.0000	0.7979
Ethane 9	0.7395	0.6756	0.6016	0.8163	0.7978	0.4921	0.6707	0.7979	1.0000

There are 279 missing values. The correlations are estimated by REML method.

Table E.21: Multivariate correlations of Ethylene concentrations between 9 monitoring stations

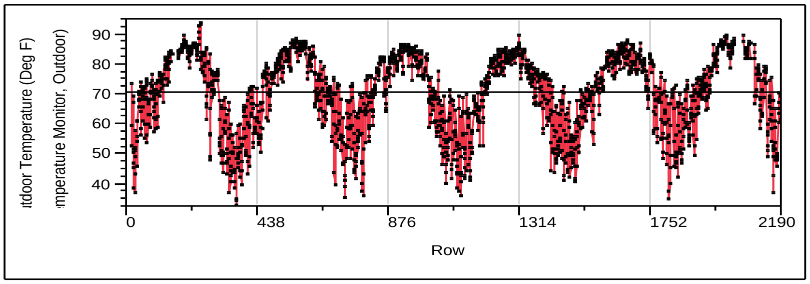
	Ethylene 1	Ethylene 2	Ethylene 3	Ethylene 4	Ethylene 5	Ethylene 6	Ethylene 7	Ethylene 8	Ethylene 9
Ethylene 1	1.0000	0.4060	0.6410	0.6241	0.2146	0.0754	0.5025	0.4427	0.3923
Ethylene 2	0.4060	1.0000	0.2913	0.3262	0.2629	0.0869	0.3364	0.1700	0.2418
Ethylene 3	0.6410	0.2913	1.0000	0.4882	0.0217	–0.0930	0.3092	0.4030	0.2470
Ethylene 4	0.6241	0.3262	0.4882	1.0000	0.0980	0.1708	0.2637	0.5239	0.5617
Ethylene 5	0.2146	0.2629	0.0217	0.0980	1.0000	0.0889	0.1655	0.0392	0.0403
Ethylene 6	0.0754	0.0869	–0.0930	0.1708	0.0889	1.0000	0.0596	–0.0602	0.2356
Ethylene 7	0.5025	0.3364	0.3092	0.2637	0.1655	0.0596	1.0000	0.2644	–0.0262
Ethylene 8	0.4427	0.1700	0.4030	0.5239	0.0392	–0.0602	0.2644	1.0000	0.2218
Ethylene 9	0.3923	0.2418	0.2470	0.5617	0.0403	0.2356	–0.0262	0.2218	1.0000

There are 279 missing values. The correlations are estimated by REML method.

WEATHER DATA FOR CLINTON DRIVE AND HARRIS COUNTY

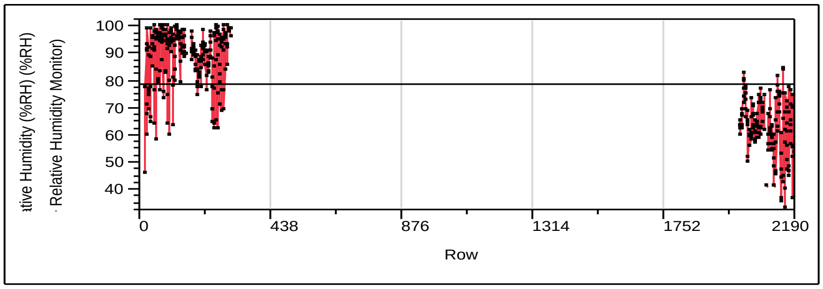
Originally, we obtained the validated hourly temperature and relative humidity data for Harris County 2000–2010 from TCEQ. However, temperature and relative humidity data from TCEQ have many missing values (e.g., the Clinton Drive monitoring station has 4% and 80% of missing values for temperature and relative humidity, respectively. See Figure E.5).

Time Series Clinton Drive Outdoor Temperature



Mean	70.329591
Std	12.53465
N	2106

Time Series Clinton Drive Relative Humidity



Mean	78.56168
Std	16.333223
N	440

Figure E.5. Time series of daily mean temperature and relative humidity data at Clinton Drive. The y-axis label for the first plot should be 'Outdoor Temperature (Deg F)' and the y-axis label for the second plot should be 'Relative Humidity (%RH)'

Due to a high proportion of missing values in relative humidity, we obtained another set of meteorological data from the National Oceanic and Atmospheric Administration (NOAA) —

National Climatic Data Center (*http://www.ncdc.noaa.gov/*). Specifically, weather data collected from the Automated Surface Observation System (ASOS) stations were primarily utilized in the study. The AOSO serves as the nation’s primary surface weather observing network, often located at airport locations, and provides daily summary of meteorological data, including minimum, maximum, and mean values of temperature, mean values of dew point, air pressure, wind speed, and precipitation (NOAA 1998). This study thus used data obtained from the two ASOS stations to represent spatial and temporal variability in weather conditions within the study area: 1) Hobby airport station covering the southern part of the study area and 2) Houston intercontinental airport station covering the northern part of the study area. The map in Figure E.6 shows the location of two weather stations. An additional station within the study area (Hooks airport) was identified, but it was not used in the study because of its close proximity to the Houston intercontinental airport station.

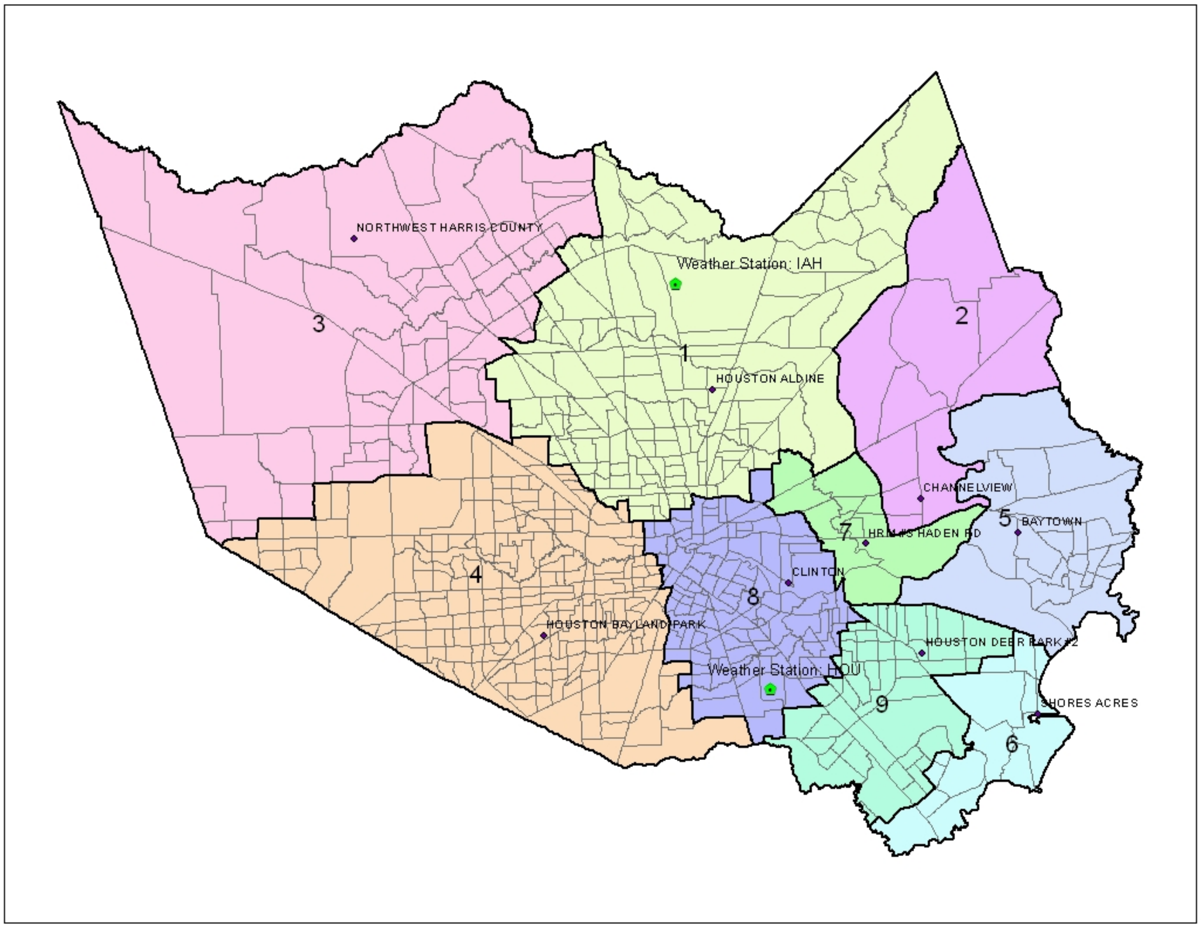
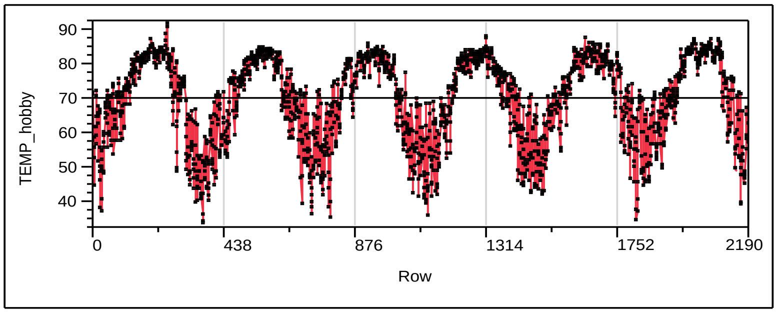


Figure E.6. Map of nine subregions in Harris County and Location of Two ASOS Weather Stations in Harris County. IAH = Houston Intercontinental airport station, HOU = Hobby airport station

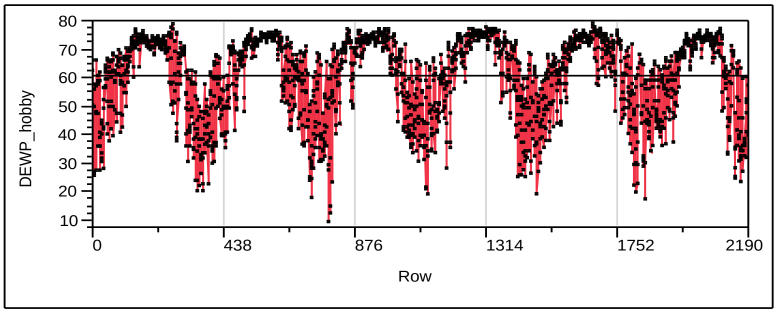
Temperature and dew point temperature data obtained from the above two stations have 100% complete values. Descriptive summary statistics (Mean and SD) of temperature and dew point temperature measured at the Houston Intercontinental airport station were within less than 1 degree of difference; temperature with mean 69.1 and SD 12.9, and dew point temperature with mean 59.6 and SD 13.9, respectively. Figure E.7 shows time-series plots of temperature and dew point temperature measured at the Hobby station. Daily 24-hour average temperature and dew point temperature data for the time periods between 2000 and 2005 were used as confounding variables in the subsequent analyses of air pollution and mortality.

Time Series TEMP_hobby



Mean	69.931661
Std	12.177313
N	2192

Time Series DEWP_hobby



Mean	60.917518
Std	13.802336
N	2192

Figure E.7. Time series of daily mean temperature and dewpoint temperature data at Hobby

Reference

NOAA 1998. Automated Surface Observation System (ASOS) User’s Guide. National Oceanic and Atmospheric Administration.