Air Quality Project

By: Zhenghan Li, Yichen Wang

The air quality is one of the issues that captures our concerns. With the rapid development of technology innovation, the side effect becomes more and more serious: the air pollution becomes a problem to humans. To reach the goal of sustainability, solving air pollution is the essential step for humans.

However, America’s air quality seems to increase since 2016, just like what Nadja said in the article *America’s Air Quality Worsens,Ending Years of Gains, Study Says*: “New data reveals that damaging air pollution has increased nationally since 2016, reversing a decades-long trend toward cleaner air.” But Nadja just analyzed the data about PM2.5 to simply reach this result, which is quite hasty. Nowadays, there are always some misunderstandings about the equation of PM2.5 and air quality. Many people consider that the lower the PM2.5 index is, the higher the air quality, which is quite incomprehensible. So we try to fix this misunderstanding and tell people which is the real main pollutant to different counties in America and predict the trend of the pollutants.

Most people think that the main cause effect is some tiny particles like PM2.5, but nobody knows what the certain places is suffering what kind of air pollution, which means we can’t know the targeted way to improve the air quality specifically. So we write some programs to first analyze the data with AQI to find out the counties which are suffering from serious air pollution in America. Then, by analyzing the statistics about the day when the main pollutant is PM25, PM10, NO2, SO2, OZONE and so on to find out which pollutant results in air pollution most in these counties, and then we can use specific ways to govern the specific pollutant. And then the people in these seriously-polluted counties may not suffer from the bad air any more.

Also,using some statistics from 2008 to 2018 makes it possible to predict the trend of such pollutants concentration, which can help to remind us to govern them in advance.

The major reason to measure the pollutant concentration is because a high concentration of pollutant can become extremely hazardous to people with asthma, heart disease, and etc. So A goal for this project will be to use existing data and to predict where and when a high Air Quality Index will occur, or a high concentration of pollutant. Thus enabling people to get advance notice on whether or not they should leave their home.

Link to data download: <https://aqs.epa.gov/aqsweb/airdata/download_files.html>

Article Cited:

# America’s Air Quality Worsens,

# Ending Years of Gains, Study Says

By [Nadja Popovich](https://www.nytimes.com/by/nadja-popovich)Oct. 24, 2019

New data reveals that damaging air pollution has increased nationally since 2016, reversing a decades-long trend toward cleaner air.

An analysis of Environmental Protection Agency data published this week by researchers at Carnegie Mellon University found that fine particulate pollution increased 5.5 percent on average across the country between 2016 and 2018, after decreasing nearly 25 percent over the previous seven years.

“After a decade or so of reductions,” said Nick Muller, a professor of economics, engineering and public policy at Carnegie Mellon, and one of the study’s co-authors, “this increase is a real about-face.”

The research identified recent increases in driving and the burning of natural gas as likely contributors to the uptick in unhealthy air, even as coal use and related pollution have declined. In the West, wildfires contributed to the rise in particulate matter.

Researchers also suggested that a decrease in enforcement of the Clean Air Act may have contributed to the recent rise in pollution. That law and its subsequent updates put in place strict air pollution standards for power plants, factories, vehicles and other sources, and has been credited with dramatically improving air quality across the country and saving hundreds of thousands of lives.

The new analysis estimated that the increase of slightly more than 5 percent in fine particulate pollution nationwide between 2016 and 2018 was associated with nearly 10,000 additional premature deaths during that time.

Fine particulate pollution – known as PM2.5 because the particles are less than 2.5 micrometers in diameter, or one-thirtieth the size of a human hair – has been linked to a range of health problems including asthma and respiratory inflammation, lung cancer, heart attack and stroke. A recent study found a significant link between air pollution and the risk of miscarriage.

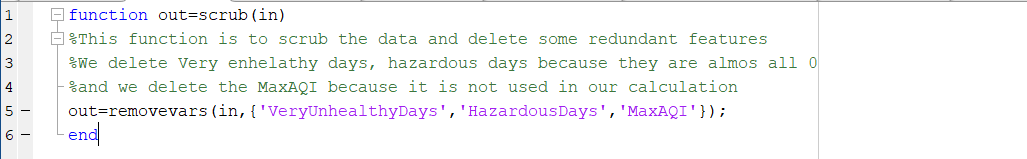
Articles from: <https://www.nytimes.com/interactive/2019/10/24/climate/air-pollution-increase.html>

Synopsis:

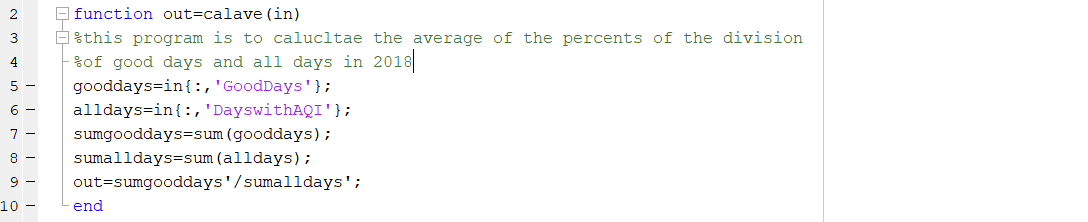
I strongly agree with the article that air quality worsens in America, but I think the author’s way to reach this conclusion is not accurate enough. There are a lot of pollutants may influence the AQI, but the author just measures the concentration of PM2.5 and before reaching a conclusion, which may mislead the readers that the lower the PM2.5, the higher the air quality is. And just reaching the conclusion can’t solve anything; we should know which pollutant is needed to be governed in certain counties where we should take actions, and that’s what we want to do in our project.

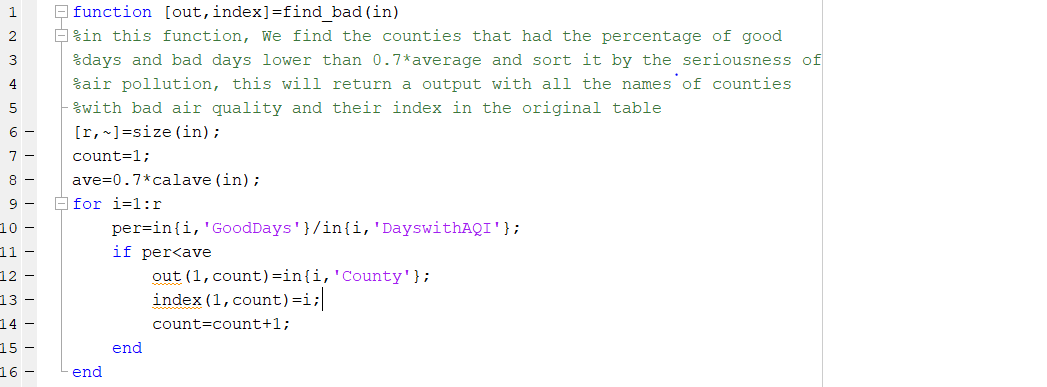
1. The first part is to first analyze the data of 2018 to figure out the current situation of air quality of most counties in America.

There are some important features in the data. Dayswithaqi represents the days when AQI is recorded. The Gooddays represent the days when air quality is good. And the columns with days plus particles like Daysco, DaysNo2 represent the days when these certain pollutants are the main pollutants.

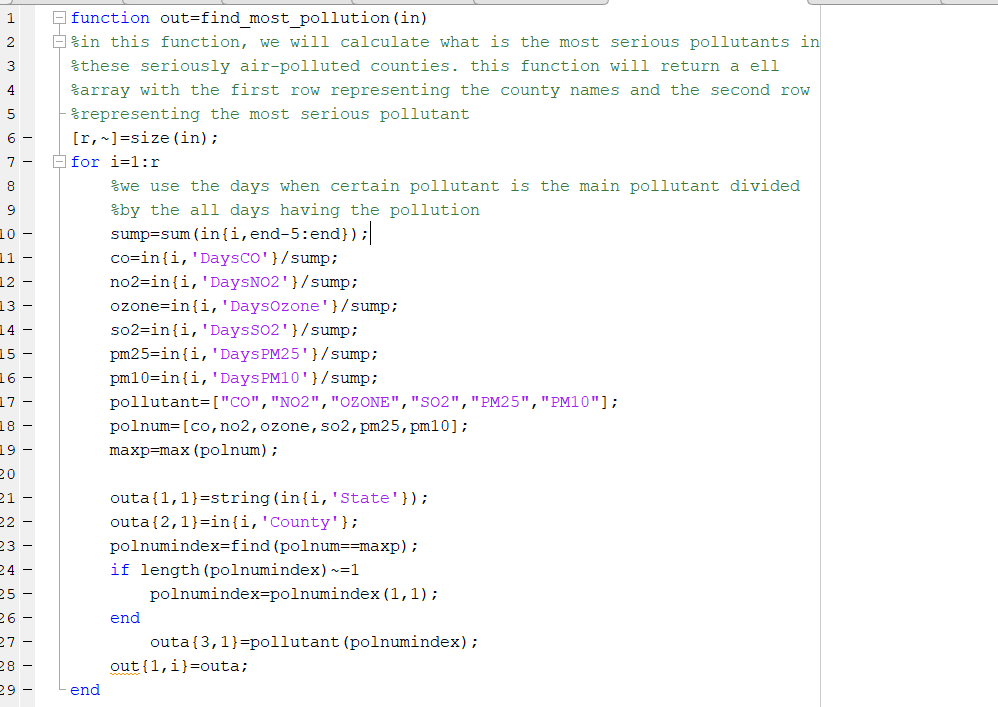
1. The first step we do is to scrub the data.(We will have another complex scrub with 10 files of statistics later)

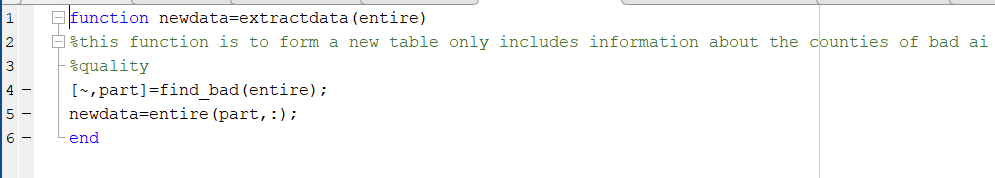
We delete the columns of very unhealthy days and hazardous days because almost all of them are zero, which is quite a good thing. And the MaxAQI column is deleted because there are too many contingencies when just considering the MaxAQI as the degree of air quality pollution.

1. The second step is to calculate the average of the division of gooddays and dayswithAQI, which represents the average degree of air pollution in America.Basically, we use days to test the degree of air quality.
2. After getting the average, we want to find out the certain counties which has lower number of the division of gooddays and dayswithaqi, which means that some enforcement or laws are needed to improve air quality.



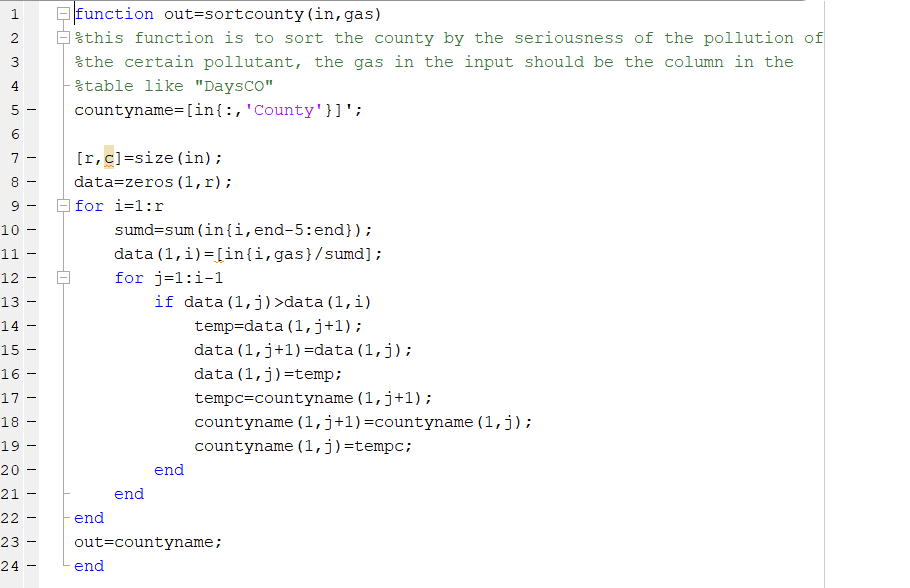
1. Then we want to find out in all these seriously-polluted counties, which pollutant results in the bad air quality most, and then these certain pollutants can be resolved more efficiently.





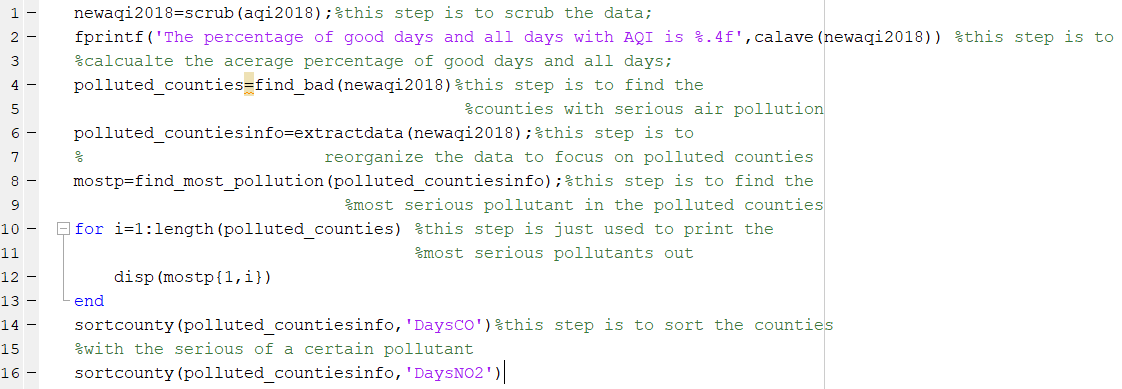
(This function is to form a new table which only includes the seriously-polluted counties)

1. Then we use these data to sort the counties according to their seriousness of certain particle/gas pollution.



(For example, if this function receives a table of air quality and PM25, it will then sort all the counties by the seriousness of PM25 pollution from high to low)

1. And we put all of these functions together to see the outcomes.



Outcome:

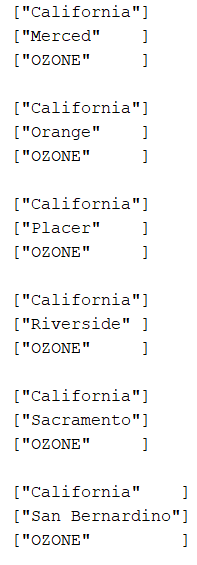
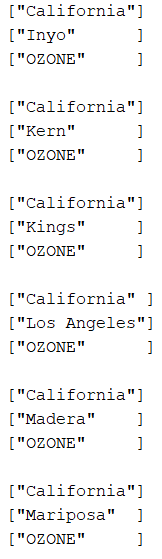
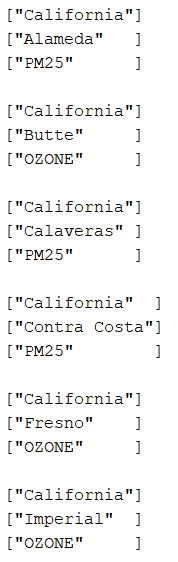
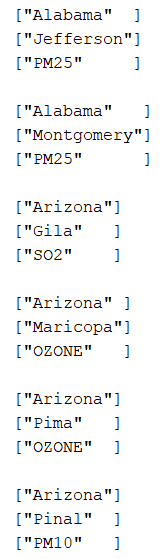
1. 

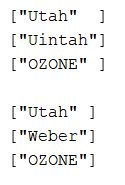
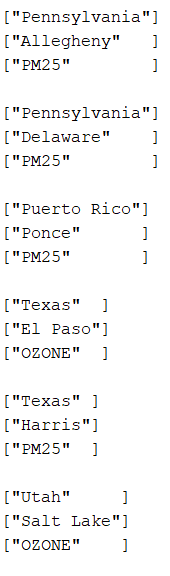
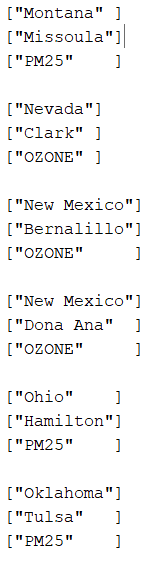
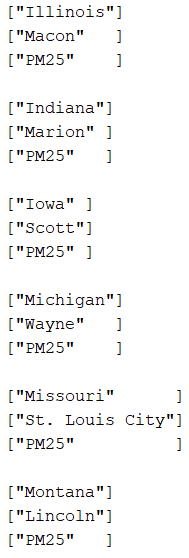
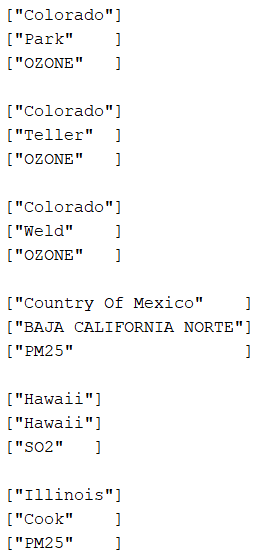
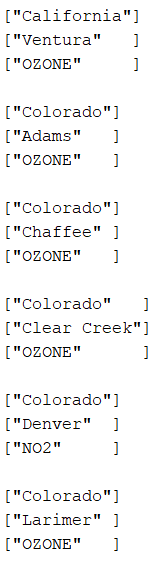
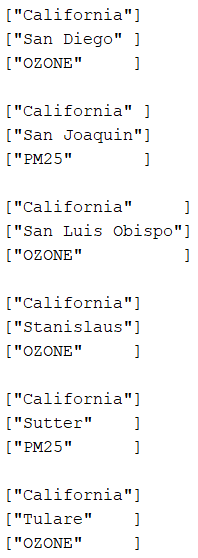
Which means in America in 2018, 80 percent of days have good air quality.

1. This output shows all of the counties which have the division of gooddays and dayswithaqi less than 0.7\*average.



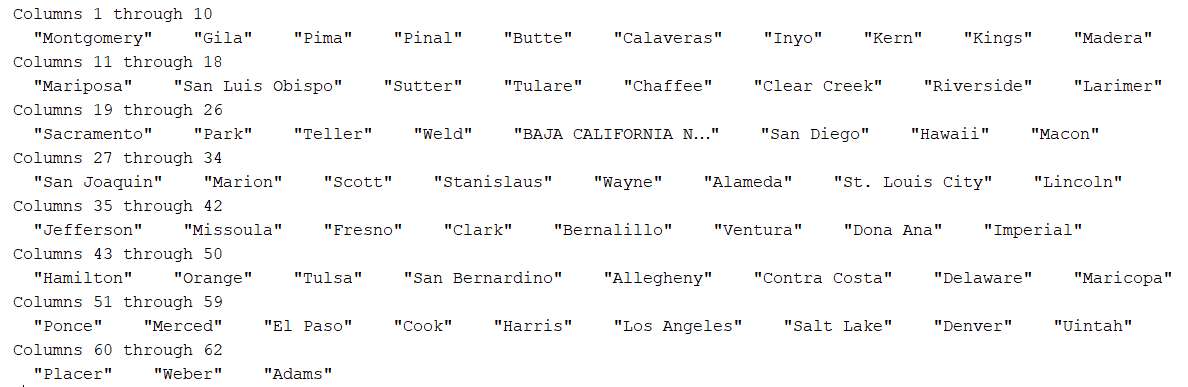
1. This shows the state, the counties, and the most serious pollution needed to be taken actions.





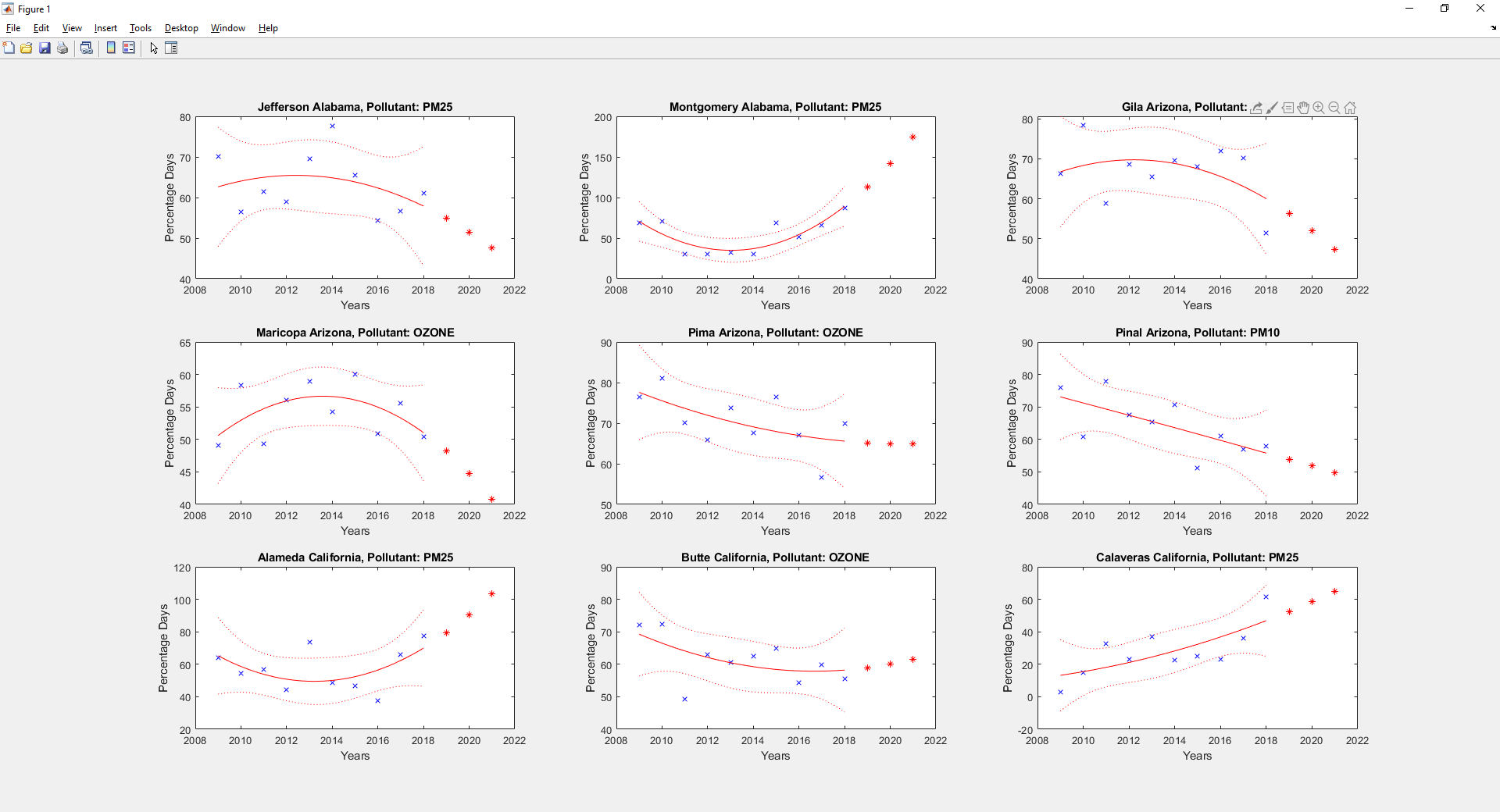
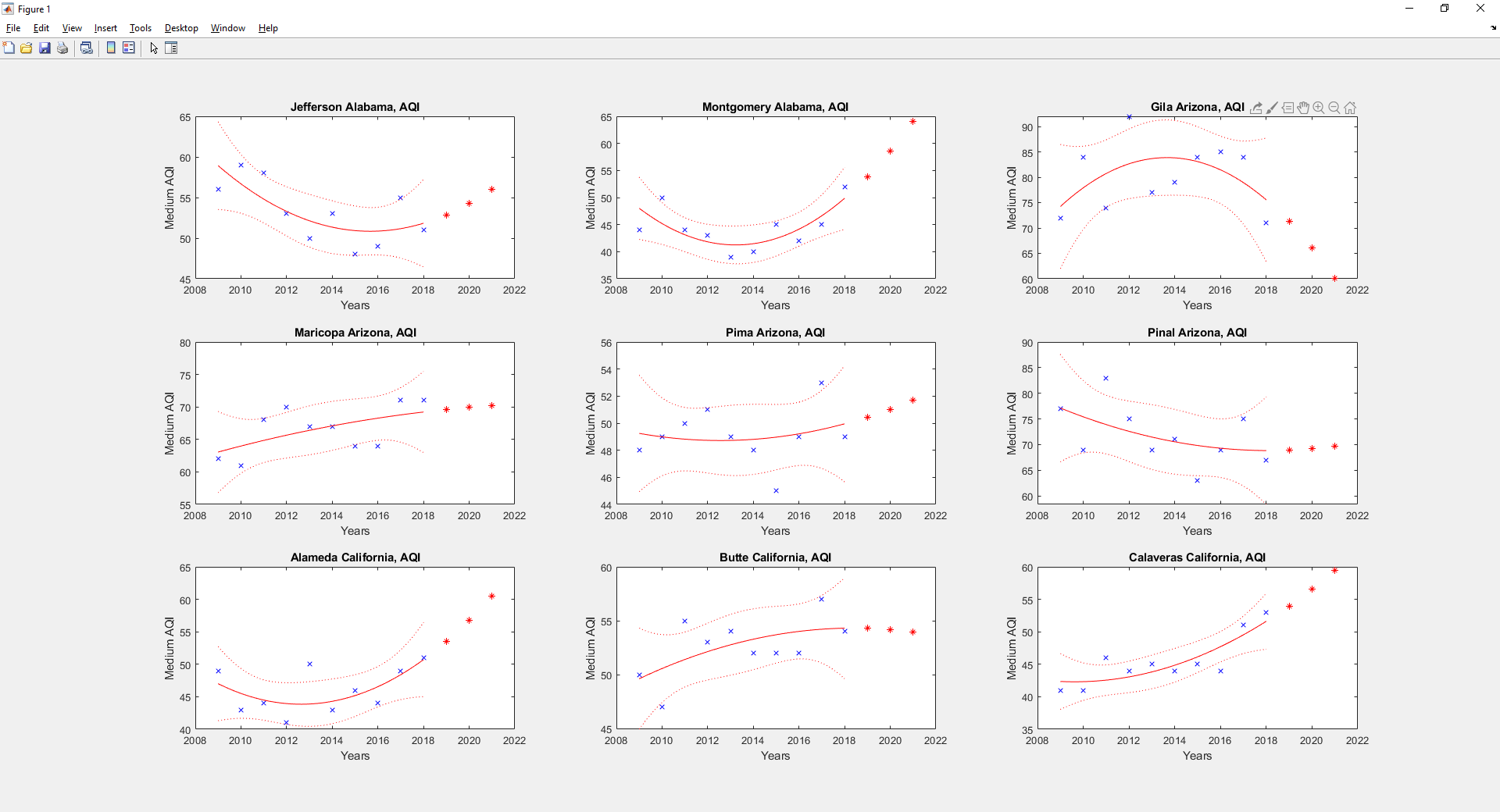
1. In these outputs, it sorts the counties by the seriousness of CO and NO2 from high to low.





f. So as we can see, counties in California struggle a lot with the bad air qualities, and the most serious pollutant is not the PM25, but the ozone does.

B. The second part is using the data from 2008 to 2018 to predict the trend of the seriousness of the certain gas pollution in seriously-polluted counties.





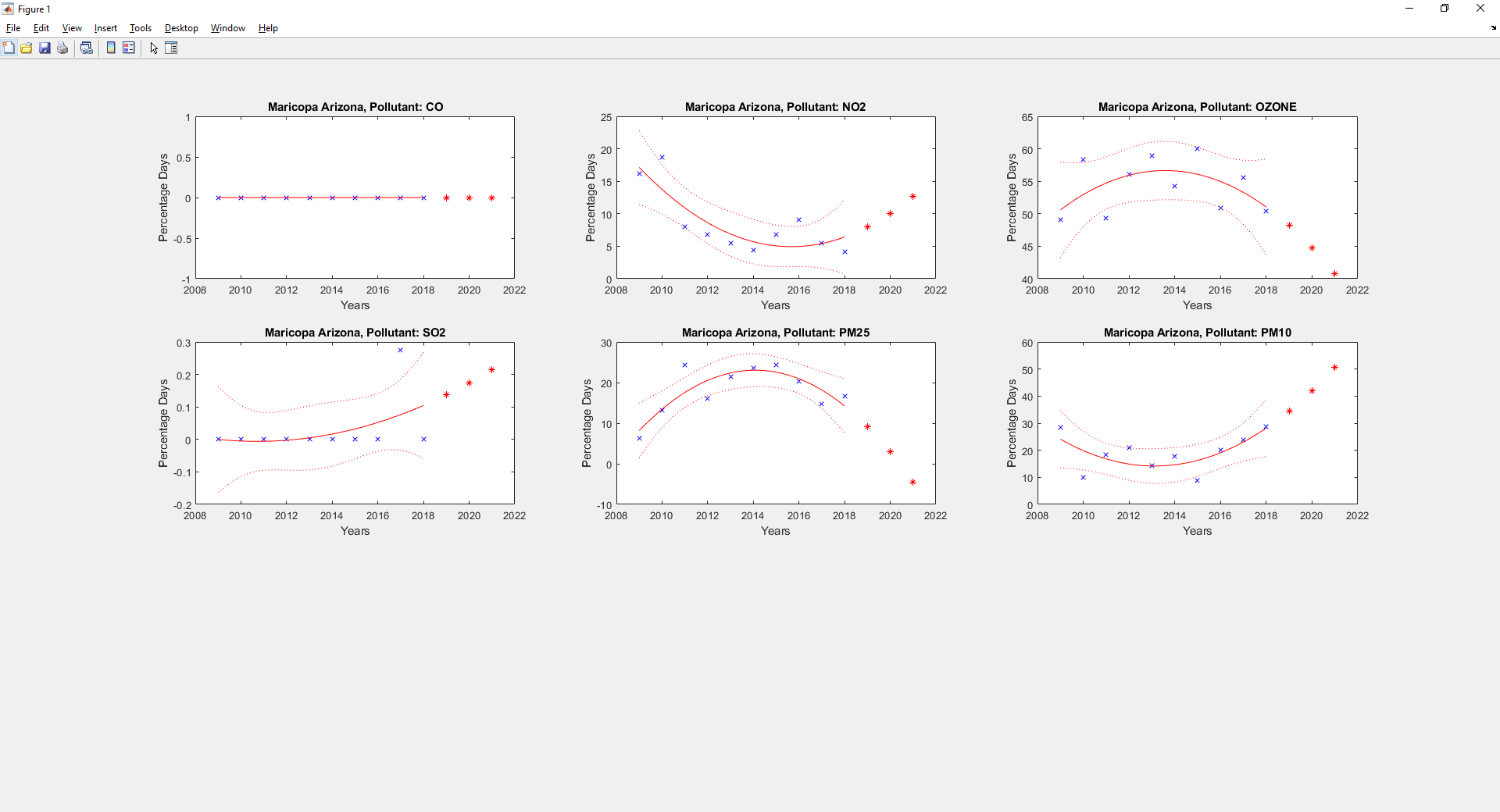
So after finding out the countries that are heavily polluted and their major pollutant. It would be interesting to try to predict the trend in air quality and dominant pollutant in the upcoming three years.

For this data analysis I am using a table that have the state and county name then it is followed by the annual air quality index statistic, like max, 90th percentile, medium. Then it is also followed by the number of days in which a certain pollutant is the dominant contributing factor to the increase of AQI. So it tells you which type of pollutant is really causing an increase in the AQI in a specific county. So first I took all ten years of data and put them in a cell array. The reason that I choose to put them into a cell array first is because the number of rows in each table is not consistent. Then by passing through the entire cell array through a function that I wrote, I was able to generate a county index. The county index included all the counties that have data in all ten years. Meaning if one of the tables, let’s say table 2013, did not have a county, let’s assume some county A, then county A will not be in the county index. Now, after cleaning up the tables and removing counties that are not recorded over 10 years. We can fetch the data from part 1 and create graphs for those counties on different variables.

I wrote various functions that can either plot multiple single plots or multiple subplots which will give a much more organized feeling when trying to visualize a lot of counties. Not only it can just plotting historical data I also included a polynomial regression model fitter by using fitlm. This helped me to generate the lines you can see on each graph. The lines are useful to showcase the general trend of the data over the ten years and I was also able to use the model to predict the value of three years in the future, and those are also plotted on the graph as red stars. Another feature is that you can choose the y axis to be either a specific pollutant or air quality index. Both will do different things. First of all, counties can plot year against air quality index to get a prediction of whether the air quality in that given county is going to increase or decrease over the upcoming 3 years. Then the counties can plot specific pollutant against years to see which pollutant is actually contributing to the worsening of air pollution and to see whether in the future they should be worried about that certain pollutants. This will be very helpful for policy maker to decide which industry and which pollutant to tackle in order to improve air quality.

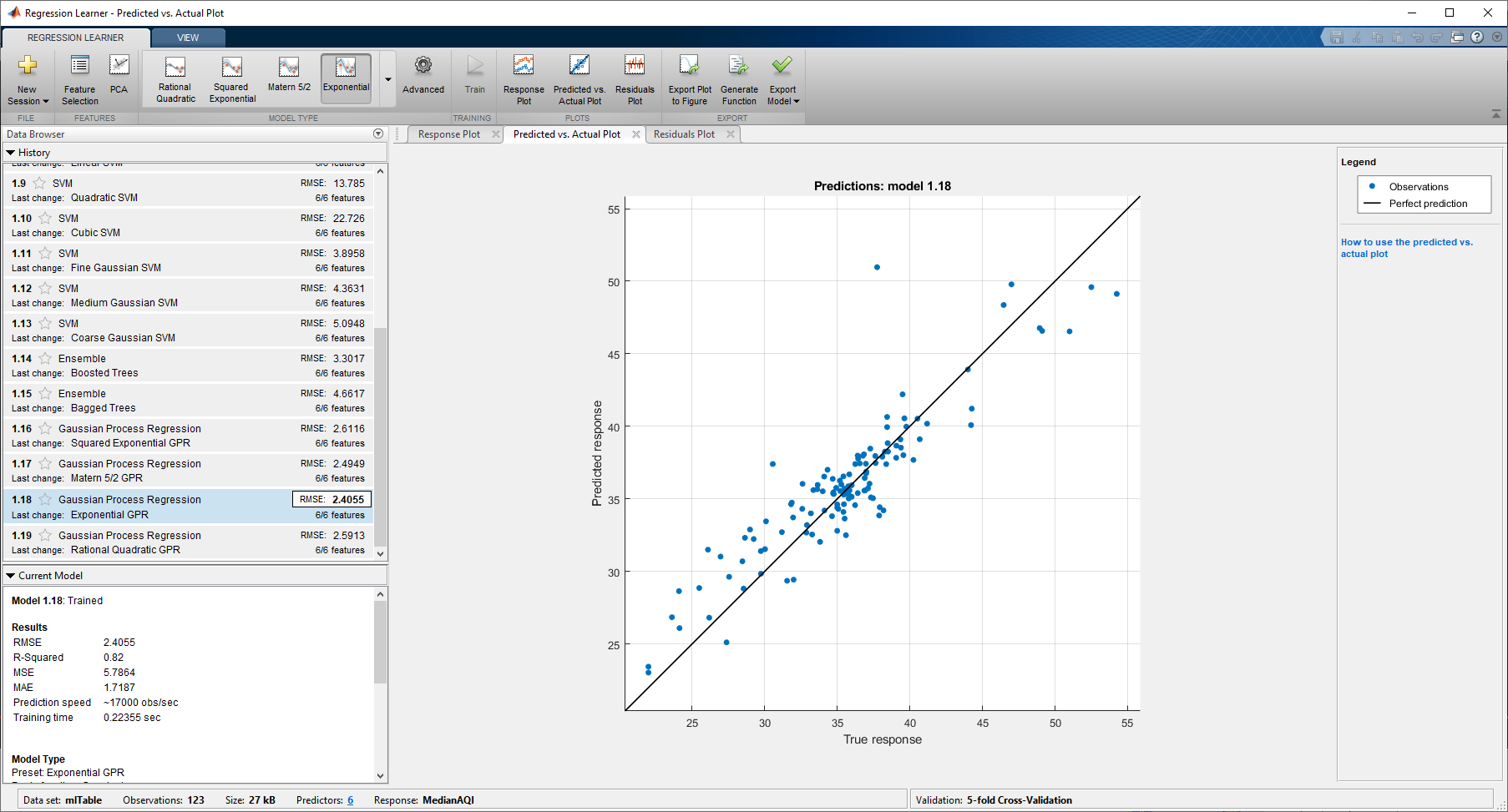
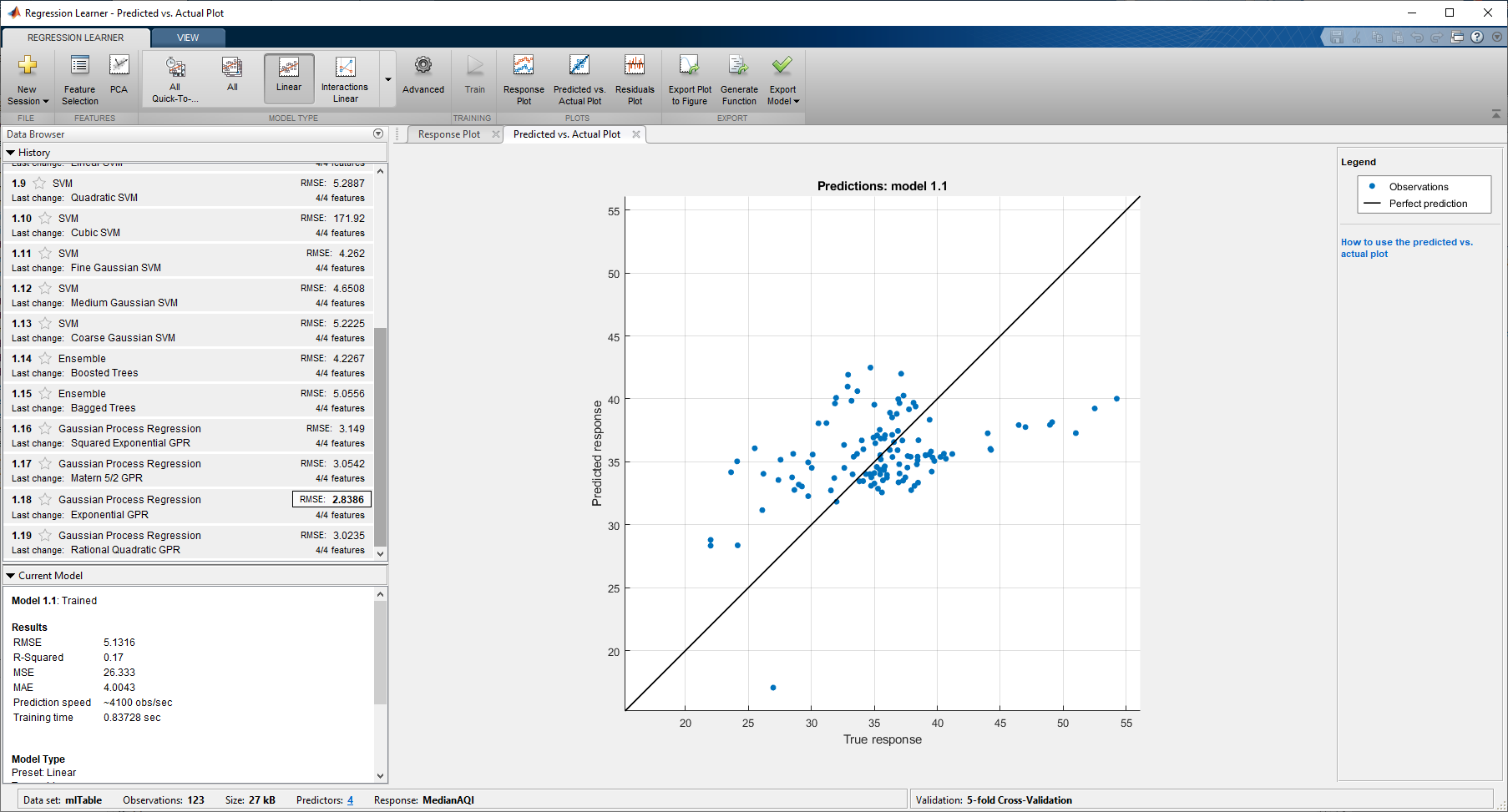
From part 1 we have a total of 62 counties that are heavily polluted. So above I have included a subplot of the first 9 states and I have included the rest of the counties in the appendix at the end of this paper. After running the program we first get a messaging saying Teller, Colorado does not have 10 year worth of data which means it will not be plotted, thus leaving us with 61 plots. So back to analysing the subplot above. We can see that there is a solid line that showcase the best fit trend of the data and there are also dotted lines that shows the error and uncertainty due to the data. Let’s analyze some of the better fitted ones, the two that jump out to me are Montgomery county and Calaveras county as they have a very narrow band of dotted line. In both of these cases the solid line predicted an upward trend meaning that AQI is going to be higher in up coming three years meaning the air quality is getting worse. Then we can look at the second plot which showcased that first PM2.5 is the dominant pollutant in both counties. You can tell from the title of the graph. Then you can also see that there is an upward trend of solid line meaning that PM2.5 is going to become an even more dominant force in worsening the air quality.

An interesting observation to make is Marloop county in Arizona, as shown in the first graph the AQI appears to increase representing a worsening air quality however, the current dominating pollutant, ozone, is on the decline. Thus it is worth taking a close look by subplotting all of the pollutant.



As shown on the graph ozone is on a downward trend, however PM 10 is on a great upward trend so although in 2018 ozone have a higher percentage days than PM 10 it is suggestive from the graph that in the near future PM 10 is going to over take ozone and become the dominant pollutant. So policy makers should actually try to lower PM 10 level.

C. The third part is to use the ML toolbox to predict the AQI with some external factors like wind, temperature, moisture and so on.

This is a model used to predict the air quality index using feature such latitude, longitude, wind speed, outdoor temperature, barometric pressure, and relative humidity. 

The model is being trained using the Gaussian Process Regression and with exponential GPR. The model have a root mean square error of 2.41 which is fairly decent.

The data scrubbing process is quite complicated for this model. One major difficulty we faced when creating the model is that there are no tables that have air quality index, latitude, longitude, wind speed, outdoor temperature, barometric pressure, and relative humidity in a single row. What we ended up finding is a table that have the annual air quality index data of a lot of the counties in different states. One table for each of the meteorological factors (wind speed, outdoor temperature, barometric pressure, relative humidity). These tables contain daily data of those measurements at a specific place. The problem is that the tables doesn’t match with each other. For example, the table that collects wind speed data will have station at county A (I am just using A for simplicity sake) in Alabama. But the table that collects outdoor temperature will have station at county B in Alabama. And the table that collects barometric pressure will have station at another different county C in Alabama. Thus we can’t just concatenate the tables together nor can we create a table based on county since no county will have all of the variables.

So the solution that we have come up with is to treat each state as a whole and average all the value that's collected from each state. That includes air quality index, wind speed, outdoor temperature, barometric pressure, and relative humidity. So what I would do is average all the air quality index that were measured at different counties in Alabama to find the average air quality index in Alabama. Then I would also average all the daily climate factors that were collected in Alabama, doesn’t matter which county it is in. So in the end we would have one table that is the annual average of all the factors in different states. So the rows will represent each state and columns will represent, state name, air quality index, wind speed, outdoor temperature, barometric pressure, and relative humidity.

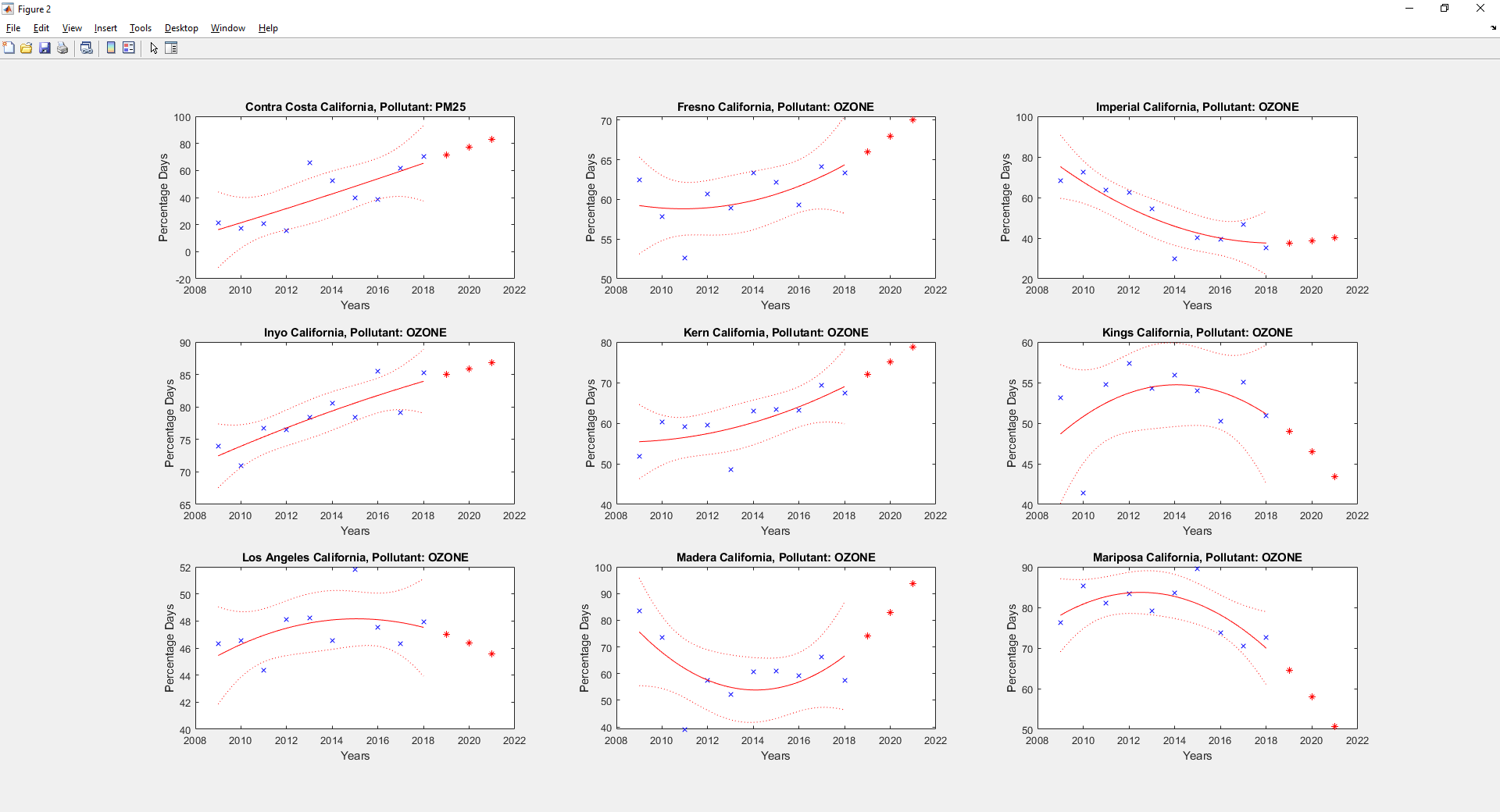
However, during the process, another problem was discovered, The problem was meteorological table did not have value from all of the states. The one that is missing the most is the barometric pressure table, which only recorded data from 41 states meaning there were no barometric pressure data for a lot of the states. So we had to remove some states from other tables to match the 41 states that are in the barometric pressure table.

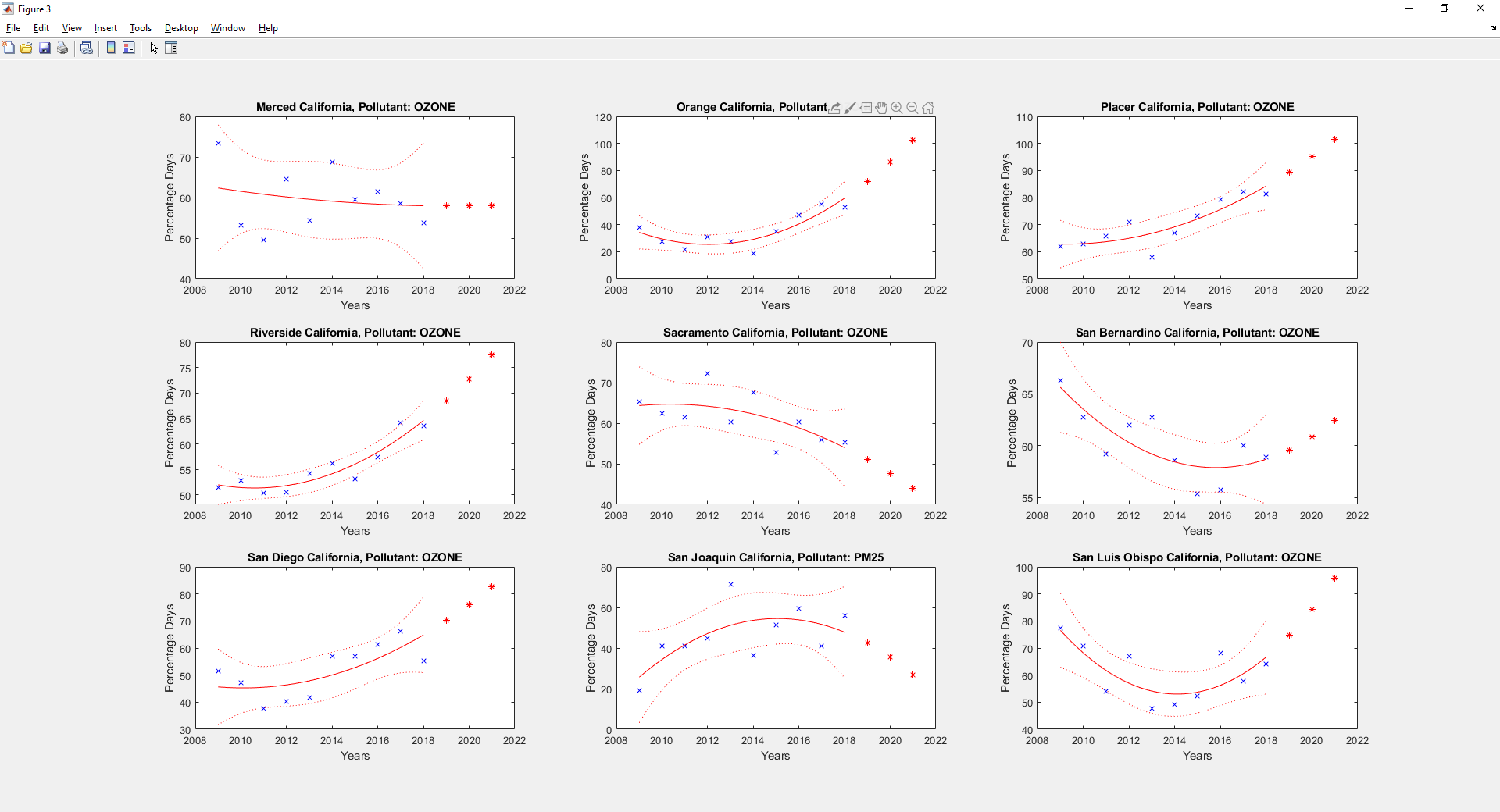
Then after cleaning through three years worth of AQI (air quality index) and meteorological data we ended up with a table of 123x6. Because each year there are data for 41 states so three years 41\*3 is 123. Then I decided to test it out in the machine learning tool box with regression learner, since we are trying to predict an AQI number. However, I found out that the correlation is not the best. The reason because different states are at different location so it would be helpful to include some sort of location data as a feature to the model. Then I found a list of all the latitude and longitude for each of the state. After that, the model was able to predict the AQI much more accurately. As shown on the graphs above the first graph is without state coordinate and the second one is with state coordinate. The second one visually fits much better and have have smaller root mean square error

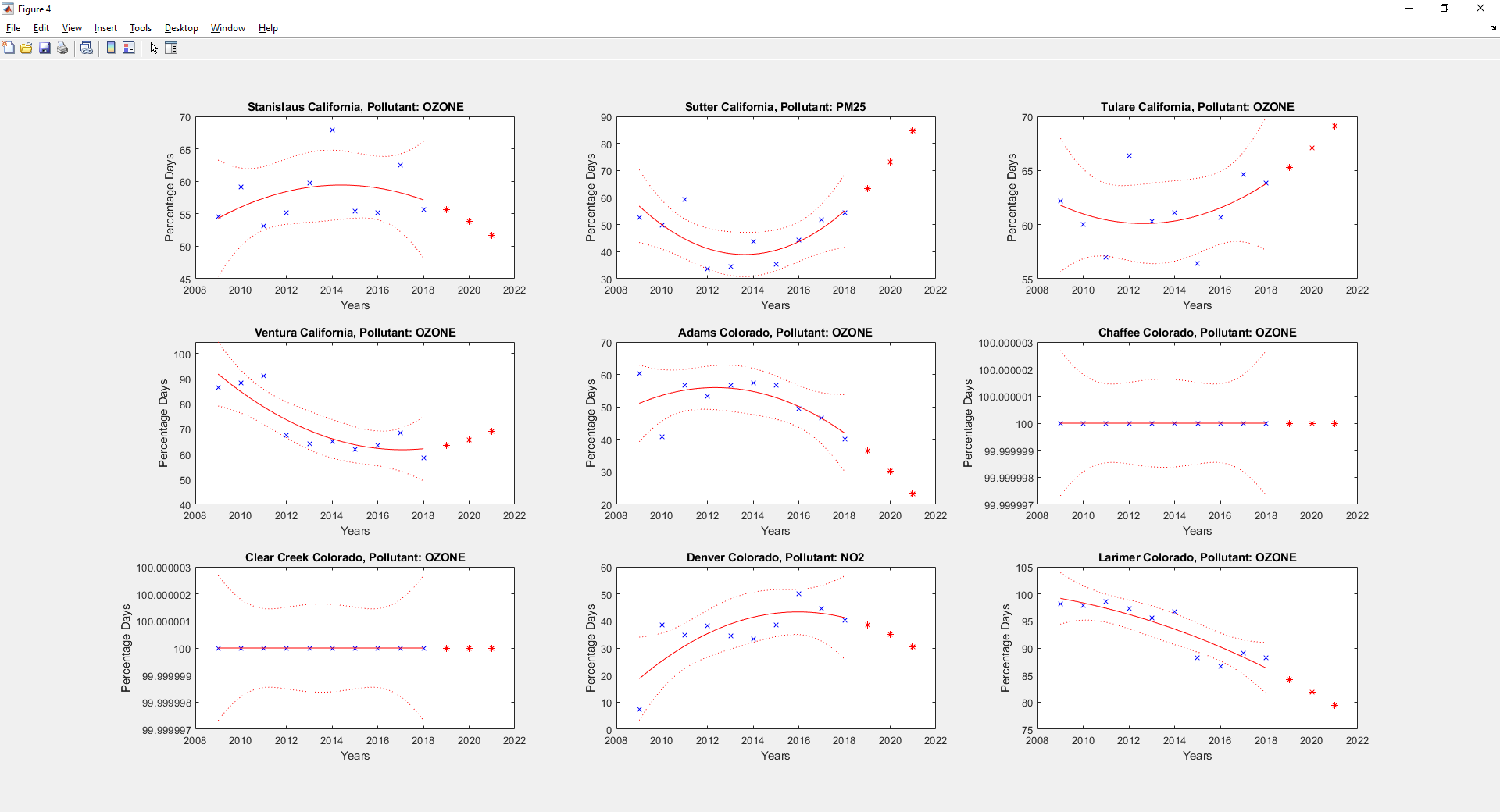
The reason that we made this model is that it is an easier way for people to estimate the AQI in their area. Getting the temperature, humidity, pressure, and wind speed is fairly easy and an instrument that will take these measurements will be fairly cheap as well. Compare instrument that will detect the amount of pollutant gas like CO, NO2, Ozone, and etc. Thus by creating this machine learning model it creates a cheaper and easier alternative for people to gain access to measuring the AQI.

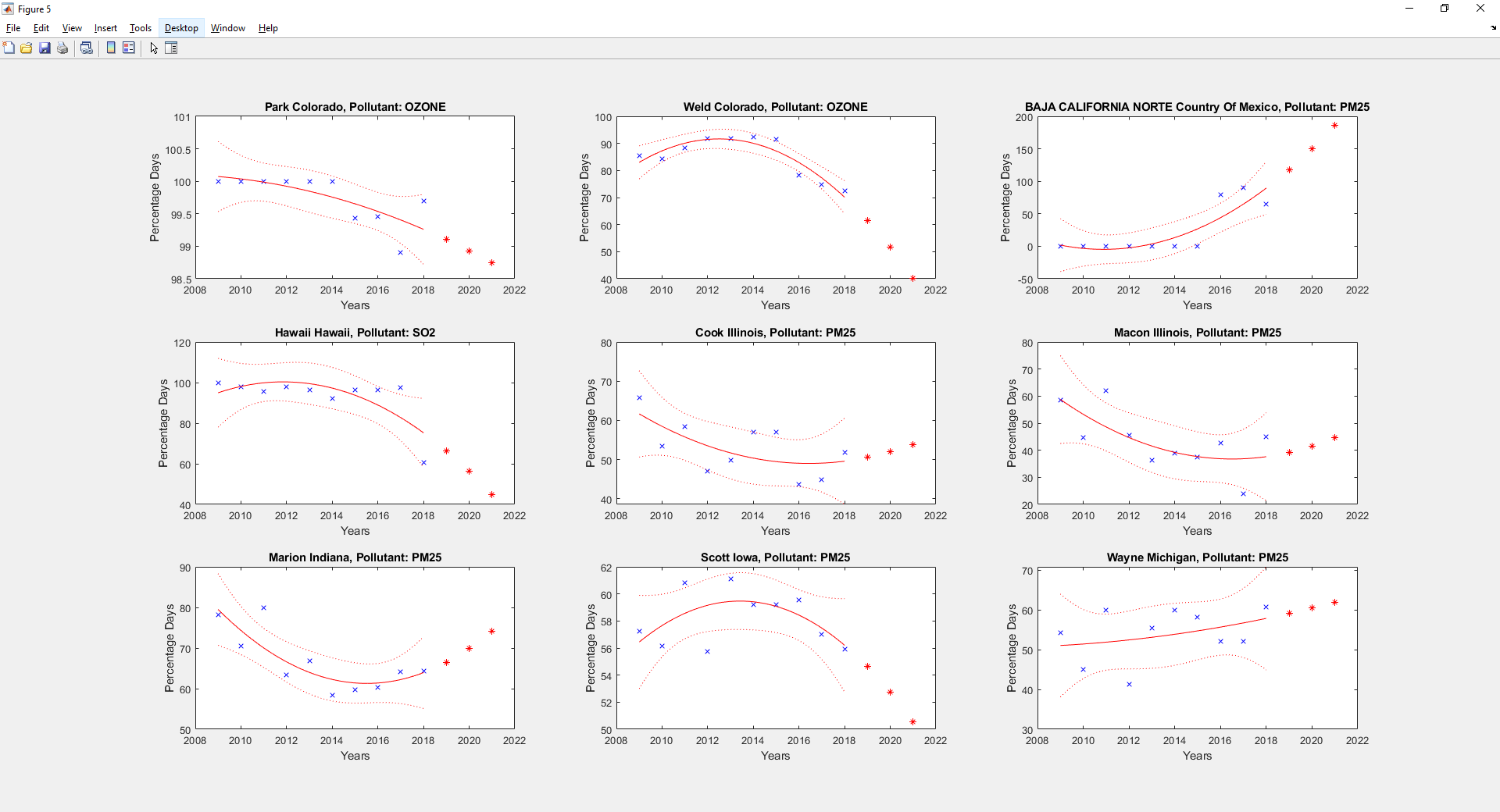
Appendix:

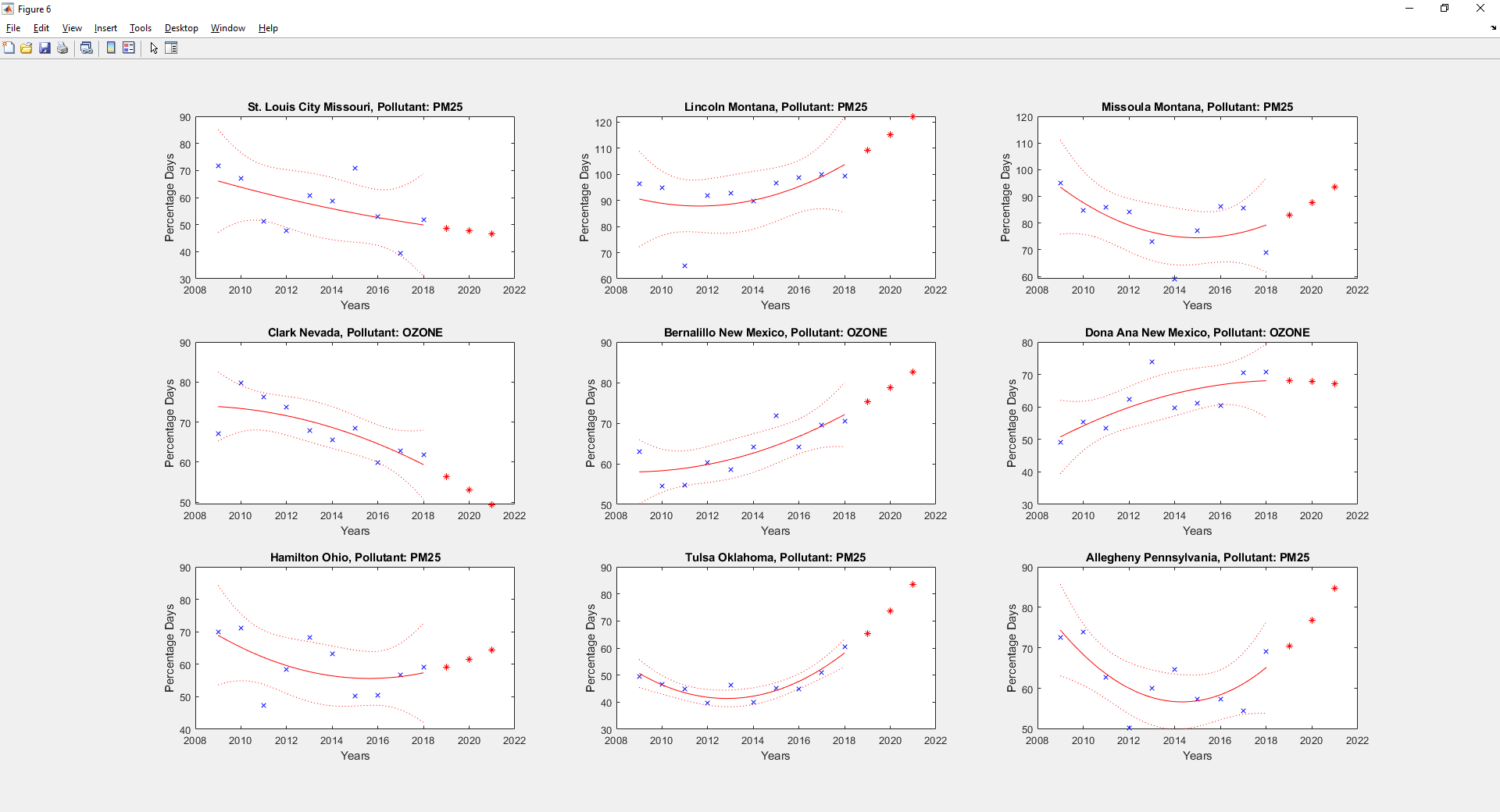
Graphs for 10 years against most dominant pollutant

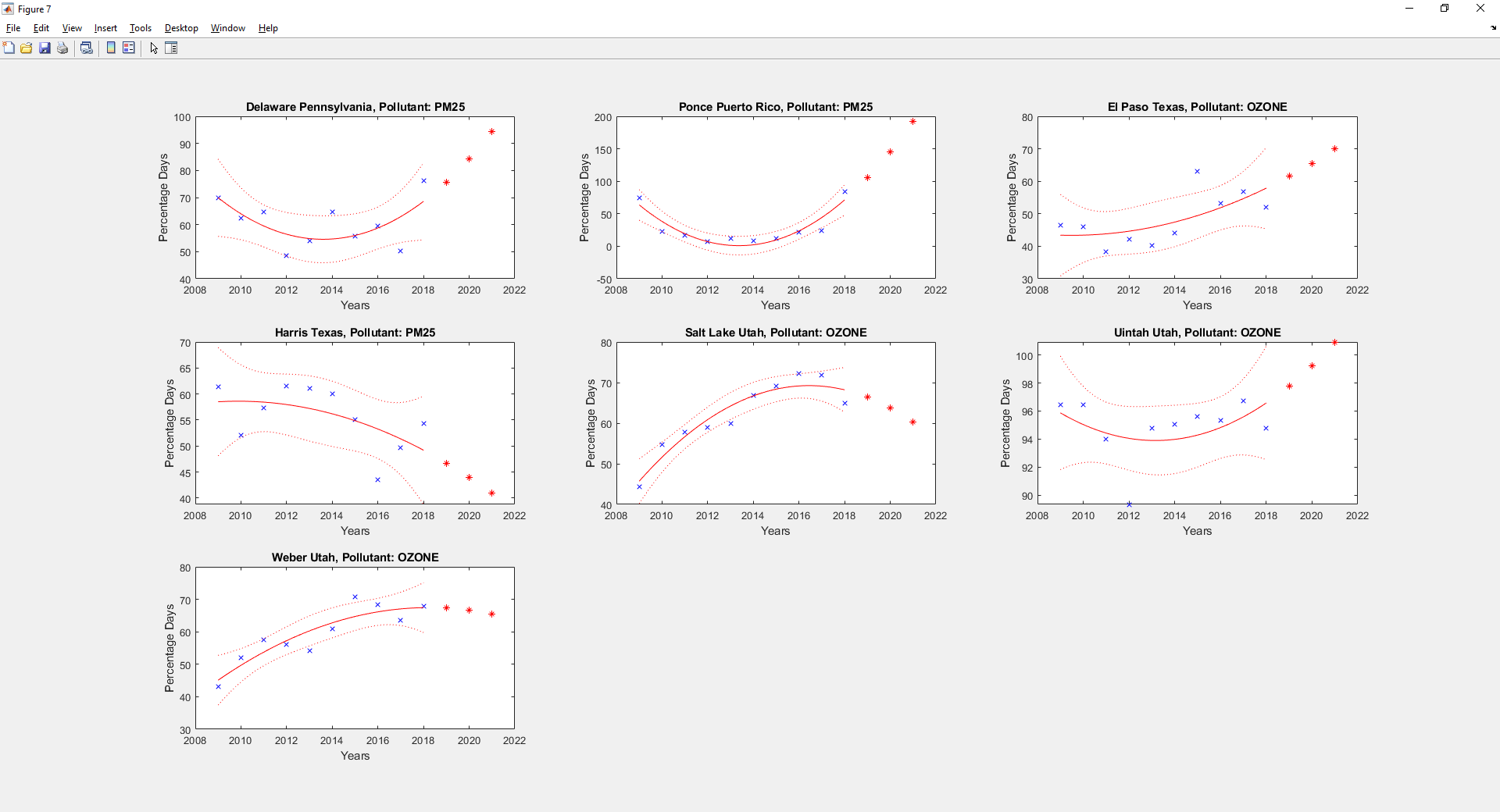




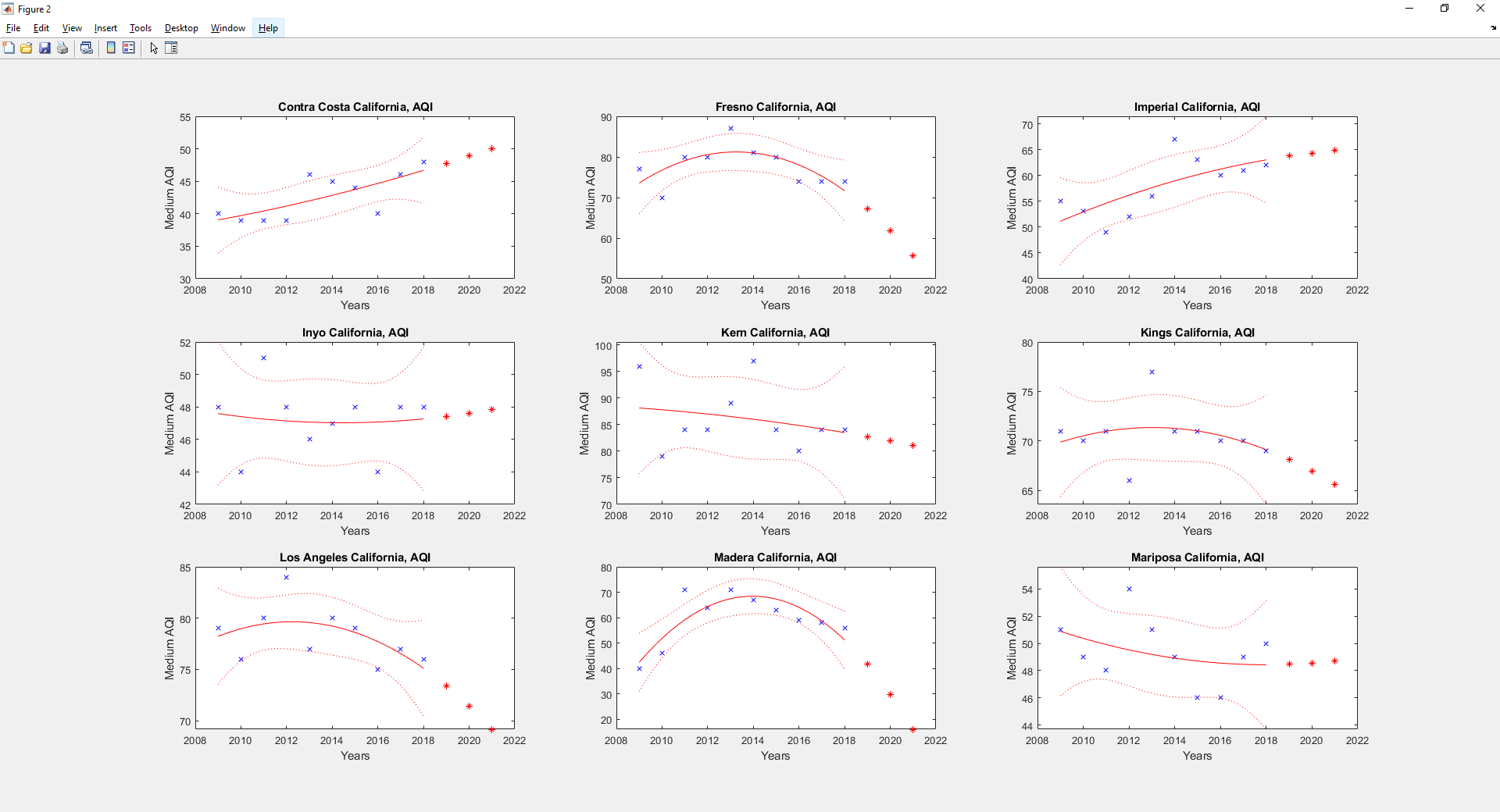


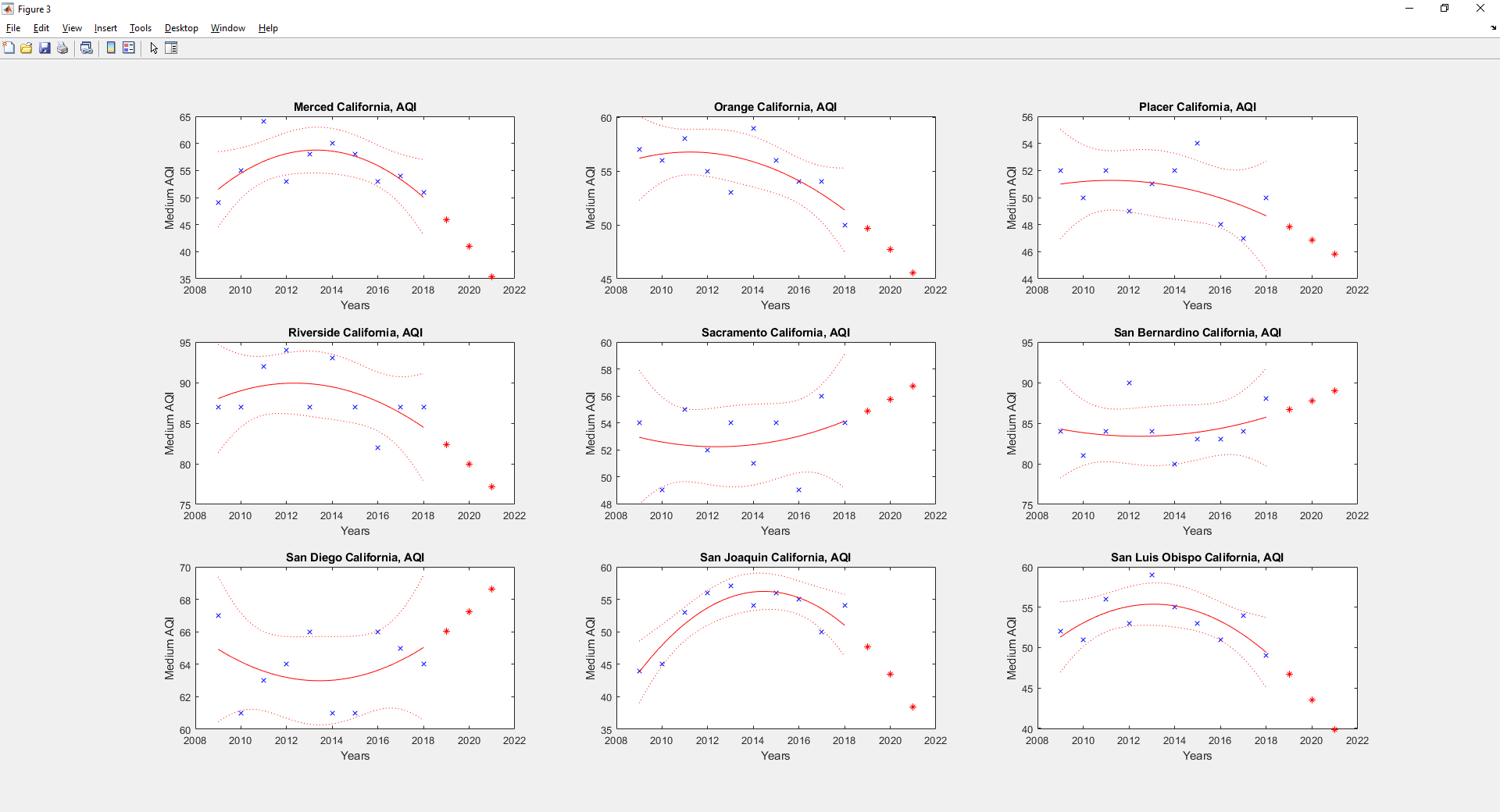


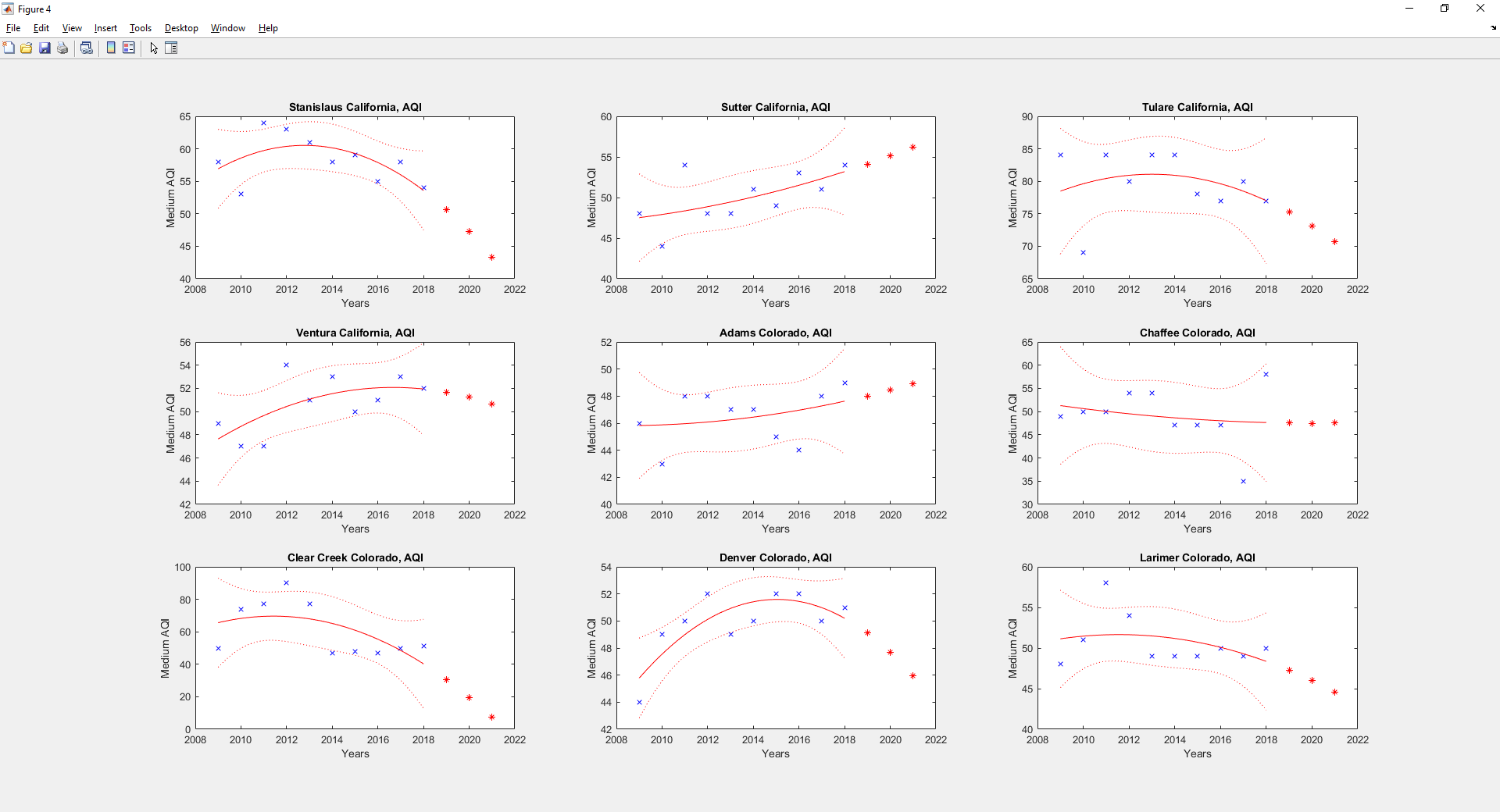


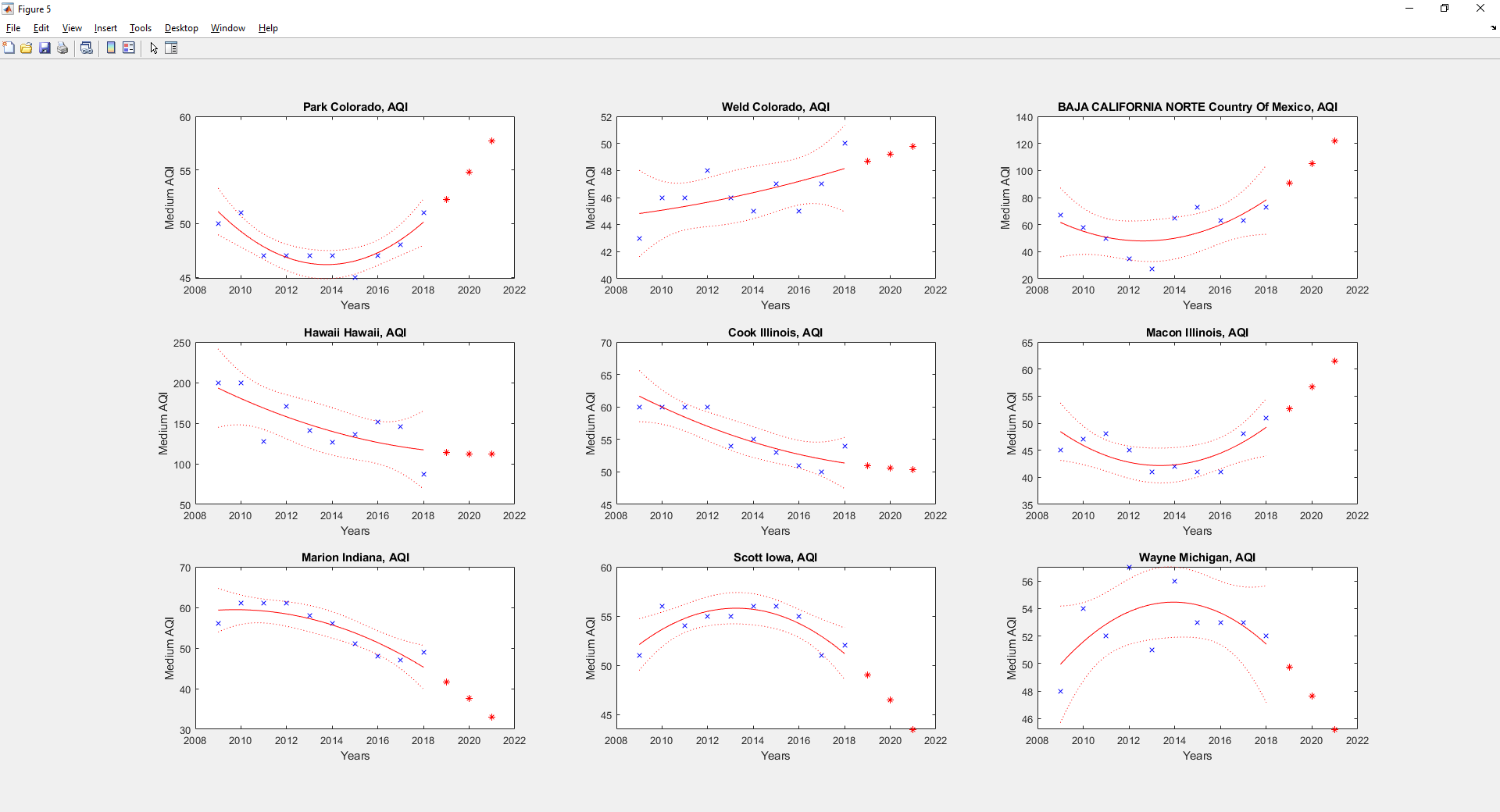


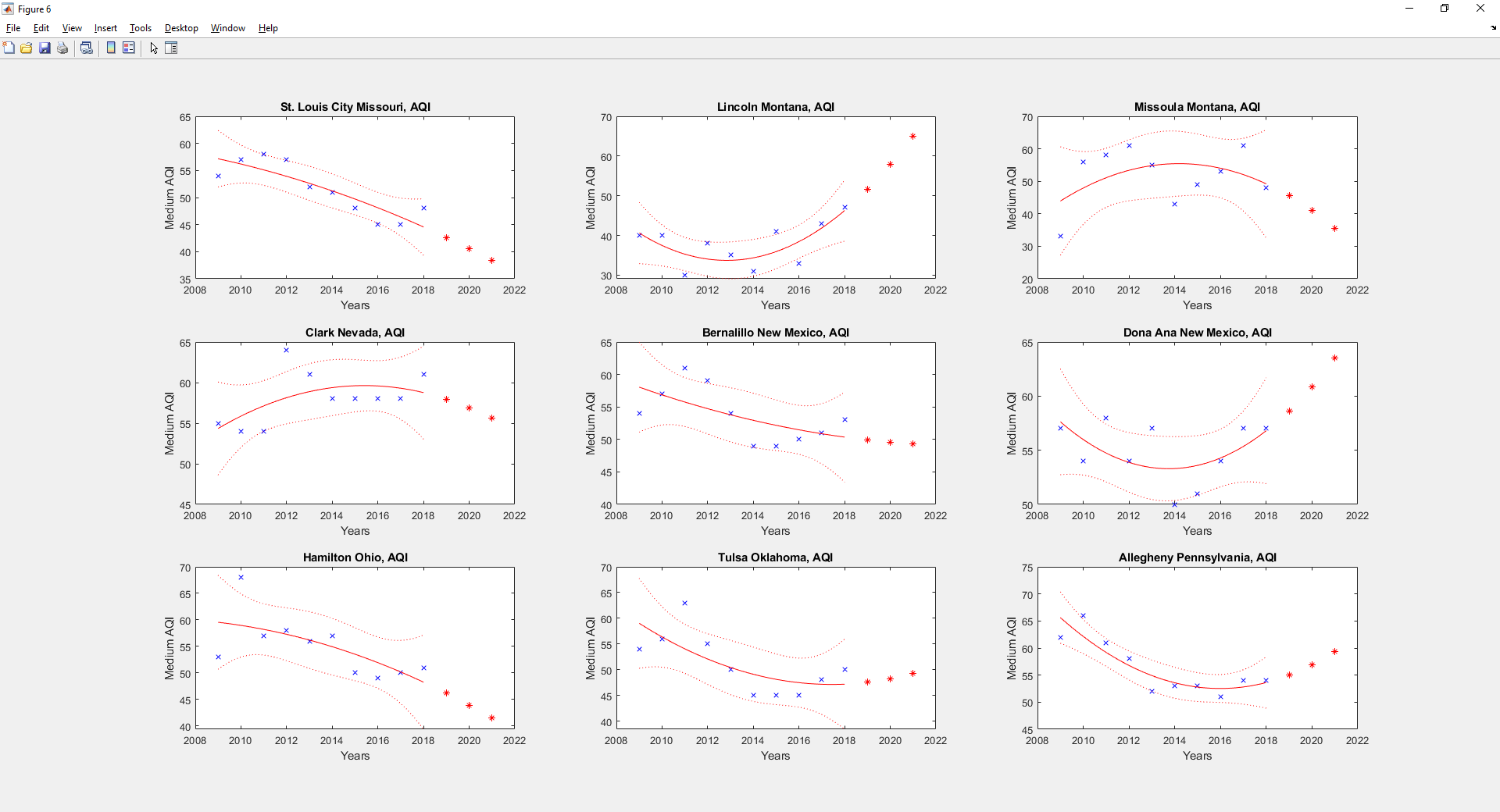
Graphs for 10 years against AQI

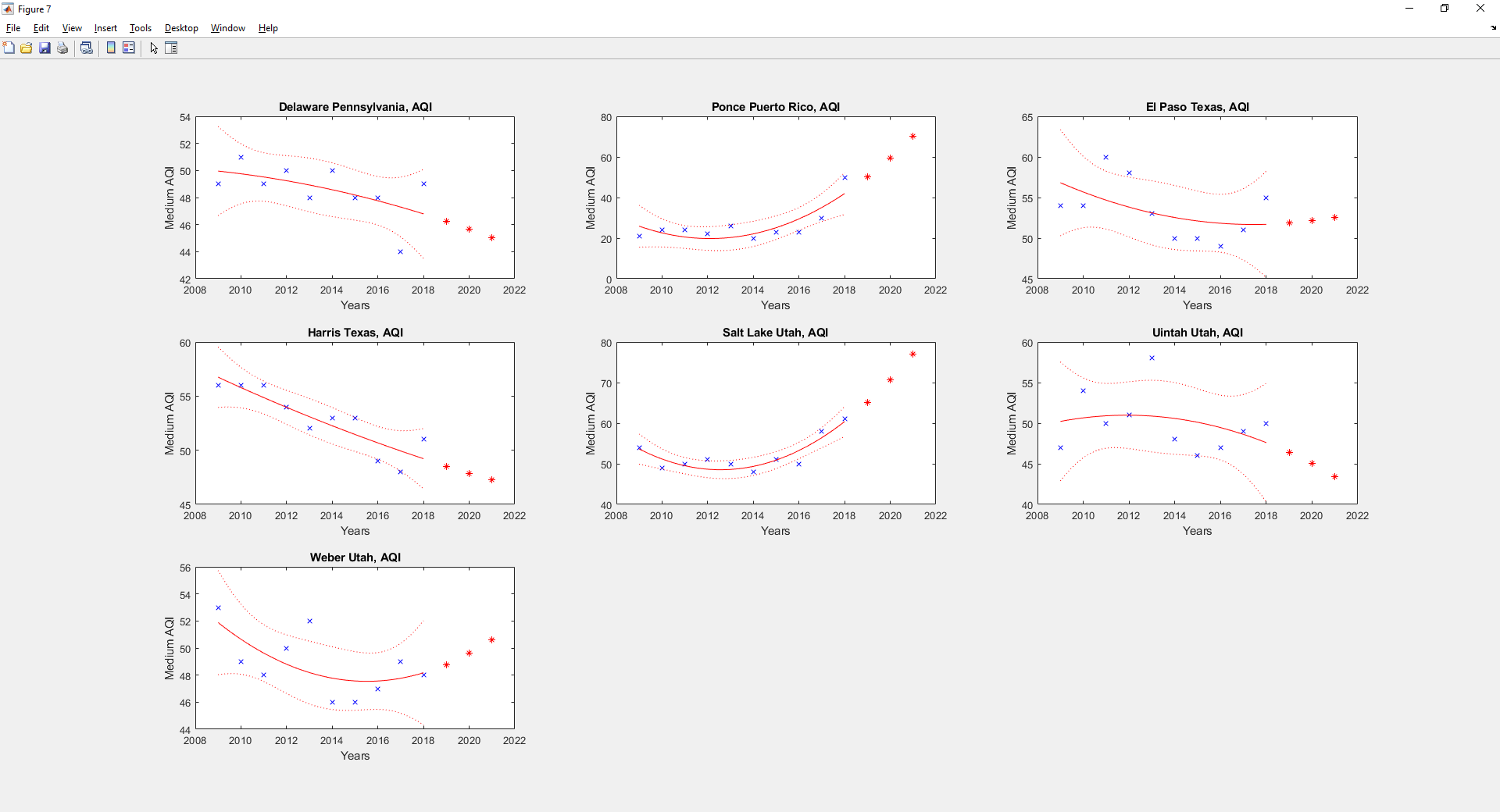












Code:

Part 2

%fromt 2009 - 2018

tablesInCell = getTablesInCell10year();

index = getIndex(tablesInCell);

tablesInCellScrubbed = scrubCell(tablesInCell, index);

%totalSearch = mostp;

%totalSearch = {{"Arizona"; "Maricopa"; "CO"}, {"Arizona"; "Maricopa"; "NO2"}, {"Arizona"; "Maricopa"; "OZONE"}...

{"Arizona"; "Maricopa"; "SO2"}, {"Arizona"; "Maricopa"; "PM25"}, {"Arizona"; "Maricopa"; "PM10"}};

multiPlotOverYear(tablesInCellScrubbed, index, totalSearch, 3);

clear tablesInCell;

function out = getTablesInCell10year()

t2018 = readtable('aqi2018.csv');

t2017 = readtable('aqi2017.csv');

t2016 = readtable('aqi2016.csv');

t2015 = readtable('aqi2015.csv');

t2014 = readtable('aqi2014.csv');

t2013 = readtable('aqi2013.csv');

t2012 = readtable('aqi2012.csv');

t2011 = readtable('aqi2011.csv');

t2010 = readtable('aqi2010.csv');

t2009 = readtable('aqi2009.csv');

out = {t2009, t2010, t2011, t2012, t2013, t2014, t2015, t2016, t2017, t2018};

end

function out = getIndex(tablesInCell)

columns{1, length(tablesInCell)} = [];

for i = 1:length(tablesInCell)

columns{1, i} = table2cell(tablesInCell{1, i}(:,[1 2]));

end

out = columns{1, 1};

for i = 2:length(columns)

out = compareCell(out, columns{1, i});

end

end

function out = compareCell(cell1, cell2)

i = 1;

while i <= length(cell1)

[state, county] = cell1{i, :};

if isInCell(state, county, cell2) == 0

cell1(i, :) = [];

else

i = i+1;

end

end

out = cell1;

end

function out = isInCell(state, county, inCell)

%output 0 or 1, 0 means the state, county combination not in inCell

out = 0;

range = 1:length(inCell);

for i = range

if and(strcmp(inCell{i, 1}, state), strcmp(inCell{i, 2}, county))

out = 1;

break;

end

end

end

function out = scrubCell(tablesInCell, index)

out{1, length(tablesInCell)} = [];

for i = 1:length(tablesInCell)

out{1, i} = scrubTable(tablesInCell{1, i}, index);

end

end

function out = scrubTable(inTable, index)

i = 1;

while i <= height(inTable)

state = inTable{i, 1};

county = inTable{i, 2};

if isInCell(state, county, index) == 0

inTable(i, :) = [];

else

i = i + 1;

end

end

out = inTable;

end

function multiPlotOverYear(tablesInCellScrubbed, index, totalSearch, varargin)

figNum = 1;

subFigNum = 1;

fprintf("\n")

for i = 1:length(totalSearch)

oneSearch = totalSearch{1, i};

state = oneSearch{1, 1};

county = oneSearch{2, 1};

if isInCell(state, county, index) == 0

fprintf("%s, %s does not have 10 years of data to plot a graph!!!\n", county, state);

else

figure(figNum);

if nargin > 3

if nargin == 4

plotSubPlot(tablesInCellScrubbed, oneSearch, subFigNum, varargin{1});

else

plotSubPlot(tablesInCellScrubbed, oneSearch, subFigNum, varargin{1}, varargin{2});

end

subFigNum = subFigNum + 1;

if subFigNum > (varargin{1} \* varargin{1})

subFigNum = 1;

figNum = figNum + 1;

end

else

plotOnePlot(tablesInCellScrubbed, oneSearch);

figNum = figNum + 1;

end

end

end

end

function plotSubPlot(tablesInCellScrubbed, oneSearch, subFigNum, figSize, varargin)

subplot(figSize, figSize, subFigNum);

if nargin > 4

plotOnePlot(tablesInCellScrubbed, oneSearch, varargin{1});

else

plotOnePlot(tablesInCellScrubbed, oneSearch);

end

end

function plotOnePlot(tablesInCellScrubbed, oneSearch, varargin)

historic(10) = 0;

for i = 1:10

oneYearTable = tablesInCellScrubbed{1, i};

if nargin > 2

historic(i) = getOneYearValue(oneYearTable, oneSearch, varargin{1});

else

historic(i) = getOneYearValue(oneYearTable, oneSearch);

end

end

x = 2009:2018;

y = historic;

mdl = fitlm(x,y,'purequadratic');

xPre = 2019:2021;

yPre = predict(mdl,xPre');

plot(mdl);

hold on;

plot(xPre, yPre, 'r\*');

legend('off');

hold off;

%plot(x, y, 'bo');

state = oneSearch{1, 1};

county = oneSearch{2, 1};

pollutant = oneSearch{3, 1};

xlabel('Years');

if nargin > 2

title([char(county), ' ', char(state), ', AQI'], 'Interpreter', 'none');

ylabel('Medium AQI');

else

title([char(county), ' ', char(state), ', Pollutant: ', char(pollutant)], 'Interpreter', 'none');

ylabel('Percentage Days');

end

end

function out = getOneYearValue(oneYearTable, oneSearch, varargin)

state = oneSearch{1, 1};

county = oneSearch{2, 1};

pollutant = oneSearch{3, 1};

if nargin > 2

out = getMediumAQI(oneYearTable, state, county);

else

rawValue = getRawValueFromTable(oneYearTable, state, county, pollutant);

out = (rawValue(1) / sum(rawValue(2:end))) \* 100;

end

end

function out = getMediumAQI(oneYearTable, state, county)

tableIndex = getTableIndex(oneYearTable, state, county);

out = oneYearTable{tableIndex, 13};

end

function out = getRawValueFromTable(oneYearTable, state, county, pollutant)

tableIndex = getTableIndex(oneYearTable, state, county);

pollutantReference = {"CO","NO2","OZONE","SO2","PM25","PM10"};

pollutantIndex = 13 + find([pollutantReference{:}] == pollutant);

out(7) = 0;

out = [oneYearTable{tableIndex, pollutantIndex}];

for i = 1:length(pollutantReference)

out(i+1) = oneYearTable{tableIndex, i+13};

end

end

function out = getTableIndex(oneYearTable, state, county)

out = 0;

for i = 1:height(oneYearTable)

if and(strcmp(oneYearTable{i, 1}, state), strcmp(oneYearTable{i, 2}, county))

out = i;

break;

end

end

end

Part 3:

windData = initWind();

tempData = initTemp();

pressData = initPress();

rhData = initRH();

aqiData = initAQI();

searchCol = "ParameterName";

searchTermIndex = {"Wind Speed - Resultant", "Outdoor Temperature", "Barometric pressure", "Relative Humidity "};

colVarIndex = {'WindSpeed', 'Temperature', 'Pressure', 'Humidity'};

colName = {'StateName', 'ArithmeticMean'};

aqiColName = {'State', 'MaxAQI', 'x90thPercentileAQI', 'MedianAQI'};

windTable = processData(windData, colName, searchCol, searchTermIndex{1}, colVarIndex{1});

clear windData;

tempTable = processData(tempData, colName, searchCol, searchTermIndex{2}, colVarIndex{2});

clear tempData;

pressTable = processData(pressData, colName, searchCol, searchTermIndex{3}, colVarIndex{3});

clear pressData;

rhTable = processData(rhData, colName, searchCol, searchTermIndex{4}, colVarIndex{4});

clear rhData;

aqiTable = processData(aqiData, aqiColName);

clear aqiData;

clear searchCol searchTermIndex colVarIndex colName aqiColName

function out = processData(dataIn, colName, varargin)

%varargin = searchCol, searchTerm

out{length(dataIn)} = 0;

for i = 1:length(dataIn)

oneYearTable = dataIn{i};

if nargin > 2

tempTable = createVarTable(oneYearTable, colName, varargin{1}, varargin{2});

stateAverage = avergeByState(tempTable, colName, cellstr(varargin{3}));

else

tempTable = createColTable(oneYearTable, colName);

stateAverage = avergeByState(tempTable, colName, colName(2:end));

end

out{i} = stateAverage;

end

end

function out = avergeByState(tableIn, colName, colVar)

statesTable = tableIn(1:end, 1);

states = catForColumn(statesTable);

averages(length(states), length(colName) - 1) = 0;

for i = 1:length(states)

searchCol = string(colName{1});

searchTerm = states{i};

oneStateTable = createVarTable(tableIn, colName, searchCol, searchTerm);

averages(i, :) = mean(oneStateTable{:,2:end}, 1);

end

averages = array2table(averages, 'VariableNames', colVar);

states = states';

out = table(states);

out = [out averages];

end

function out = createVarTable(tableIn, colName, searchCol, searchTerm)

col(length(colName)) = 0;

for i = 1:length(colName)

col(i) = find(string(tableIn.Properties.VariableNames) == string(colName{i}));

end

searchCol = find(string(tableIn.Properties.VariableNames) == searchCol);

out = tableIn(strcmp(tableIn.(searchCol), char(searchTerm)), col);

end

function out = createColTable(tableIn, colName)

col(length(colName)) = 0;

for i = 1:length(colName)

col(i) = find(string(tableIn.Properties.VariableNames) == string(colName{i}));

end

out = tableIn(1:end, col);

end

function out = catForColumn(tableIn)

diction = {};

for i = 1:height(tableIn)

temp = string(table2array(tableIn(i, 1)));

inDic = 0;

if isempty([diction{:}]) == 0

if isempty(find([diction{:}] == temp{1})) == 0

inDic = 1;

end

end

if inDic == 0

diction{length(diction)+1} = temp;

end

end

out = diction;

end

function out = initWind()

out = {readtable('wind2016.csv'), readtable('wind2017.csv'), readtable('wind2018.csv')};

end

function out = initTemp()

out = {readtable('temp2016.csv'), readtable('temp2017.csv'), readtable('temp2018.csv')};

end

function out = initPress()

out = {readtable('press2016.csv'), readtable('press2017.csv'), readtable('press2018.csv')};

end

function out = initRH()

out = {readtable('rhdp2016.csv'), readtable('rhdp2017.csv'), readtable('rhdp2018.csv')};

end

function out = initAQI()

out = {readtable('aqi2016.csv'), readtable('aqi2017.csv'), readtable('aqi2018.csv')};

end