CART Analysis for Ozone Trends and Meteorological Similarity

May 10, 2007
Donna Kenski
Lake Michigan Air Directors Consortium

ABSTRACT

A hierarchical model (regression tree) was applied to ozone and meteorological data from 7 Midwestern cities. The model classified all summer days into bins based on meteorologically similar characteristics. Meteorological variables included temperature, dewpoint, pressure, relative humidity, solar radiation, cloud cover, morning and afternoon mixing height, wind direction (as north-south and east-west component vectors), wind speed, lake breeze indicator (where relevant), day of week, temperature increase or decrease from previous day, pressure increase or decrease from previous day, and previous-day temperature, pressure, wind speed, wind direction, and ozone. The years from 1990-2006 were modeled for each city. Trends in ozone concentrations were examined by comparing the change in average bin concentrations in an effort to control for the effect of meteorological variability. An index of 'ozone conduciveness' was developed based on the regression results in order to compare the meteorological similarity of various years.

INTRODUCTION

Meteorological conditions affect pollutant concentrations in myriad ways, complicating the process of modeling formation and transport and determining responses to control measures. Ozone is particularly dependent on meteorology, since its production is driven by high temperatures and sunlight as well as concentrations of its precursors, nitrogen oxides and hydrocarbons. The nonlinearity of these relationships adds to the difficulty in developing predictive models.

One novel method of quantifying the relationship between multiple meteorological variables and ozone is Classification and Regression Tree (CART) analysis. This technique, also known as binary recursive partitioning, was developed in 1984 by Breiman and Friedman.[1] It has several advantages as a tool for data mining and predictive modeling. The tree produced represents a model or decision tree in which each node (branch) is determined by splitting the dataset on the basis of the one variable that results in the best separation as defined by values of the dependent variable (in this case, ozone concentration). The splitting rule is expressed in natural language – for example, is temperature less than 75ºF – so the output trees are easy to interpret. At every branch, every variable is tested for its usefulness in further splitting. This exhaustive search for splitters can make CART computationally intensive.

METHODOLOGY

A CART analysis (regression tree) was applied to the 1990-2006 ozone and meteorology data for 7 Midwestern urban areas. The purpose was threefold: (1) to categorize specific ozone-conducive conditions for each city, (2) determine how representative each year was in terms of ozone-forming potential, and (3) assess ozone trends, using the CART bins as meteorologically adjusted results. The application of the regression tree was straightforward, using CART software from Salford Systems[2]. Trees selected in this analysis were generally smaller than the optimal trees (i.e., with fewer terminal nodes or branches) and contained 10 to 20 terminal nodes. Emphasis was on finding trees that were able to distinguish the extreme ozone days and also several subsets of moderately high ozone days. Low ozone conditions were of less interest. The model was constrained to include at least 25 days in each terminal node, in order to have a more robust distribution of days across the years. Trees were tested using v-fold cross-validation because the highest ozone days were infrequent, despite the reasonably large number of observations per city (~3500). The ozone monitoring data was restricted to a subset of monitors that have been running continuously or nearly continuously for this 17 year period. The average maximum daily 8-hour concentration at these monitors (by city) was used as the dependent variable.

In order to determine the 'representativeness' of each year, a metric was developed based on the number of days assigned to each node. First the average number of days per node-year was calculated for the entire 17-year period. An index of each year's variability from the average was then calculated as the sum (over all nodes) of the difference between each year and the 17-year average, divided by the average:

$$\text{Representativeness of year i} = \sum_j \left(\frac{n_{i,j} - \mu_j}{\mu_j} \right)$$

where

$n_{i,j}$ = number of days in year i in node j
μ_j = average number of days in node j over 17-year period.

DATA

The cities of interest were Chicago, Detroit, Milwaukee, St. Louis, Indianapolis, Cincinnati, Cleveland, and Minneapolis. Meteorological data were collected from National Weather Service TDL hourly observation tapes. In each city the primary airport data were used to represent daily conditions. Daily maximum ozone concentration was the dependent variable, calculated as the maximum 8-hour ozone observed at any of the monitors previously selected as having a continuous data record for the study period. Mean daily ozone was calculated as the average of all daily 8-hour ozone observations at the same set of monitors.

Meteorological and air quality variables used in the model were as follows:
Maximum and mean daily temperature (F); maximum and mean daily wind speed (mph); wind direction (vectorized to easterly and northerly components, average and

maximum); maximum and mean daily dew point (F); maximum and mean daily pressure (mb); precipitation (in); morning (7-10 am), afternoon (1-5 pm), and evening (8-10 pm) dew point, pressure, and wind direction; previous day's maximum and mean temperature, dew point, pressure, wind speed, wind direction, and ozone; and direction of temperature and pressure change from previous day (rising or falling). The model period was restricted to the months of April through October since ozone does not exceed the National Ambient Air Quality Standard during the colder months in the target cities.

RESULTS

The regression tree for Cleveland is presented as an example in Figure 1. The splitting criteria for each node are given within the blue boxes. If the condition is true (maximum temperature is less than 77 .5), follow the left branch, otherwise follow the right branch. Terminal nodes (red boxes) give an average concentration and standard deviation of all the ozone concentrations that fall into that node.

In all of the cities, maximum daily temperature was the most important variable for categorizing ozone, followed in importance by previous-day temperature, dew point, previous-day maximum ozone, and average wind speed. Not all of these variables appear as splitters in every tree; the relative importance of each variable is assessed based on its importance over all possible nodes and splits. In any one node, only one variable will be the best splitter although another may be a close second best (a good surrogate). The second-best variable may be a good surrogate for numerous splits without ever being selected as the best primary splitter. Its usefulness as a surrogate for multiple splits leads to its higher importance.

In Cleveland, as shown in Fig. 1, node 14 was the highest-concentration node, with an average of 78 ppb (8-hr) for the 175 days that met the meteorological conditions in this branch. The presence of high previous-day ozone concentrations and sustained high temperatures indicate that these days are probably part of multiday episodes. Nodes 9, 11, and 13 are characterized by more moderate concentrations in the 65-70 ppb range. Trees for the other cities are presented in the appendix.

Concentration trends in the ozone nodes are shown in Fig. 2. Because the meteorological conditions in each node are consistent, the trend over time in each node is meteorologically adjusted. That is, the annual meteorological variability that confounds our ability to determine trends in ozone as a result of decreases in precursor emissions is removed from these data. Although not all nodes are consistently downward, there is a modest decrease in most nodes and most cities. Figure 3 shows the overall trend for all of the higher-ozone nodes together. The size of each bubble plotted shows the number of days in that node/year.

Figure 4 shows the ozone conduciveness index. Of particular interest in this study were the years 2004-2006, because 2005 was selected as a new baseline year for photochemical modeling work. The 3-year average of 2004-2006 thus becomes the

new baseline design value. The ozone conduciveness index was developed to assess how similar years were, based on meteorological conditions. Using the CART results incorporates all of the CART met conditions – temperature, dew point, wind speed and direction, etc. From Fig. 4, it is apparent that the 2005 was more ozone conducive than the long term average. Both 2004 and 2006 were quite low in terms of their ozone forming potential. Previous years that were highly conducive include 2002, 1999, 1995, and 1991. When put in context with ozone concentrations, the results seem to show that the recent improvements may be due to emissions decreases. Ozone concentrations in 2005 were only moderate, despite the meteorologically conducive conditions. The year 2004 was exceptionally cool, with the lowest ozone of the past 27 years. Ozone concentrations in 2006 were the second lowest of the last 27 years, much lower than the index implies. Because ozone in 2005 and 2006 was lower than in previous meteorologically similar years, we conclude that emissions reductions played a role in the decrease.

REFERENCES

1. Breiman, L., J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees, Pacific Grove, CA: Wadsworth (1984).

2. Steinberg, D. and P. Colla, CART—Classification and Regression Trees, San Diego, CA, Salford Systems (1997)

ACKNOWLEDGEMENTS

Many thanks to Jim Heywood of Michigan Department of Environmental Quality for providing formatted meteorological data for this analysis.

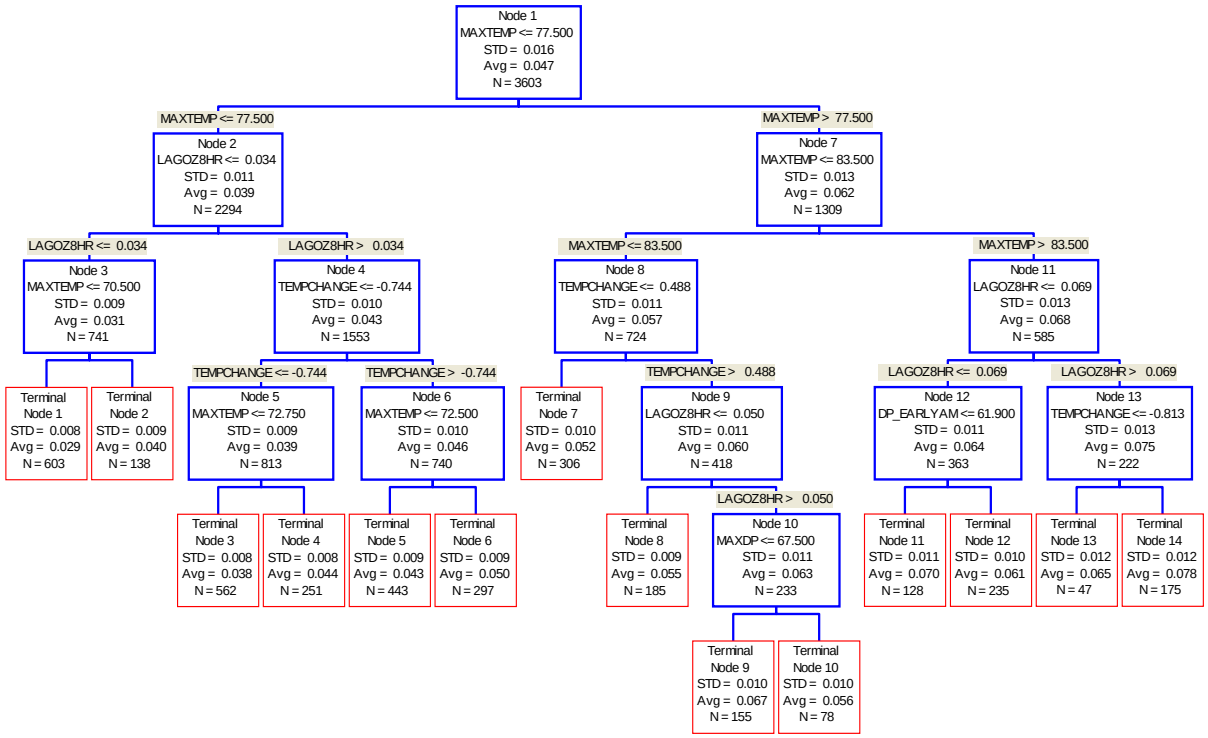


Figure 1 Regression tree for Cleveland

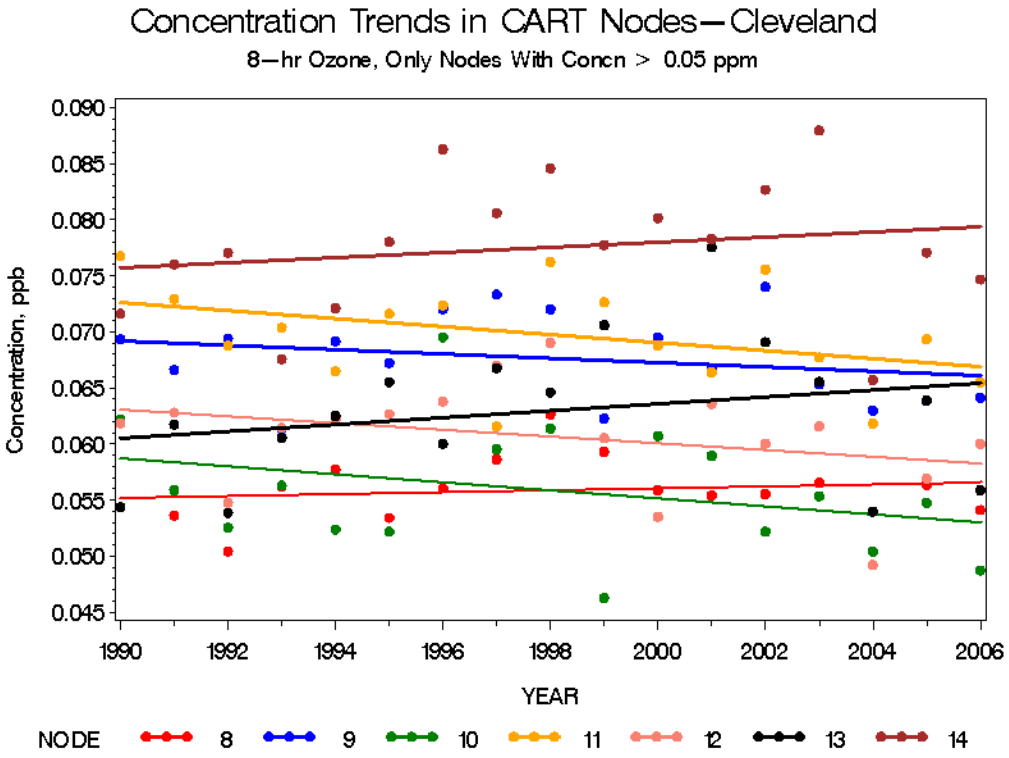


Figure 2 Concentration Trends in CART Nodes---Cleveland

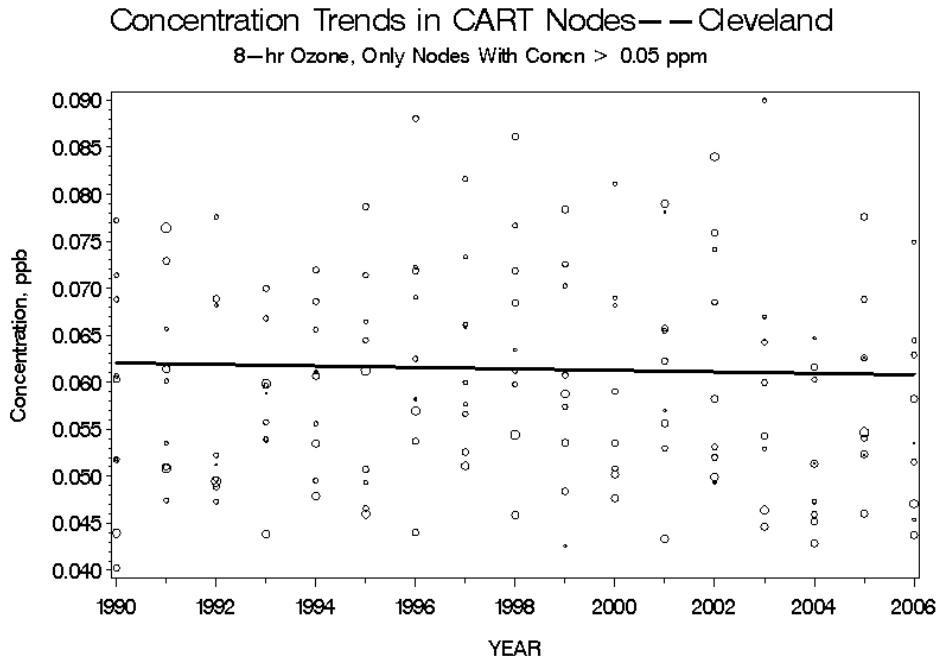


Figure 3. Overall Concentration Trend in High Ozone Nodes, Cleveland

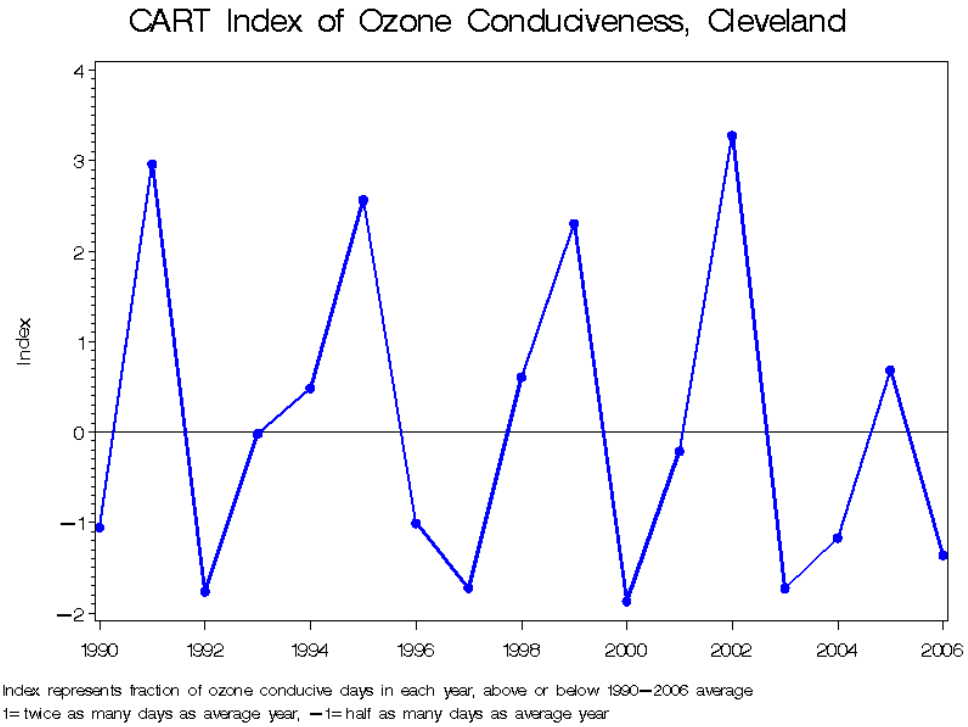
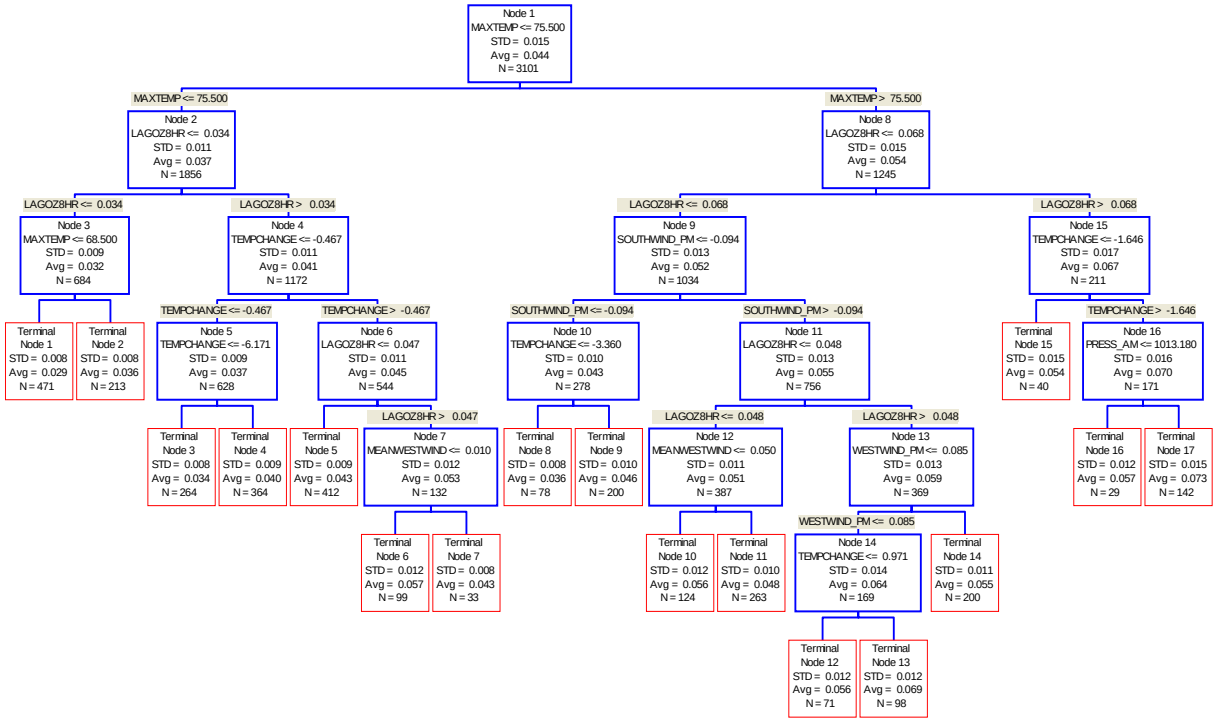
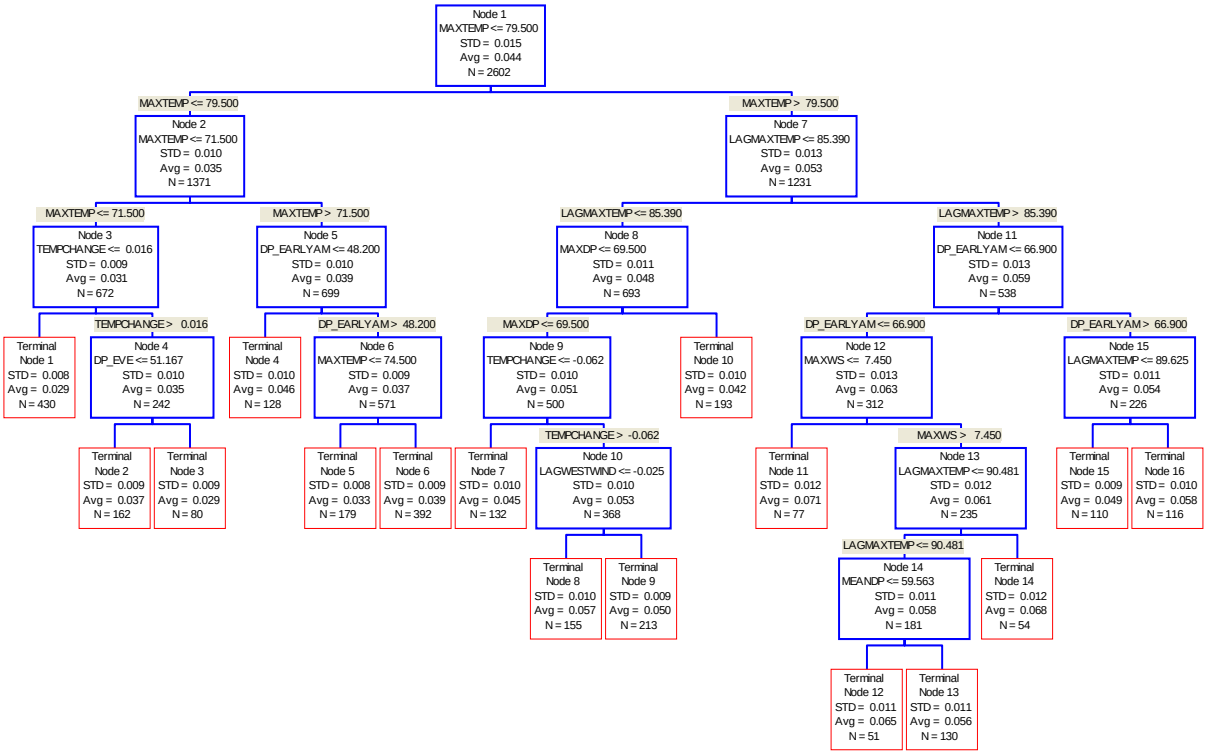


Figure 4. CART Index of Ozone Conduciveness, Cleveland

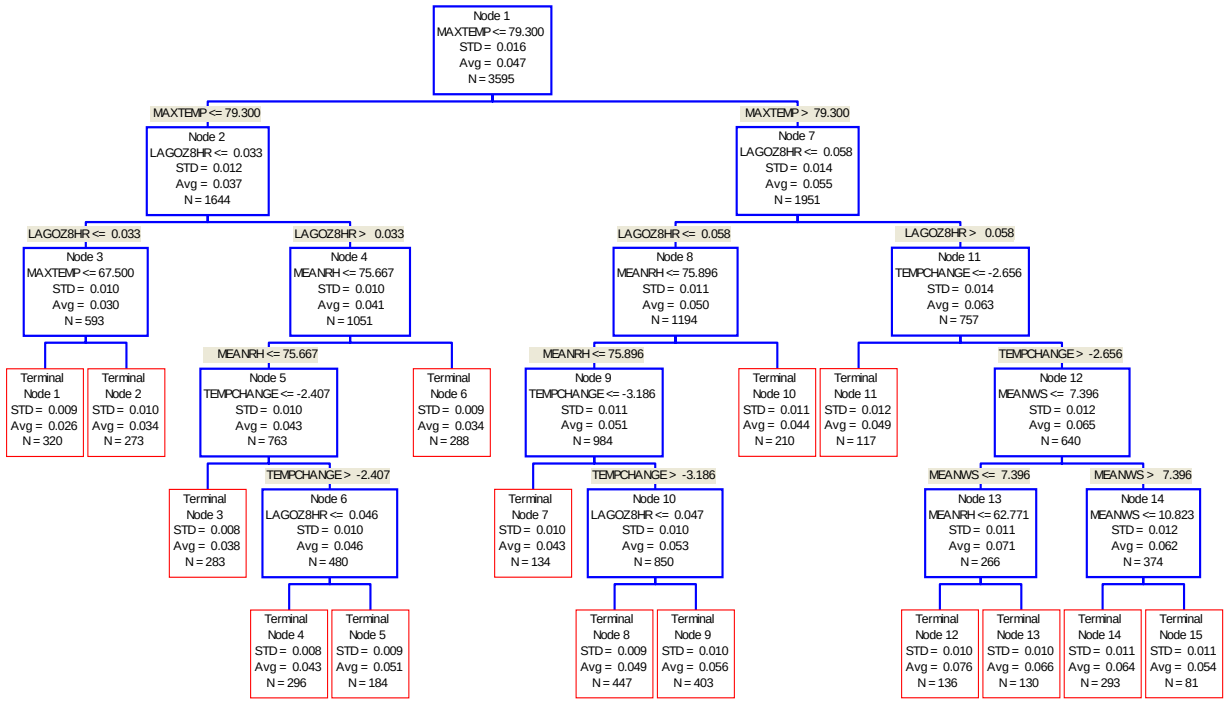
APPENDIX



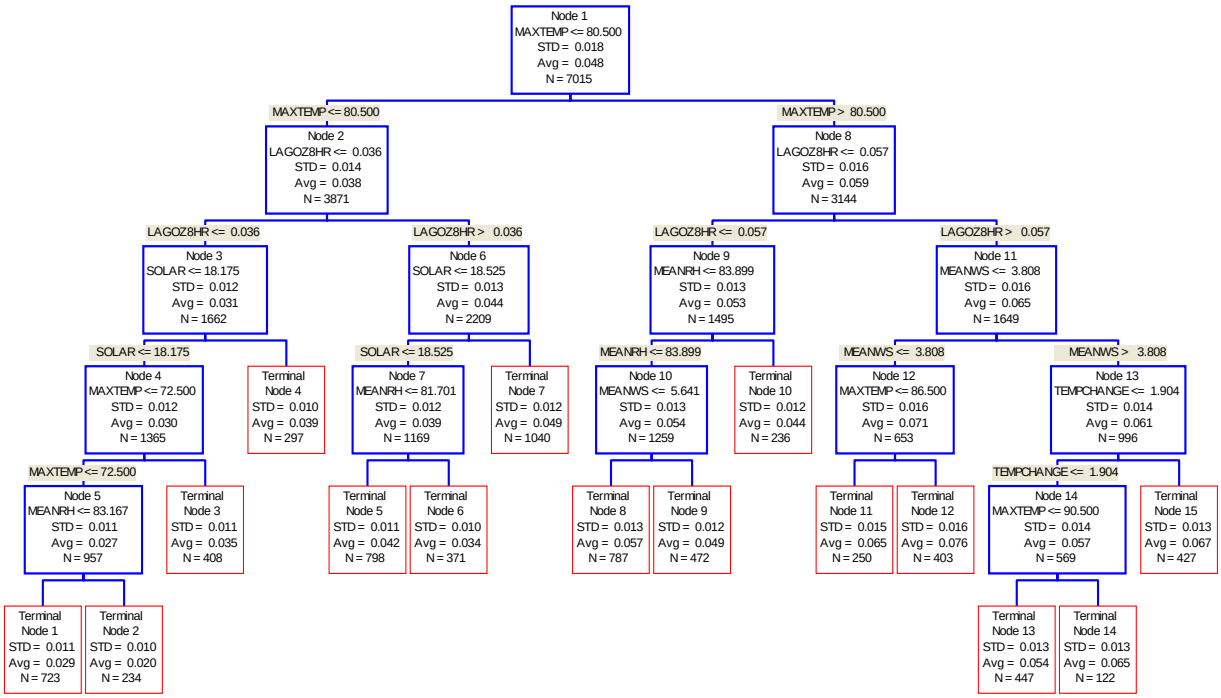
A.1 Regression Tree for Milwaukee



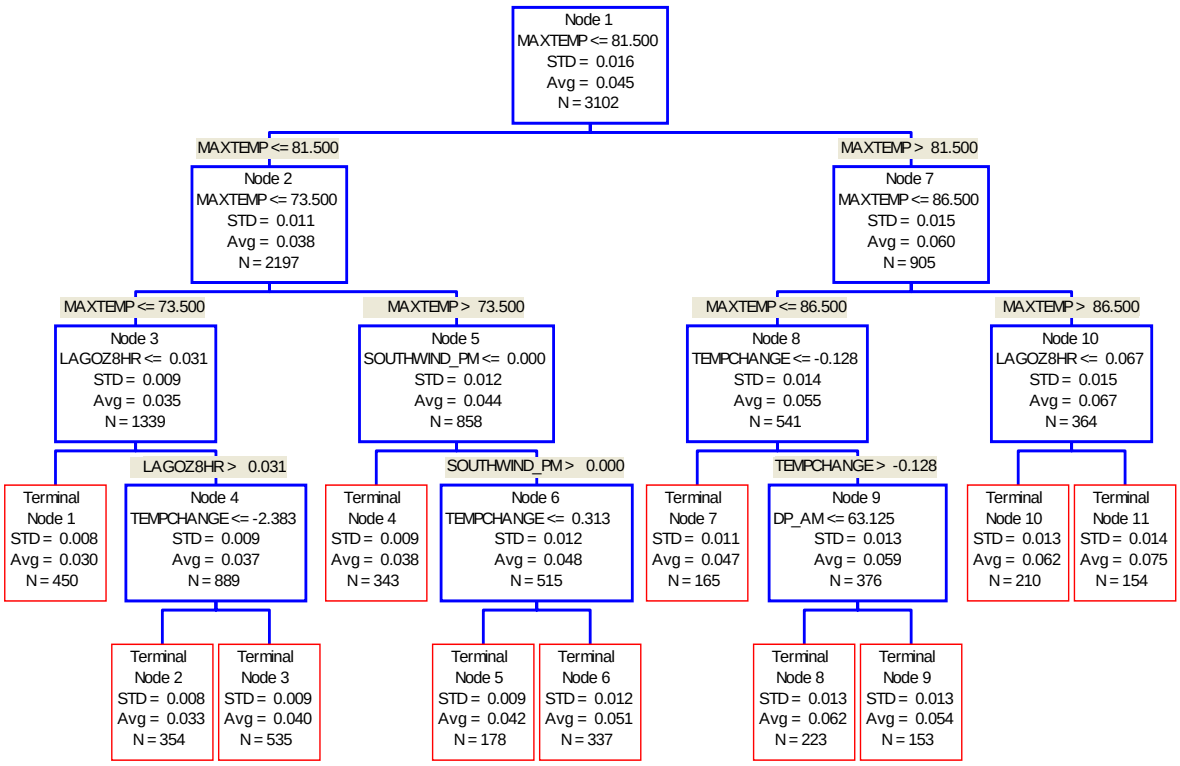
A.2 Regression Tree for Chicago



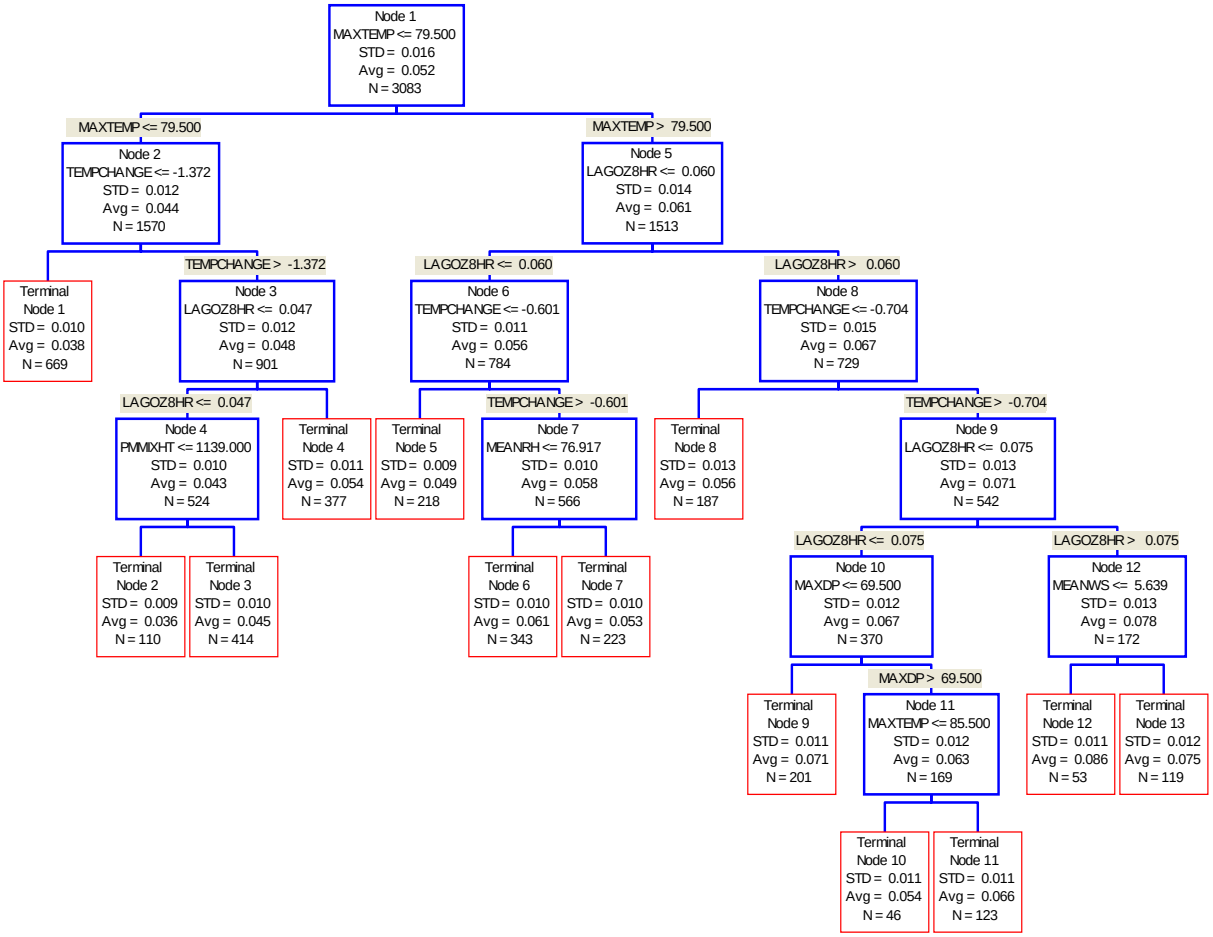
A.3 Regression Tree for St. Louis



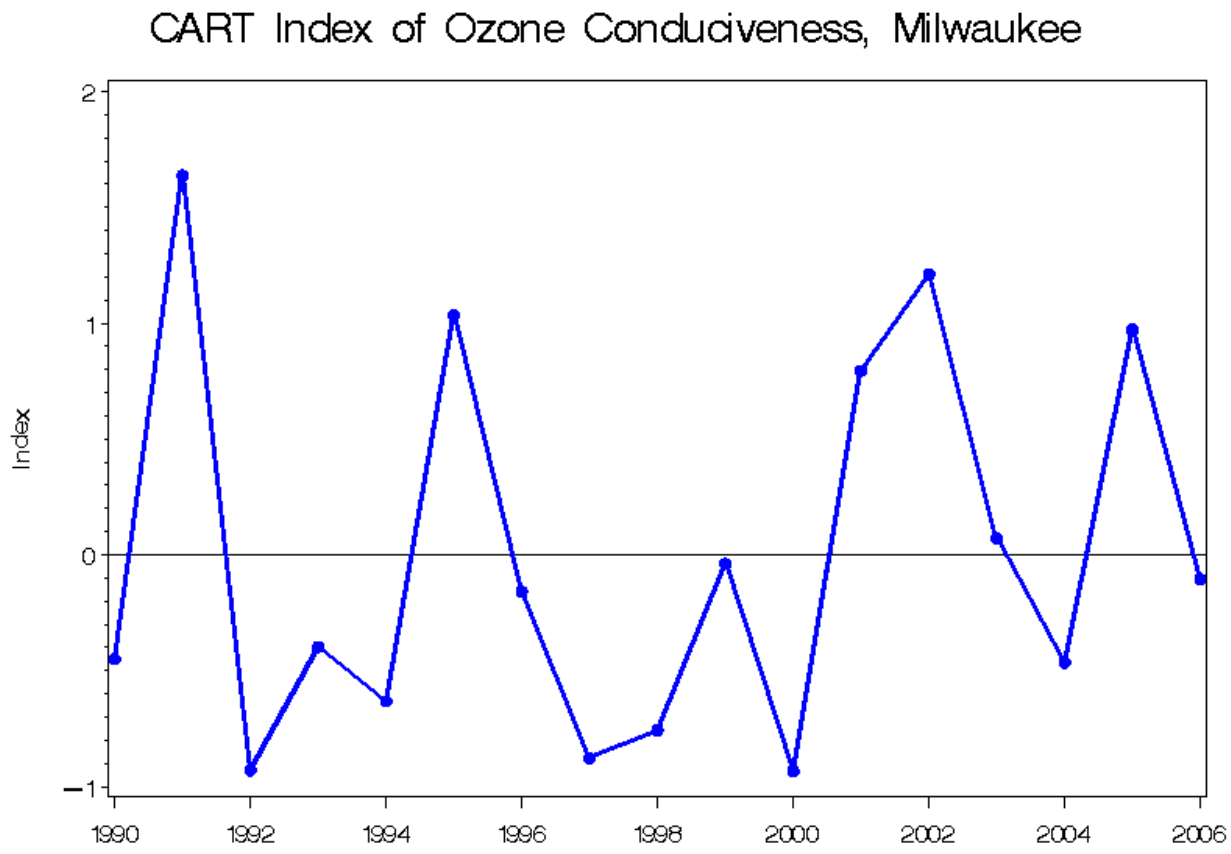
A.4 Regression Tree for Cincinnati



A.5 Regression Tree for Detroit

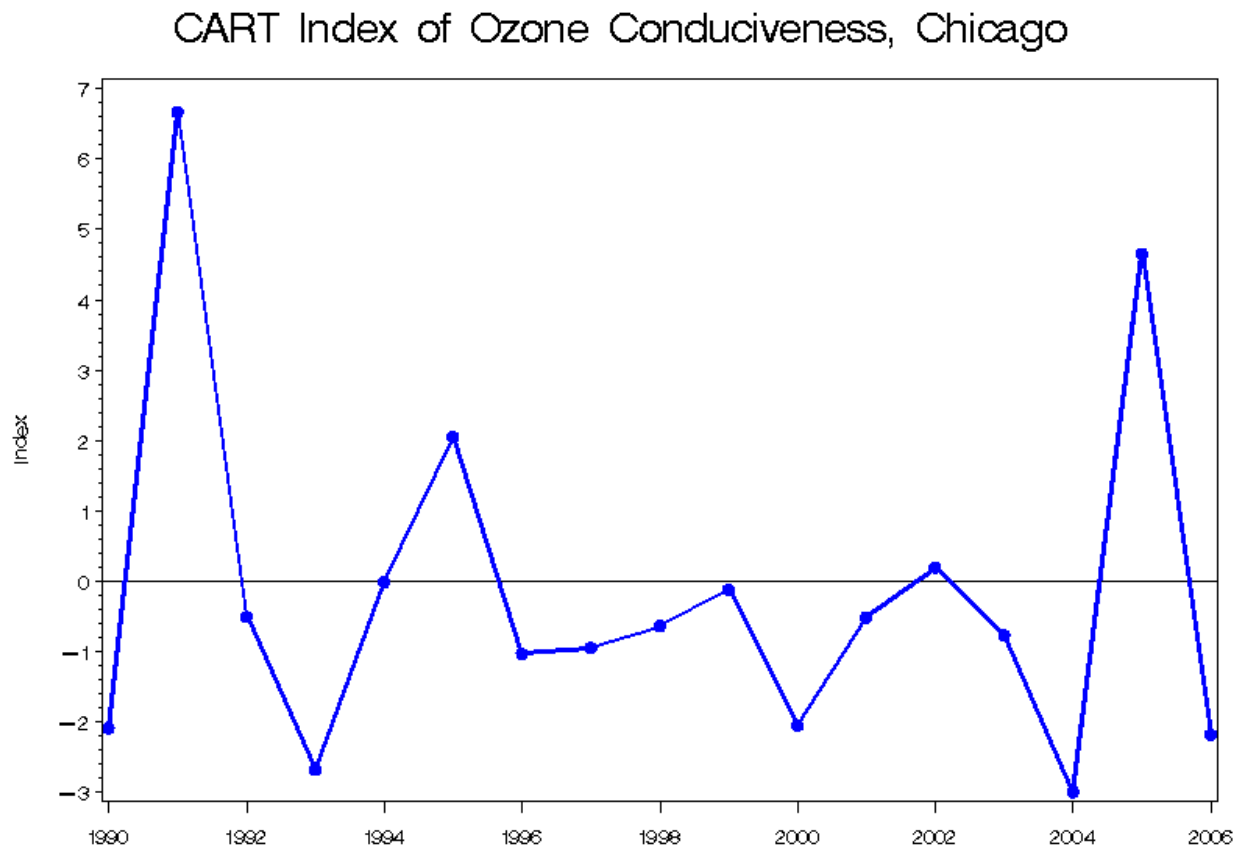


A.6 Regression Tree for Indianapolis



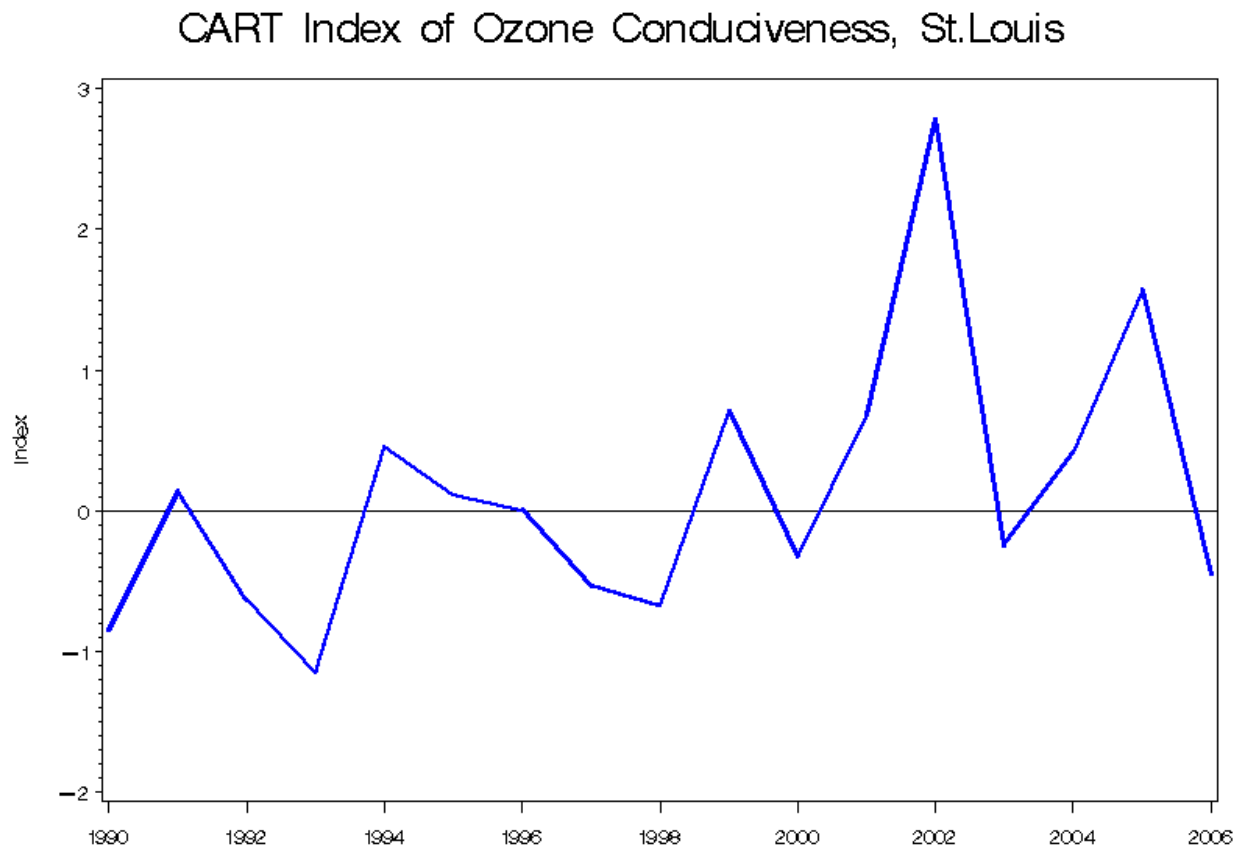
Index represents fraction of ozone conducive days in each year, above or below 1990−2006 average
1= twice as many days as average year, −1= half as many days as average year

A.7 Index of Ozone Conduciveness, Milwaukee



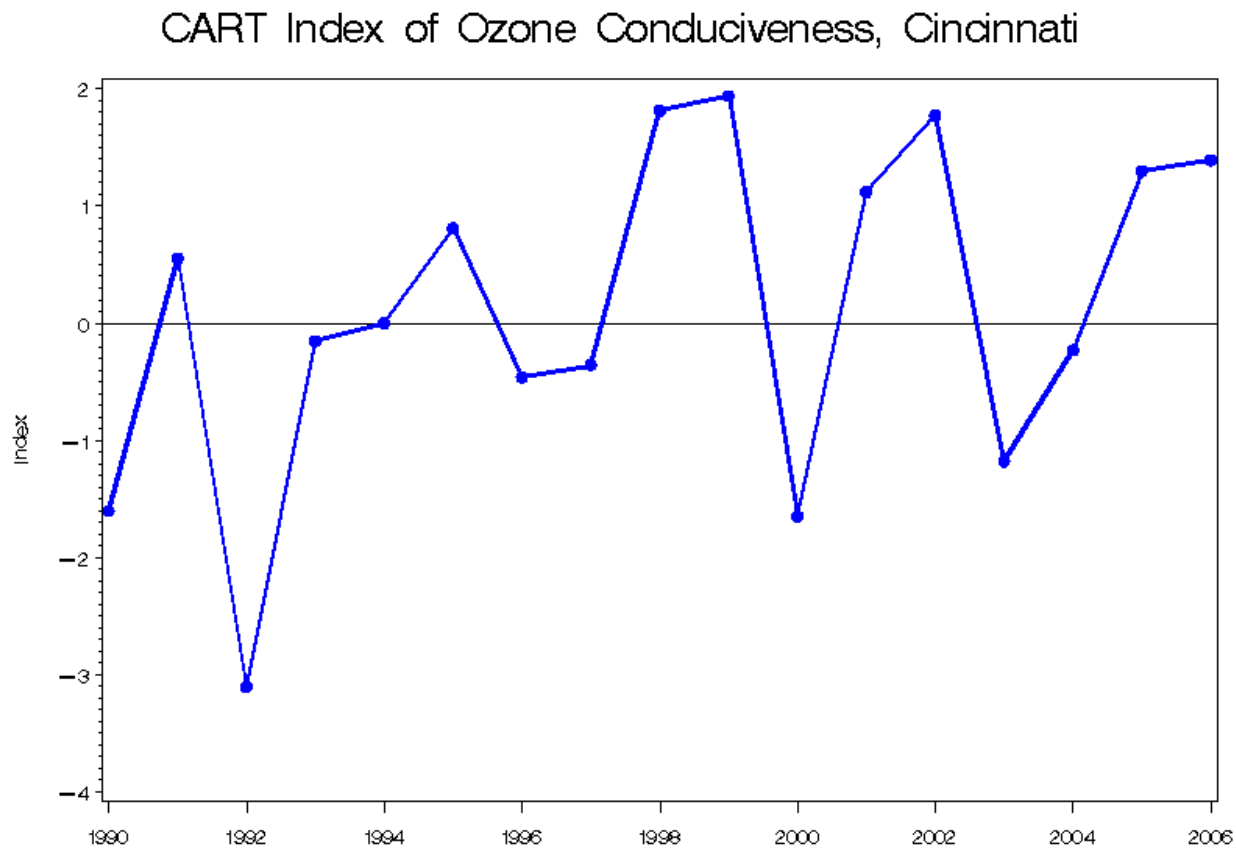
Index represents fraction of ozone conducive days in each year, above or below 1990−2006 average
1= twice as many days as average year, −1= half as many days as average year

A.8 Index of Ozone Conduciveness, Chicago

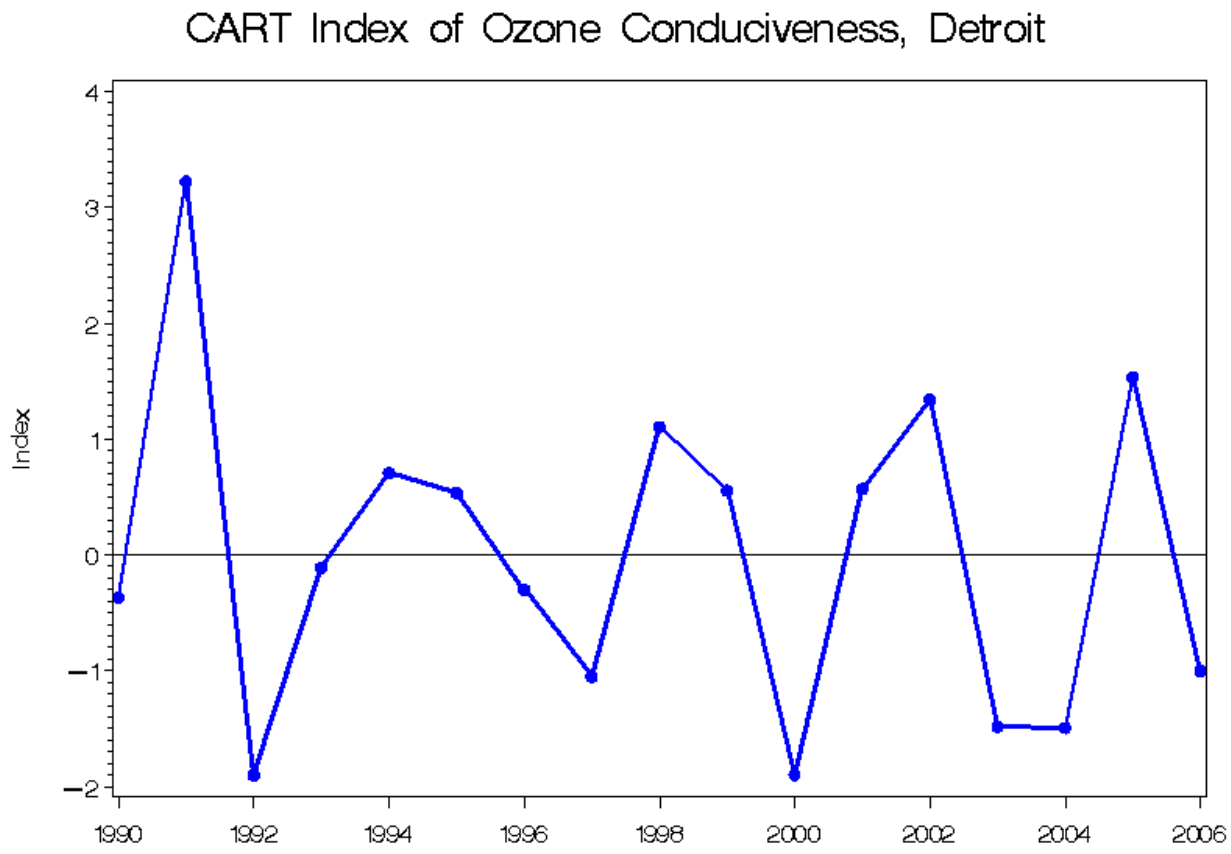


Index represents fraction of ozone conducive days in each year, above or below 1990−2006 average
1= twice as many days as average year, −1= half as many days as average year

A.9 Index of Ozone Conduciveness, St. Louis

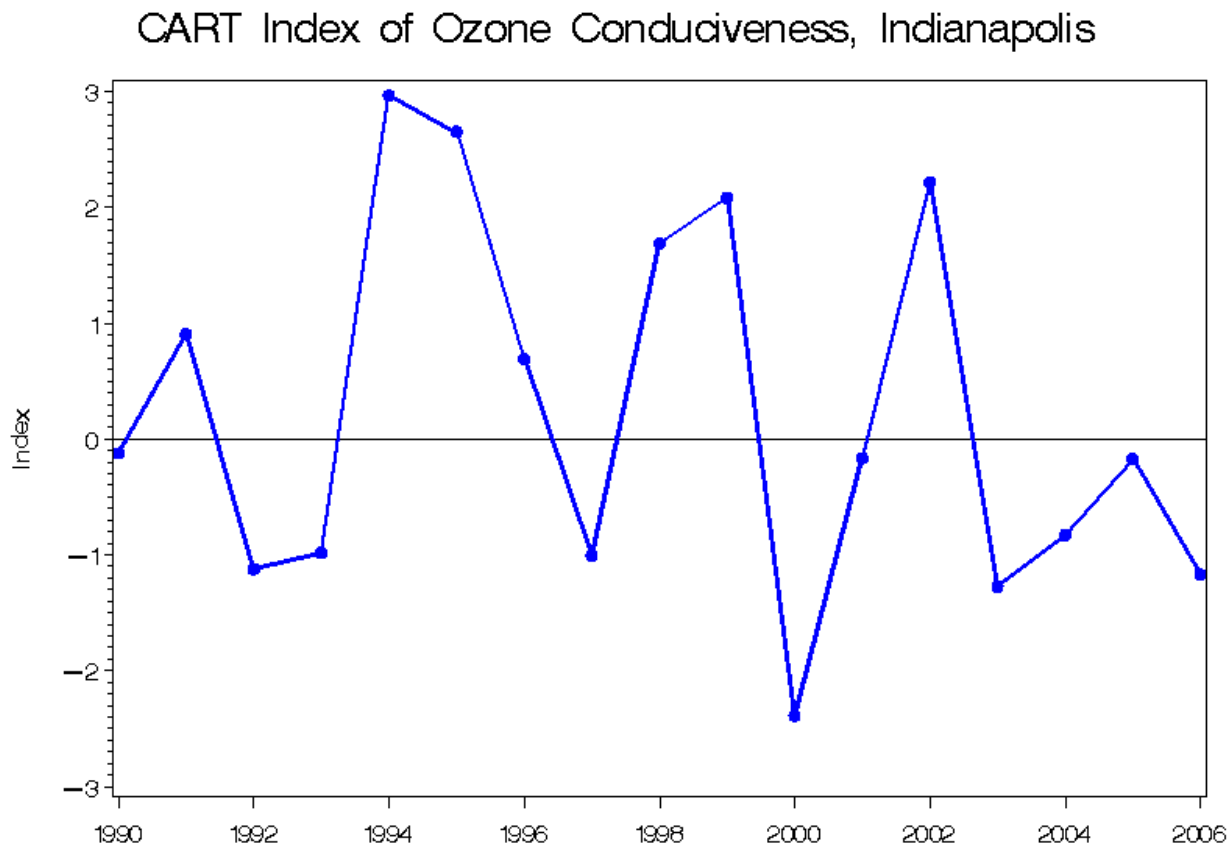


A.10 Index of Ozone Conduciveness, Cincinnati

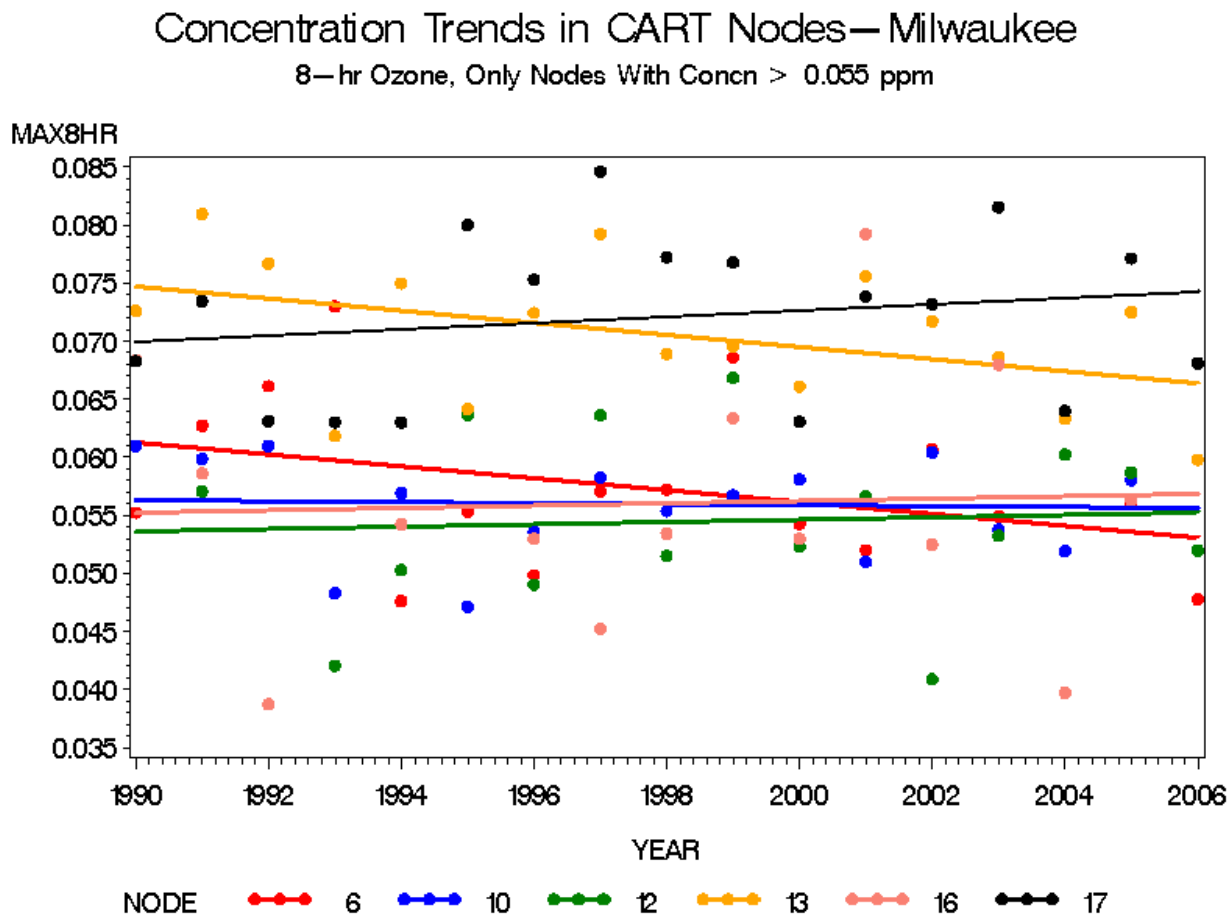


Index represents fraction of ozone conducive days in each year, above or below 1990–2006 average
1= twice as many days as average year, −1= half as many days as average year

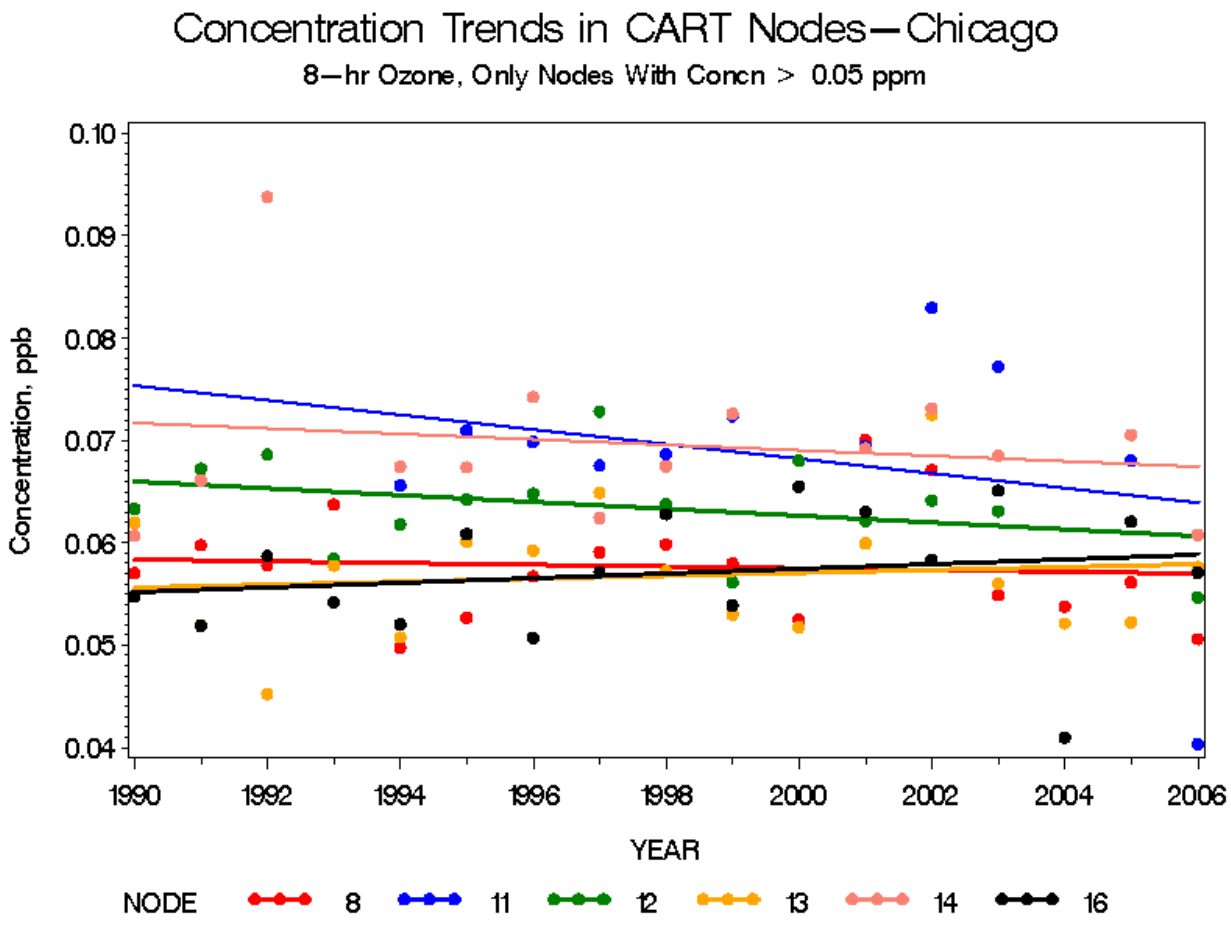
A.11 Index of Ozone Conduciveness, Detroit



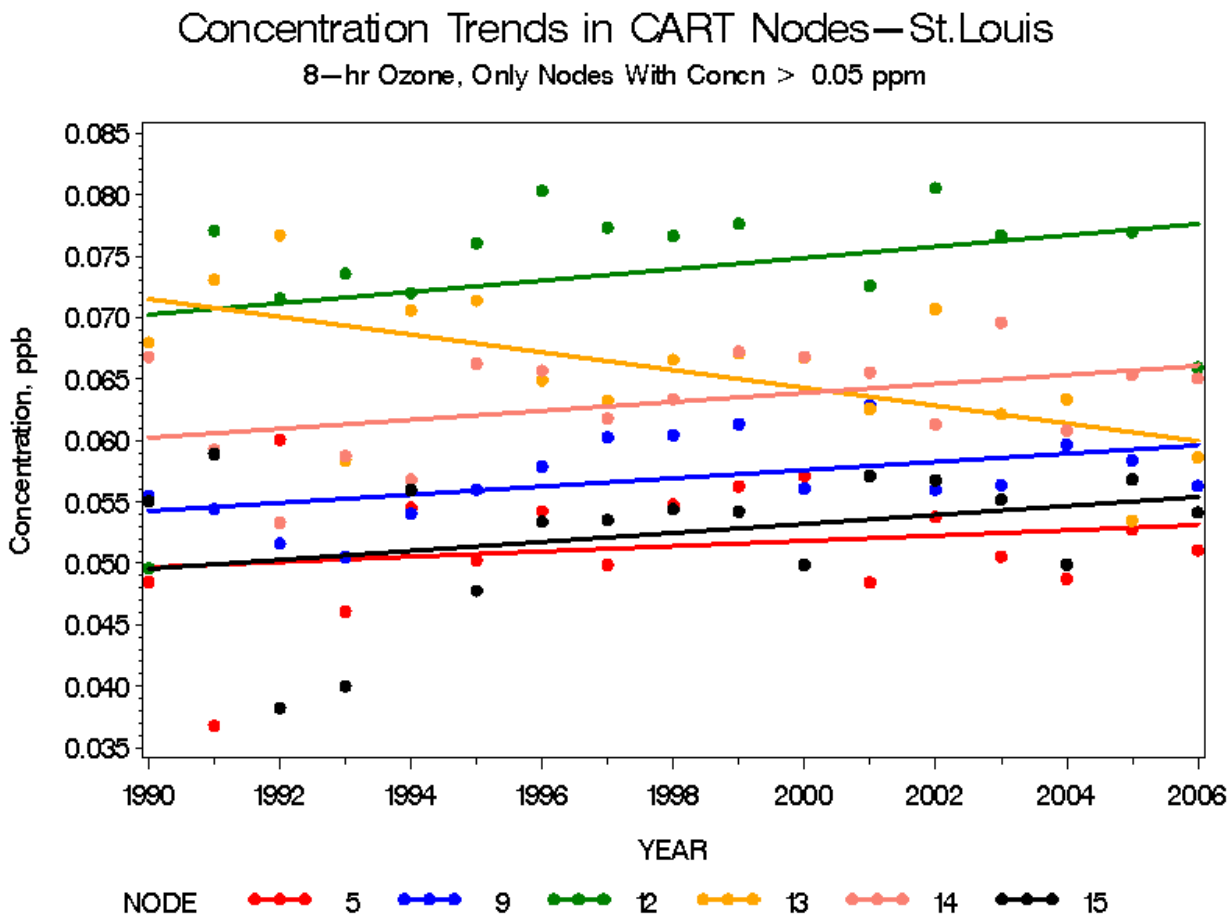
A.12 Index of Ozone Conduciveness, Indianapolis



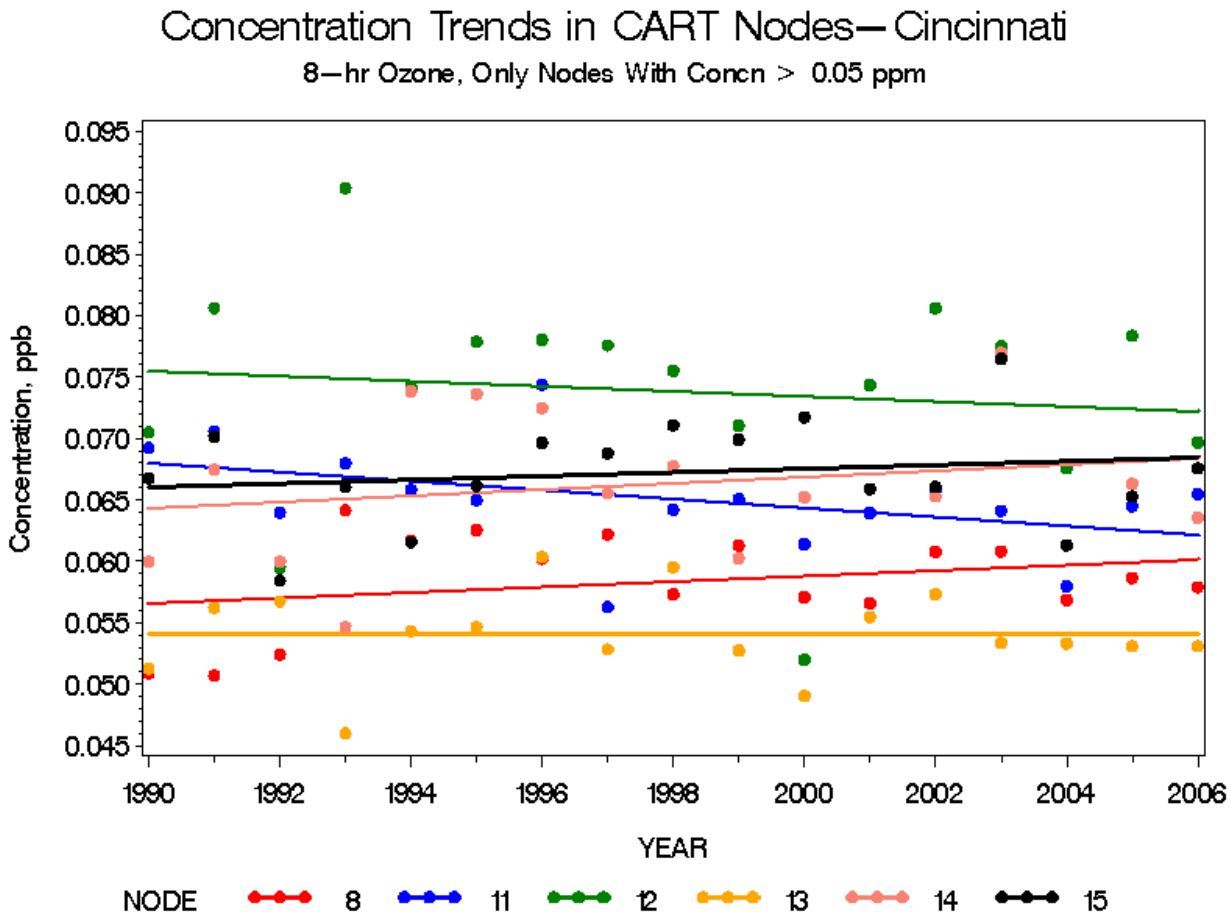
A.13 Concentration Trends in CART Nodes--Milwaukee



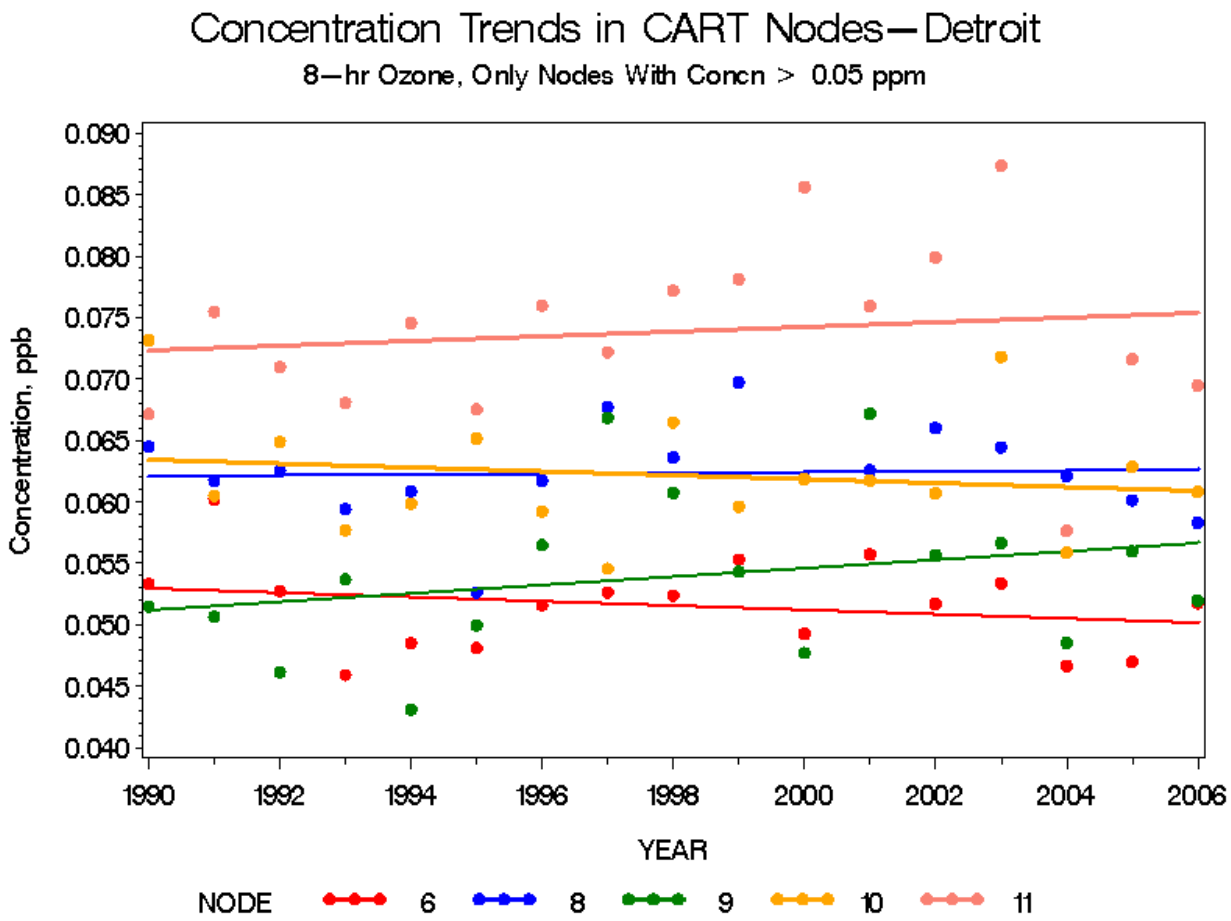
A.14 Concentration Trends in CART Nodes—Chicago



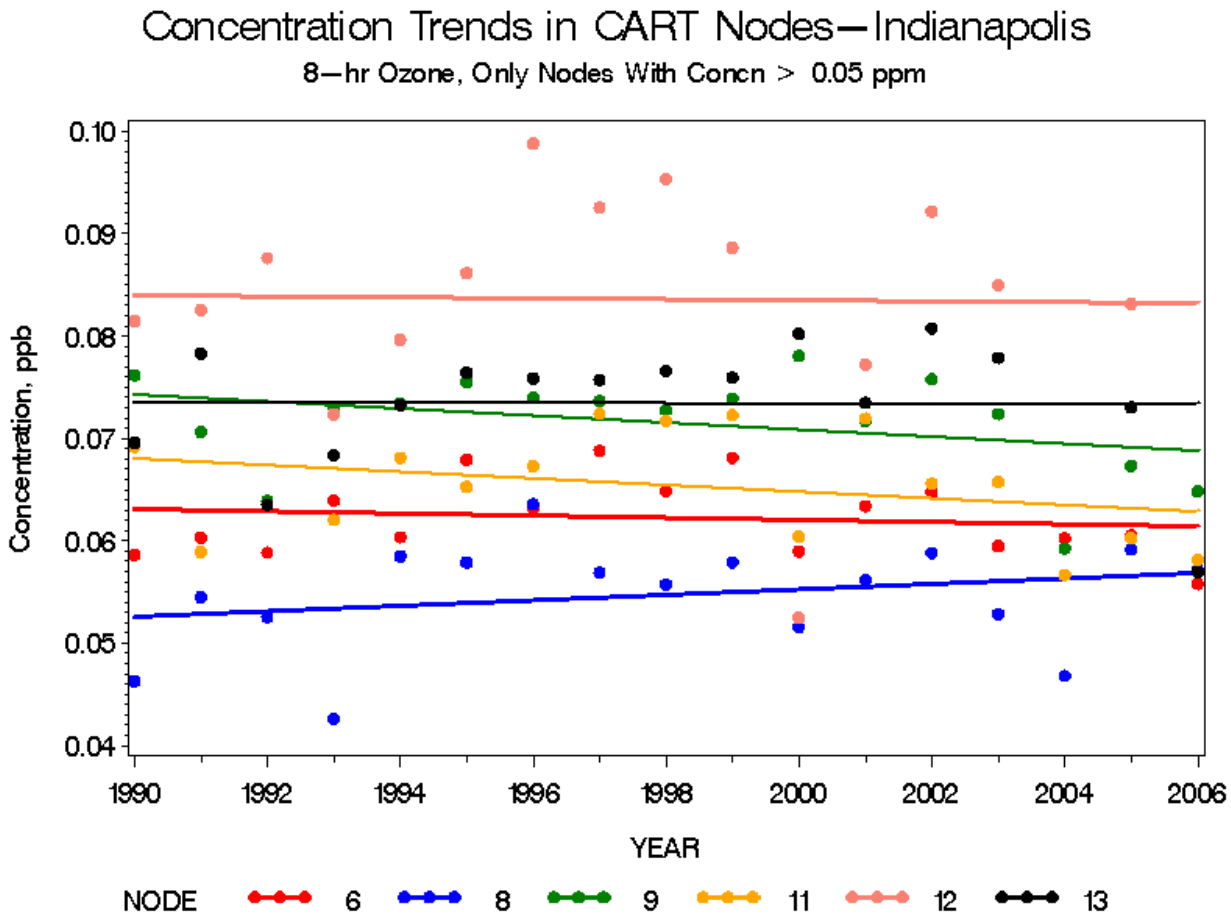
A.15 Concentration Trends in CART Nodes—St. Louis



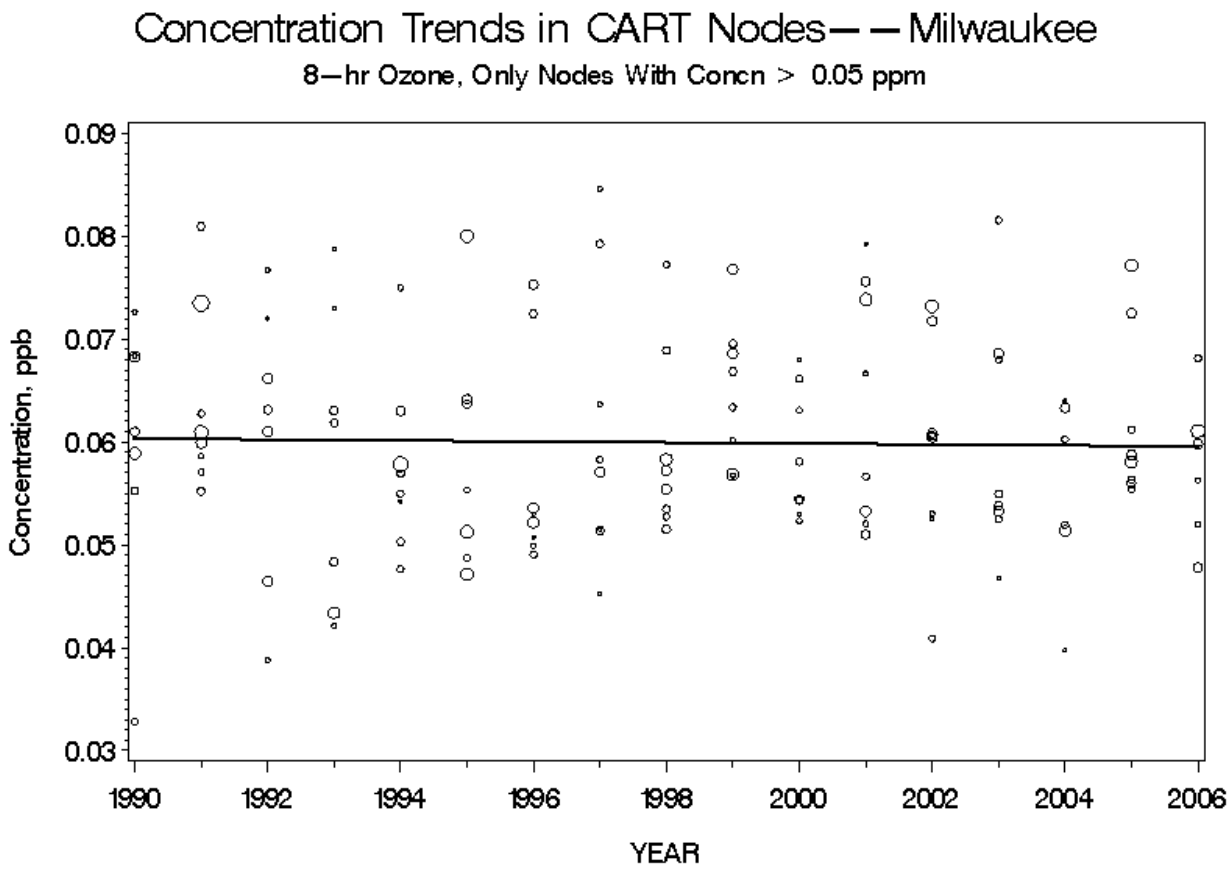
A.16 Concentration Trends in CART Nodes—Cincinnati



A. 17 Concentration Trends in CART Nodes--Cetroit

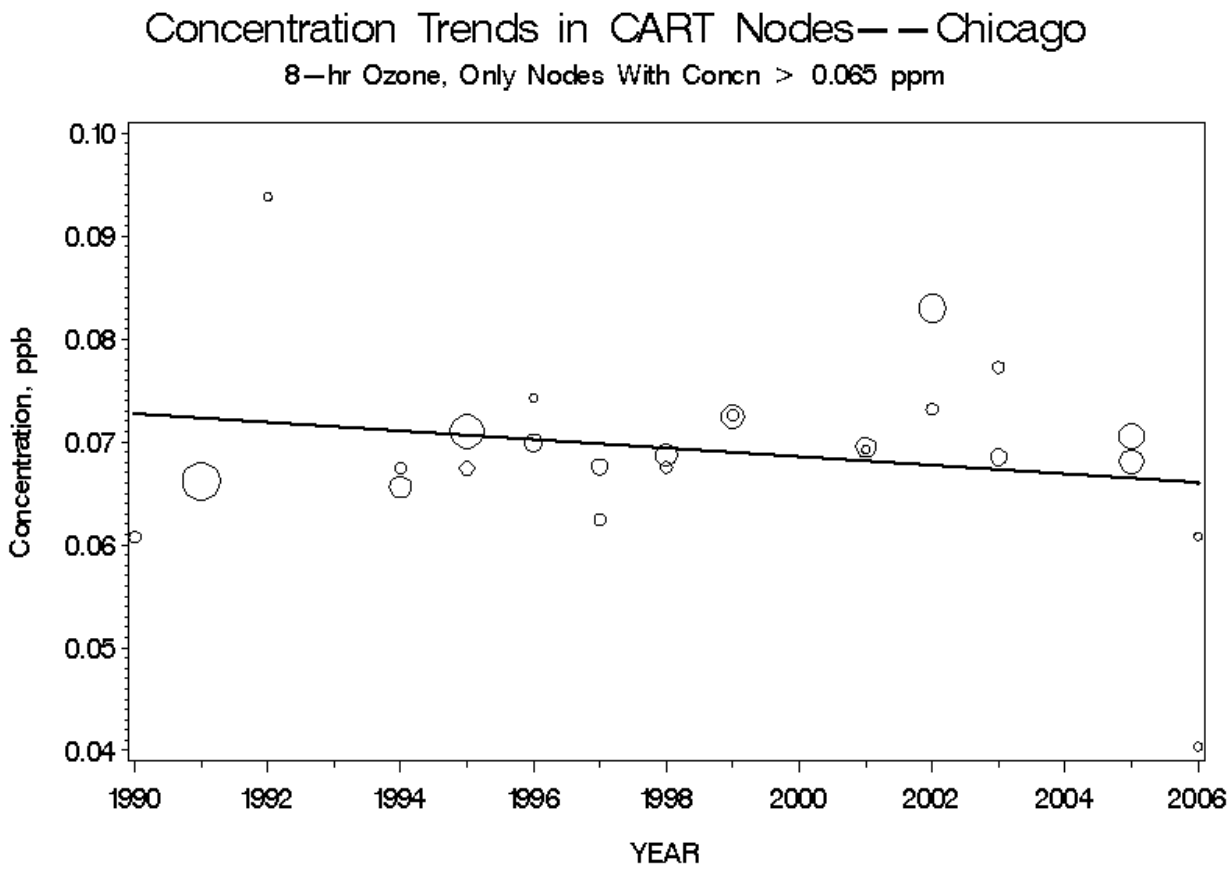


A.18 Concentration Trends in CART Nodes--Indianapolis



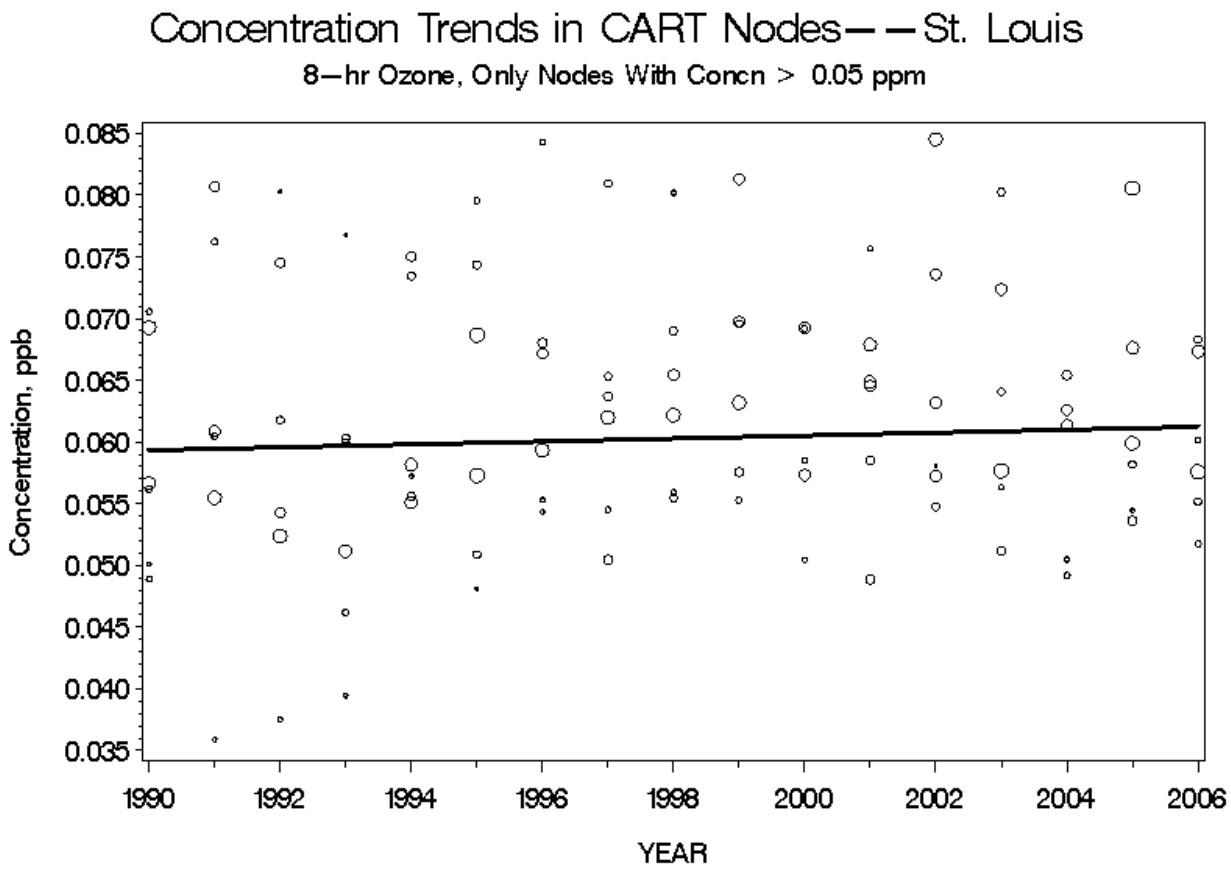
Size of bubble is proportional to number of days in node.

A.19 Overall Concentration Trend, High Ozone Nodes, Milwaukee



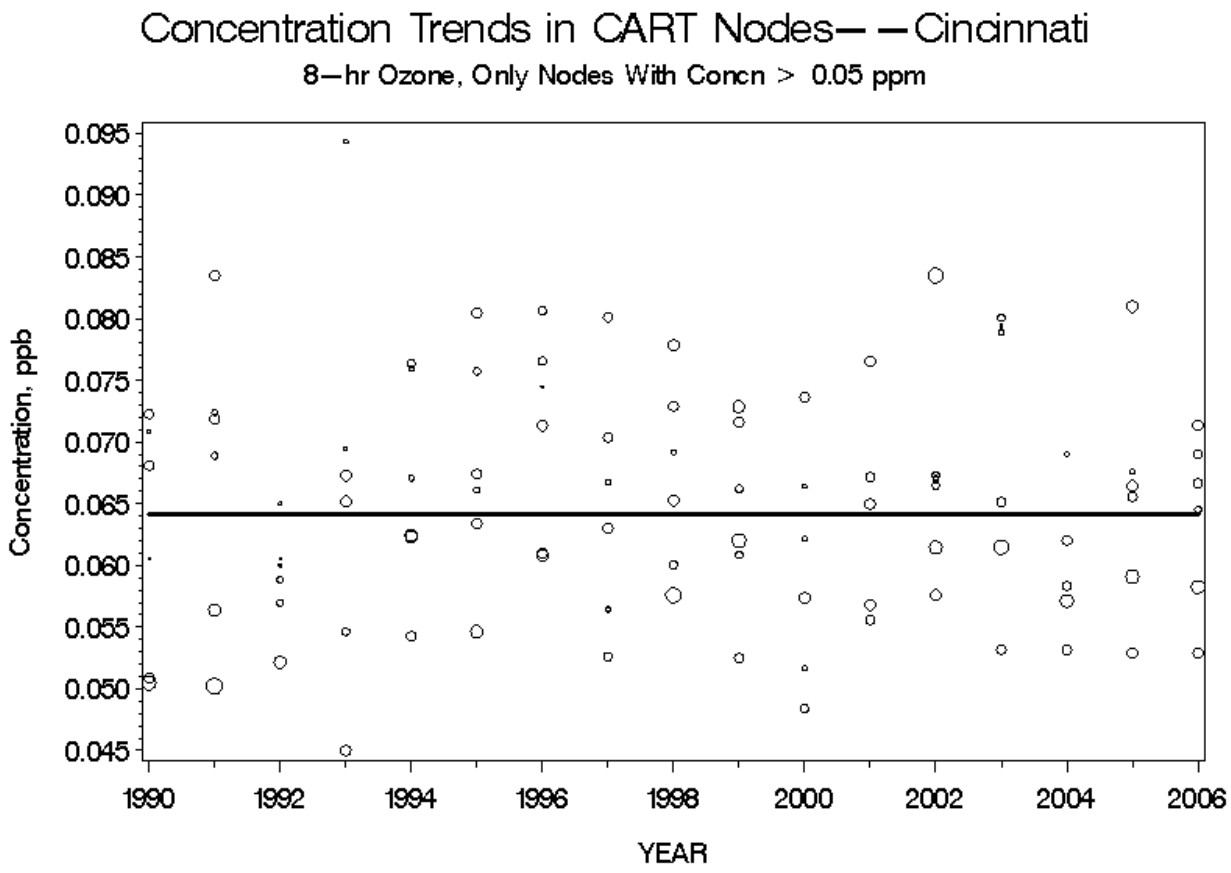
Size of bubble is proportional to number of days in node.

A.20 Overall Concentration Trend, High Ozone Nodes, Chicago

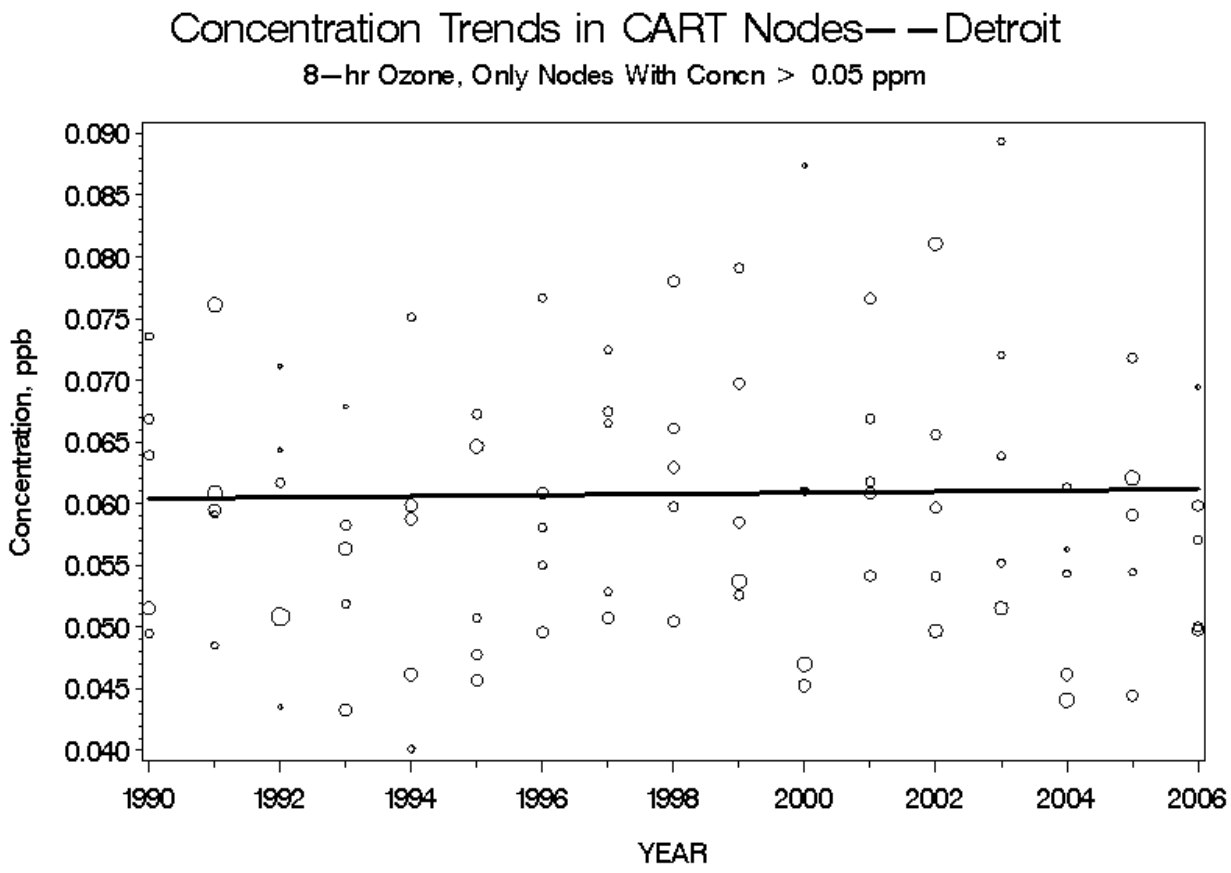


Size of bubble is proportional to number of days in node.

A.21 Overall Concentration Trend, High Ozone Nodes, St. Louis

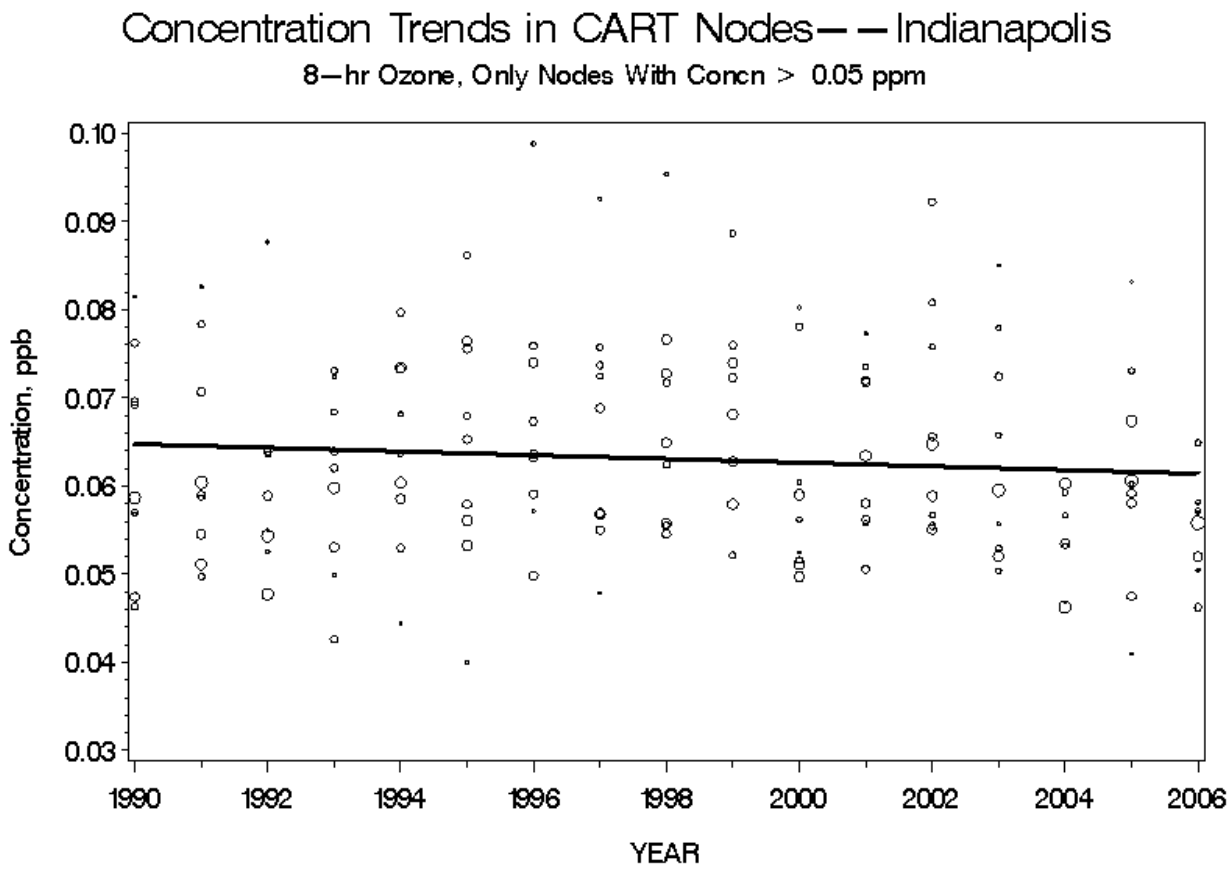


A.21 Overall Concentration Trend, High Ozone Nodes, Cincinnati



Size of bubble is proportional to number of days in node.

A.22 Overall Concentration Trend, High Ozone Nodes, Detroit



Size of bubble is proportional to number of days in node.

A.23 Overall Concentration Trend, High Ozone Nodes, Indianapolis