SOCI 20253 Project Report

Huanye Liu

1. Introduction

California, the most populated state in the US, has unfortunately seen several runs of Infectious disease these years. Many kinds of infectious disease were found to co-occur frequently and were reported to be correlated with not only the natural environment but also with some social-economic factors. In this project, we explore and analyze the spatial pattern of infectious disease outbreaks in California at the county level with the main purpose to identify both the environmental and social-economic factors which may be potentially related to the infectious disease rates there.

2. Data sources

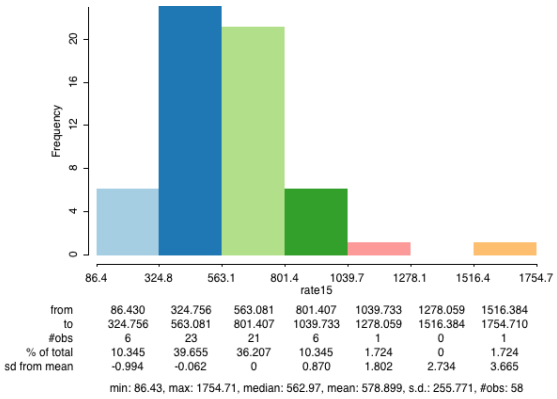
Two datasets are used in this project. The first dataset records the per capita rates of infectious disease at the county level in California in 2010 and 2015, which can be downloaded from California Health and Human Services Open Data Portal (https://data.chhs.ca.gov/dataset/infectious-disease-cases-by-county-year-and-sex). The other dataset is on natural environment and population characteristics in California in 2010 and 2015, which includes several pollutant burden variables and population characteristic variables by the county level. The pollutant burden variables used here are ozone, PM2.5, diesel PM, drinking water, pesticides, toxic release, traffic

density, cleanup sites, groundwater threats, hazardous waste, impaired water bodies, solid waste sites, and the population characteristics variables used in the project include asthma rate, education level, linguistic isolation index, unemployment rate and poverty rate. This dataset can be downloaded from California Open Data Portal (https://data.ca.gov/dataset/calenviroscreen-30, https://data.ca.gov/dataset/calenviroscreen-20).

3. Exploratory data analysis

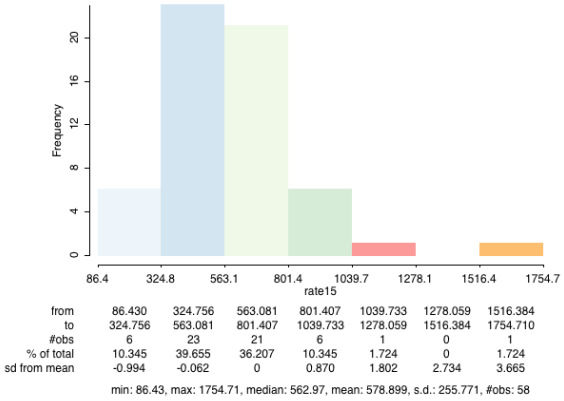
3.1 Initial data exploration

First, using the 2015 data, we carry out some data explorations to detect any potential features among variables. We first use the histogram to find out the range of infectious disease rate as below:



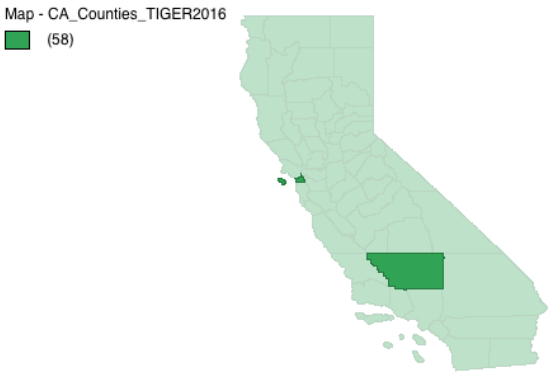
From the histogram above, we can see two counties have extremely high infectious rates (2 standard deviations above the mean). Using the linking the brushing function,

we identify the two counties are San Francisco County and Kern County, and their locations are shown as below:

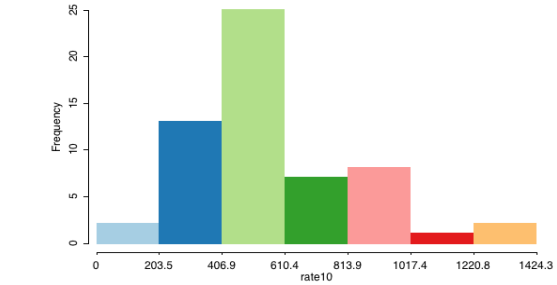


from	86.430	324.756	563.081	801.407	1039.733	1278.059	1516.384
to	324.756	563.081	801.407	1039.733	1278.059	1516.384	1754.710
#obs	6	23	21	6	1	0	1
% of total	10.345	39.655	36.207	10.345	1.724	0	1.724
sd from mean	-0.994	-0.062	0	0.870	1.802	2.734	3.665

min: 86.43, max: 1754.71, median: 562.97, mean: 578.899, s.d.: 255.771, #obs: 58



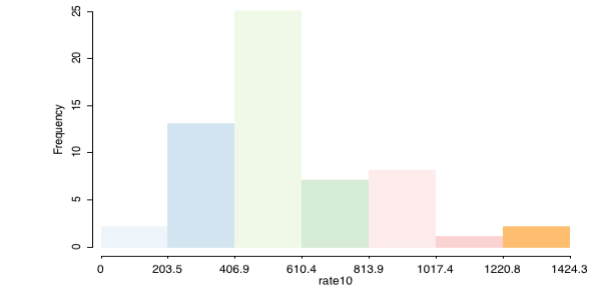
Next, we use the 2010 data to plot the histogram of infectious disease rate:



from	0	203.471	406.943	610.414	813.886	1017.357	1220.829
to	203.471	406.943	610.414	813.886	1017.357	1220.829	1424.300
#obs	2	13	25	7	8	1	2
% of total	3.448	22.414	43.103	12.069	13.793	1.724	3.448
sd from mean	-1.285	-0.554	0	0.177	0.908	1.638	2.369

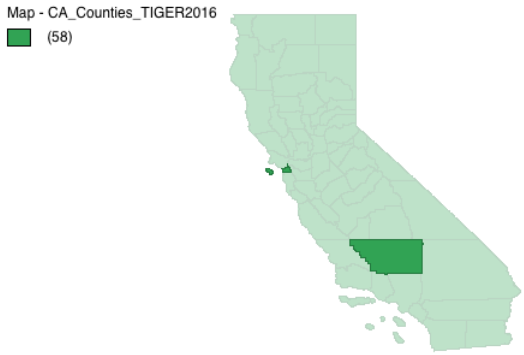
min: 0, max: 1424.3, median: 522.01, mean: 561.224, s.d.: 278.39, #obs: 58

and we find that the overall level of disease rate is less than that in 2015 and that there are still two counties with extremely high values (2 standard deviations above the mean). So using the linking and brushing function again, we identify the two counties are the same as in year 2015:

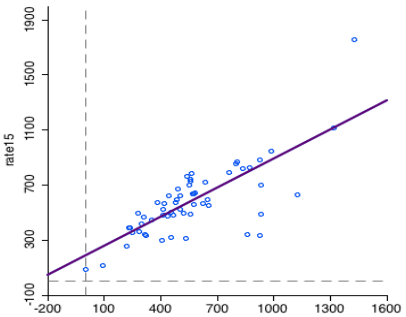


from	0	203.471	406.943	610.414	813.886	1017.357	1220.829
to	203.471	406.943	610.414	813.886	1017.357	1220.829	1424.300
#obs	2	13	25	7	8	1	2
% of total	3.448	22.414	43.103	12.069	13.793	1.724	3.448
sd from mean	-1.285	-0.554	0	0.177	0.908	1.638	2.369

min: 0, max: 1424.3, median: 522.01, mean: 561.224, s.d.: 278.39, #obs: 58



The coincidence of extreme values prompts us to examine the general pattern similarity between infectious disease rates in 2010 and 2015. We first use a scatter plot for rate comparison between the two years:



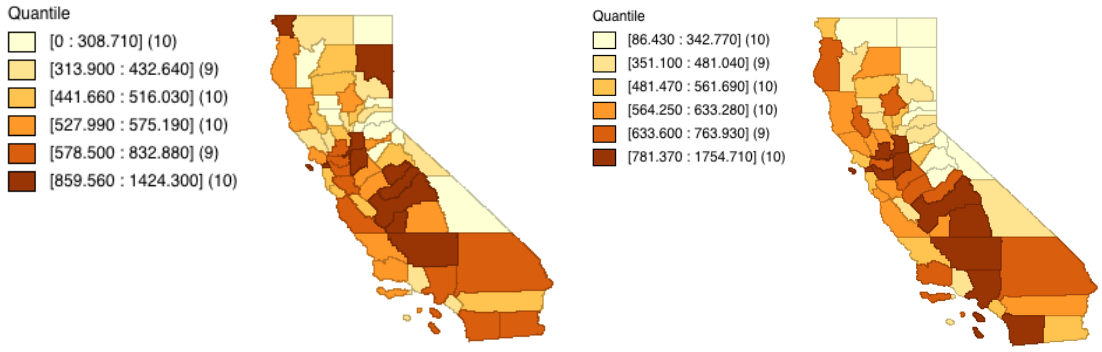
#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
58	0.591	182.443	49.094	3.716	0.000	0.706	0.078	8.999	0.000
0	0	0	0	0	0	0	0	0	0
58	0.591	182.443	49.094	3.716	0.000	0.706	0.078	8.999	0.000

Chow test for sel/unsel regression subsets: can't compute

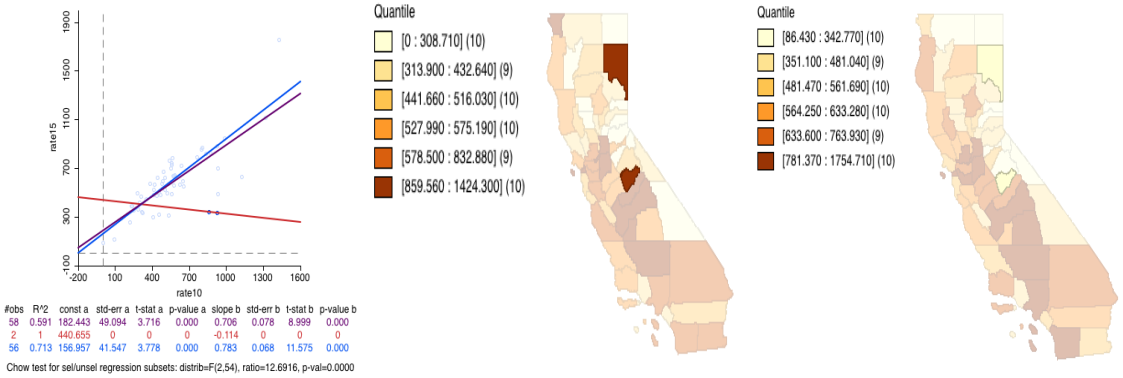
and we can see generally the two rates are positively correlated, and under the linear related assumption, one obvious outlier can be found on the top right corner which indicates an exceptionally high increase in infectious rate from 2010 to 2015. Using the linking and brushing function, we identify the outlier county is San Francisco County.

3.2 Rate mapping

Next we use quantile map to examine the spatial similarity in infectious disease rates between the two years. By choosing 6 quantiles, we can roughly learn the spatial similarity and difference between the two choropleth maps (year 2010 on the left and year 2015 on the right):

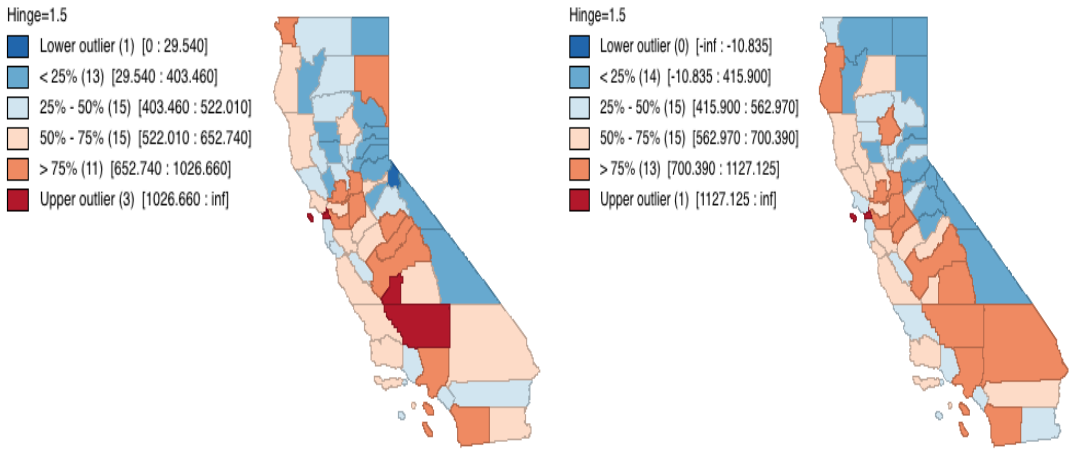


By naked eyes, we can observe the infectious disease rates in two counties drop significantly from 2010 to 2015, which correspond to the two distant points away from the regression line in the scatter plot (year 2010 in the middle and year 2015 on the right):



For the spatial autocorrelation between disease rates in 2010 and 2015, we will carry out the bivariate spatial autocorrelation analysis in Section 4.

To further examine the outliers and spatial pattern of infectious disease rates in California, we use box map with hinge = 1.5 for comparison between rate patterns in 2010 and 2015 (year 2010 on the left and year 2015 on the right):

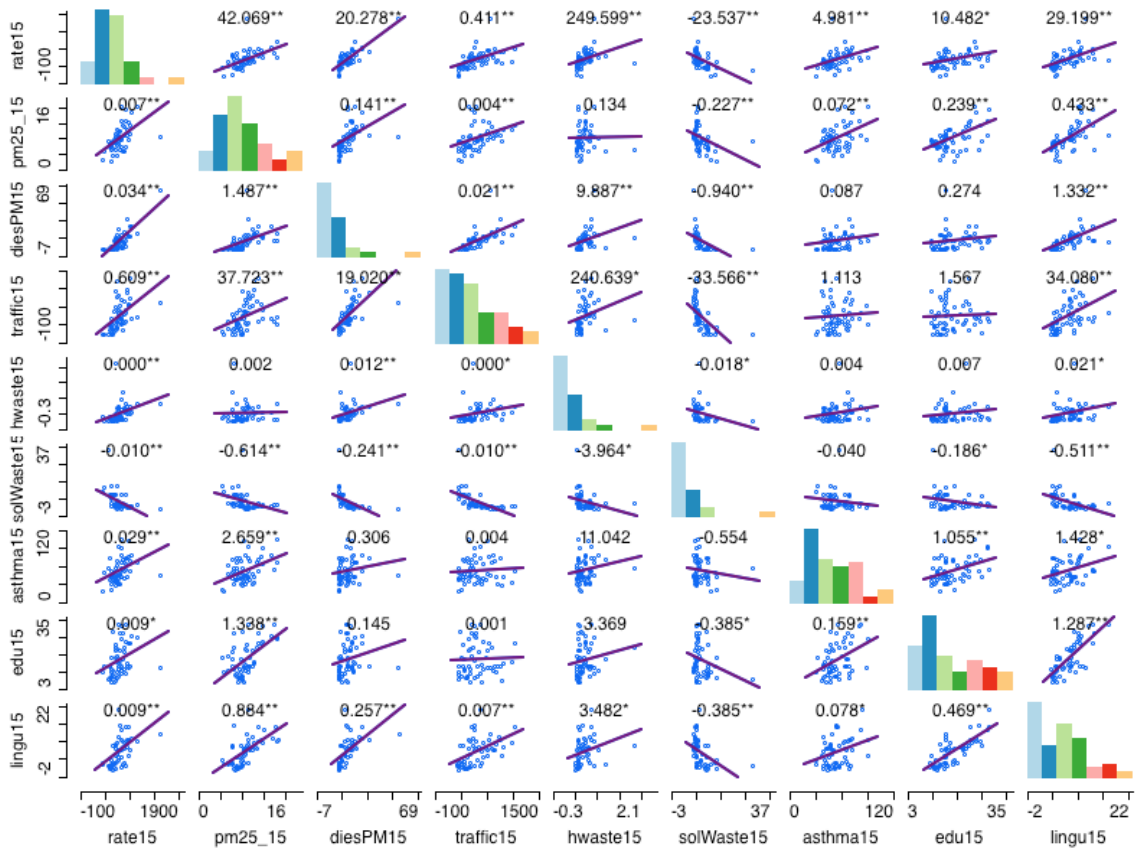


As we can see from the two box maps above, there are 3 upper outlier counties and 1 lower outlier county in 2010 while in 2015 no lower outliers and only 1 upper outlier. We can also observe in both years the northeast part of California generally had lower infectious disease rates than the southwest part, and inner counties in South California suffered from infectious diseases more than the counties in the north part in both years.

3.3 Scatter plot matrix and principle component analysis

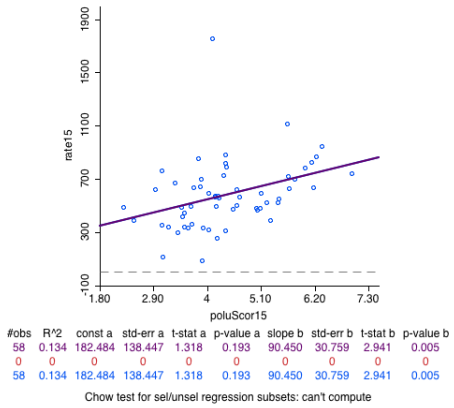
Next we turn to the potential explanatory variable analysis using a scatter plot matrix. Using the 2015 data, here we first include all natural environment variables and all population characteristic variables, and we find the following variables are significantly correlated with the infectious disease rate: PM2.5, diesel PM, traffic

density, hazardous waste, solid waste sites, asthma rate, education level, linguistic isolation index. The scatter plot matrix using these 8 variables plus the disease rate are shown below:



Using the 2010 data, we again check the correlation significant levels with infectious disease rate for each potential explanatory variables, and find exactly the same 8 significant explanatory variables. Among the 8 variables, we find many of them are highly correlated, so we reduce them to the first two principle component and find a good interpretation of these two principle components. The first component mainly summarizes the the pollution burden consisting of the first 5 variables, and the second component mainly reflects population characteristics consisting of the last 3

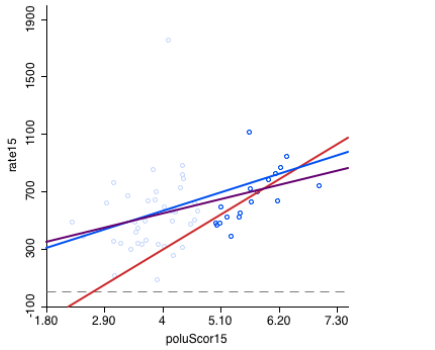
variables. Now we plot a scatterplot using the first principle component and the target disease rate in 2015:



#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
58	0.134	182.484	138.447	1.318	0.193	90.450	30.759	2.941	0.005
0	0	0	0	0	0	0	0	0	0
58	0.134	182.484	138.447	1.318	0.193	90.450	30.759	2.941	0.005

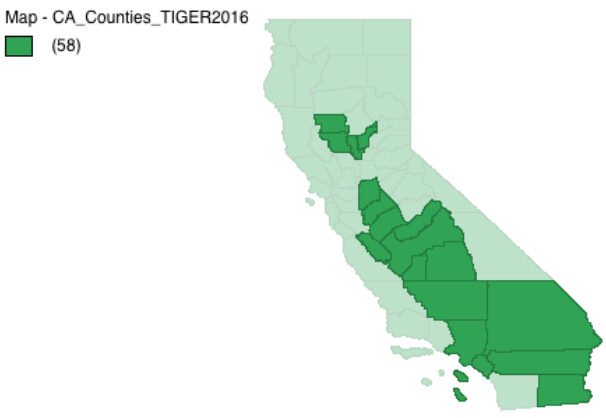
Chow test for sel/unsel regression subsets: can't compute

We can see from the scatter plot that the disease rate is positively correlated with the first component/the pollution burden score. Again, we use the linking and brushing function and check that the outlier county located high above the regression line is San Francisco County. In addition, by selecting counties with higher pollution burden scores, we detect an interesting spatial heterogeneity in terms of the regression line slope:



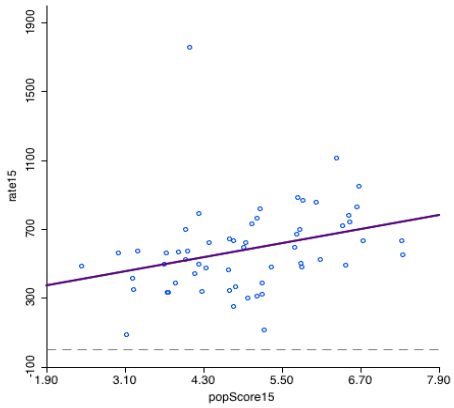
#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
58	0.134	182.484	138.447	1.318	0.193	90.450	30.759	2.941	0.005
18	0.399	-591.038	387.309	-1.526	0.147	221.645	68.032	3.258	0.005
40	0.061	92.534	288.540	0.321	0.750	117.623	75.025	1.568	0.125

Chow test for sel/unsel regression subsets: distrib=F(2,54), ratio=1.0258, p-val=0.3654

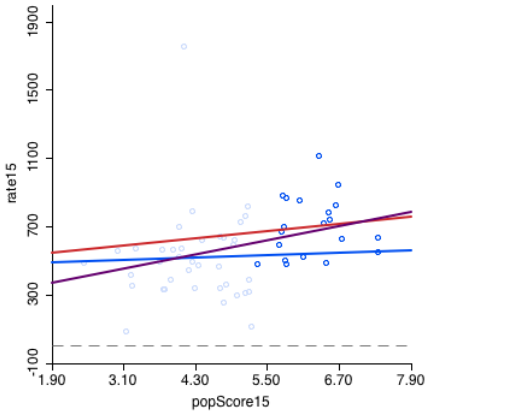


We can observe from both graphs above an interesting spatial pattern. From the scatterplot above we can see that the slope of the selected red regression line is higher than the general violet regression line, and from the right map we find that most of the selected counties with higher first principle component scores are located in the inner part of California, and the infectious disease rates in those inner counties are more sensitive to the first principle component score/the pollutant burden score. This relationship intuitively makes sense. People in those counties near the sea or near the state boundary are more likely to catch infectious disease from outside the state due to the high population mobility there. San Francisco County is a good example. From the scatterplot above we can see San Francisco County has extremely high disease rate but relatively low pollutant burden score probably due to its frequent contact with people outside the state or even people from other countries, and thus less susceptible to the local natural environment.

Next, we turn to the correlation between the disease rate and the second principle component score/population characteristic score using the scatterplot:

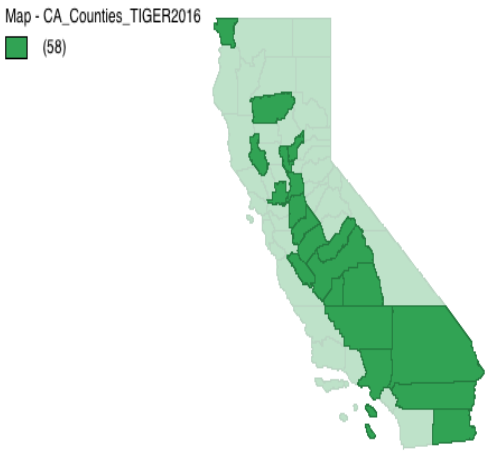


and we can see the disease rate is positively correlated with the second principle component but the correlation is weaker than the one with the first principle component, which makes sense since the second principle component summarizes the variability of the data to a less extent than the first principle component. Notice that the population characteristic scores are given in the opposite way as usual practices. The higher the score, the lower the individual' s social-economic status. Then we select counties with higher population characteristic score and observe another interesting spatial heterogeneity in terms of the regression line slope:



#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
58	0.096	240.250	142.937	1.681	0.098	68.450	28.148	2.432	0.018
20	0.012	477.161	470.574	1.014	0.324	35.735	75.137	0.476	0.640
38	0.001	467.704	269.178	1.738	0.091	11.104	62.221	0.178	0.859

Chow test for sel/unsel regression subsets: distrib=F(2,54), ratio=0.8445, p-val=0.4354

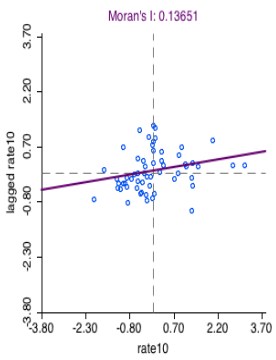


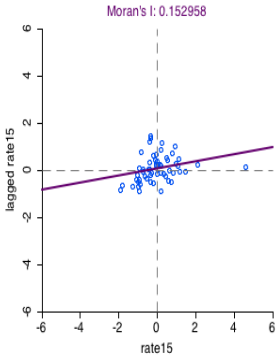
From the two graphs above we can see counties with higher second component score/population characteristic score are also mostly located in the inner part of California. The regression line slope for those counties is lower than the overall slope, and the slope for the rest counties in the boundary areas of California is even lower, which indicates that infectious disease rates have weaker correlation with the population characteristic score than with the pollutant burden score, especially in counties located in the boundary areas of California.

4. Spatial autocorrelation

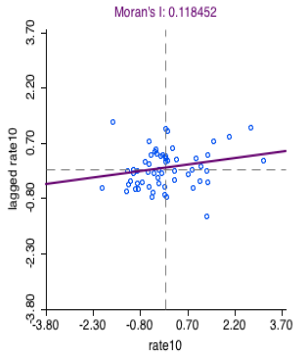
4.1 Global spatial autocorrelation

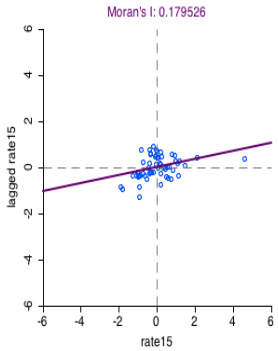
In Section 3, we observe some spatial pattern of infectious disease rates in California using quantile and box maps in 2010 and 2015. Now we first conduct global spatial autocorrelation analysis to measure the deviation from the random spatial pattern for California infectious disease rates in both 2010 and 2015 respectively. The Moran scatter plots using the first order queen contiguity weight for both years are shown below (year 2010 on the left and year 2015 on the right):





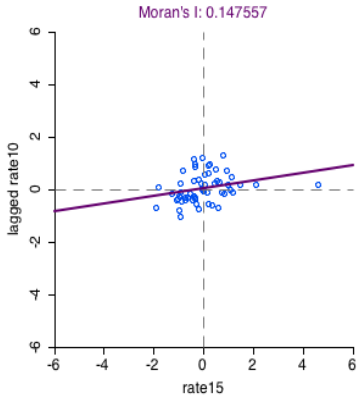
Both Moran's I scores are significantly bigger than expected values under the random spatial pattern null hypothesis, which indicates positive global spatial autocorrelation in both years. Next we choose the distance based weight using the default distance threshold for Moran scatter plots in both years (year 2010 on the left and year 2015 on the right):





The two Moran scatter plots above confirms the non-randomness of the spatial patterns of the infectious disease rates in 2010 and 2015.

Now we use Bivariate Moran Scatter plot to measure the space-time correlation between disease rates in 2010 and 2015:

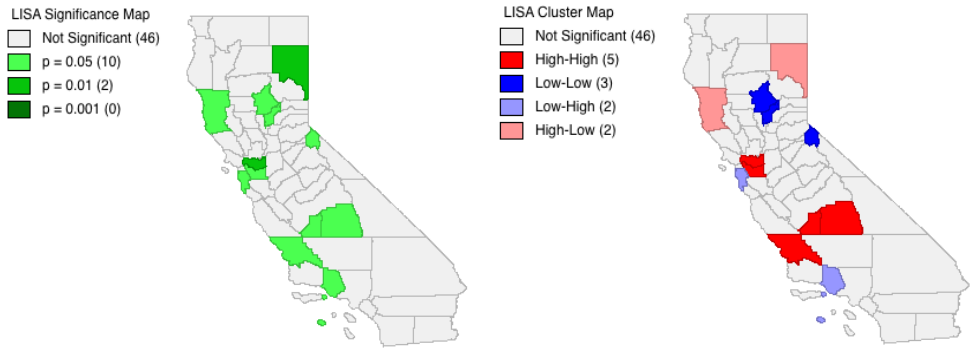


AS the map above shows, the Moran' s I confirms the hypothesis of global spatial autocorrelation between infectious disease rates in 2010 and 2015 as the quantile and box maps in Section 2 indicate.

4.2 Local spatial autocorrelation

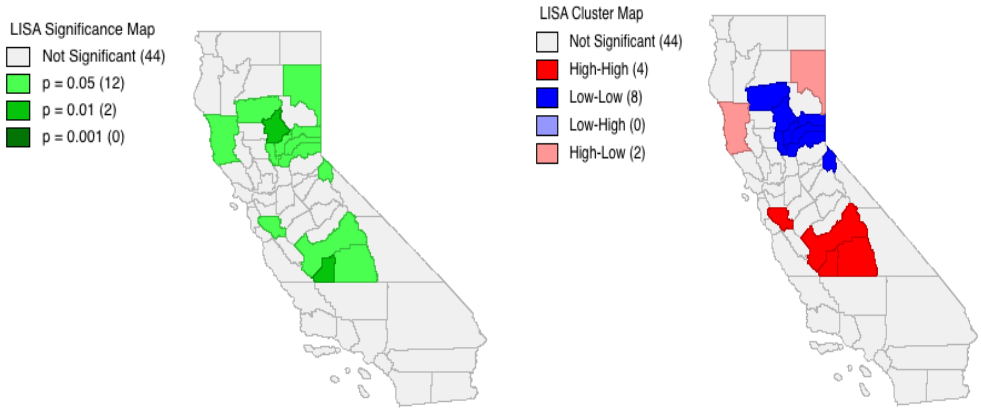
Now we turn to local cluster identification for infectious disease rates and the two principle components by using Local Moran cluster maps and significant maps.

First we examine the local clusters of disease rates in 2010 using the queen contiguous weight:



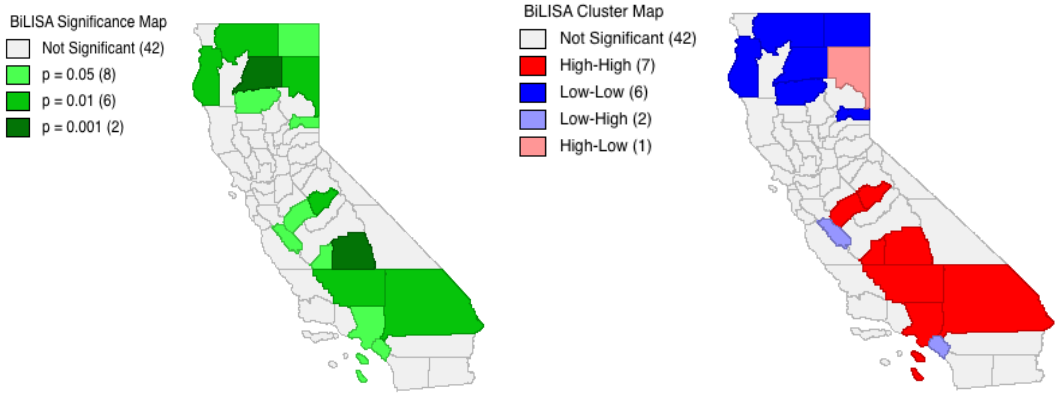
The left map shows the locations with a significant local statistic. We can see there is no county significant at $p<0.001$, 2 at $p<0.01$, and 10 at $p<0.05$. From the local cluster map on the right, we find an interesting spatial pattern that all high-low and low-high outlier counties are located at the boundary area of the state. All high-high clusters are located in Central California and all low-low clusters located are located in North California.

Next we try to identify the local clusters for infectious disease rates in 2010 using the distance based weight:

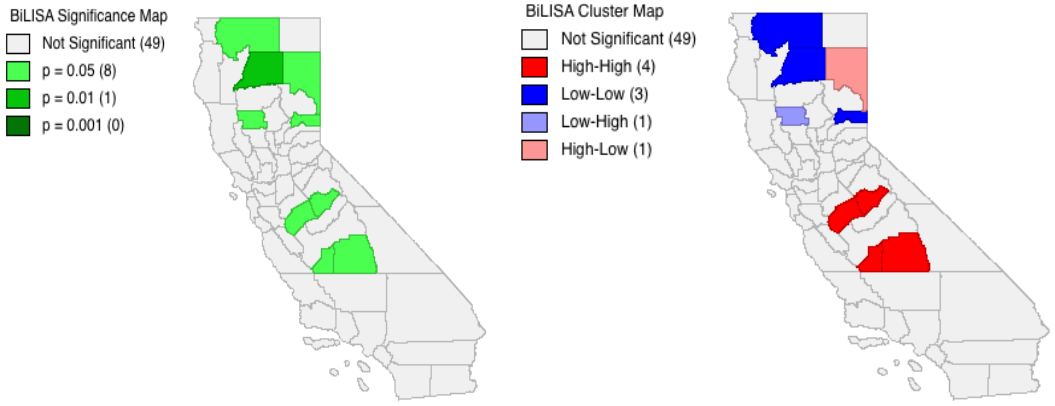


Compared to the local cluster map using queen contiguity weight, the one using distance based weight comes with more significant counties at level $p<0.05$. The number of low-low clusters increases from 3 to 8 and high-high clusters decreases from 5 to 4, and low-high outliers disappear. Again all low-low clusters are located in North California while all high-high clusters are now located in Central California, and most of low-low clusters and high-high clusters are located in the inner part of California.

To find out potential explanation for the local cluster patterns, we carry out bivariate local cluster analysis between the first principle component/the pollutant burden score and the disease rates in 2010 using the queen contiguity weight:

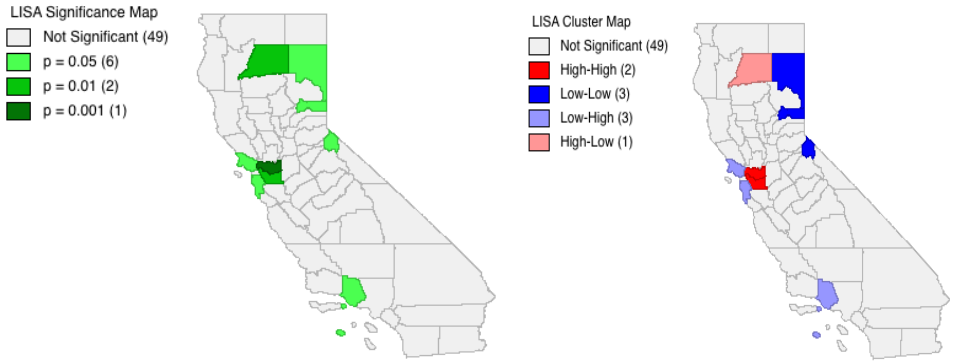


From the two maps above we can find out some spatial features in the bivariate local cluster map: all the low-low clusters are located in North California and all high-high clusters in Central and South California. Totally 16 counties are significant with 2 at $p<0.001$ level. Similarly, we carry out bivariate local cluster analysis between the second principle component/the pollutant burden score and the disease rates in 2010:



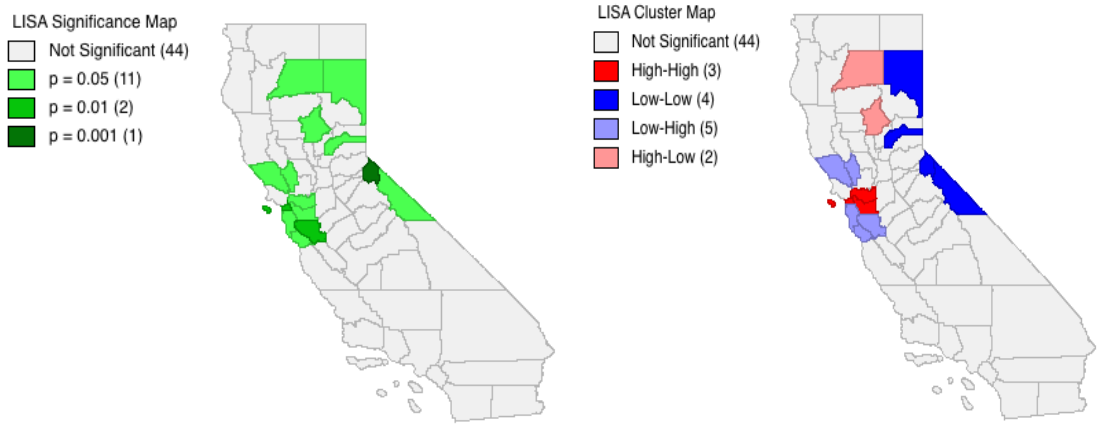
Now the number of significant counties decreases to 9, which confirms the weaker correlation between the infectious disease rate and the second principle component. Again we find some spatial patterns of the correlation: all the low-low clusters are located in North California and all high-high clusters in Central California.

To see any temporal change of the local cluster pattern, next we check the local cluster pattern of infectious disease rates and its correlation with the first two components in 2015. First we examine the local cluster map of infectious disease rates in 2015 using the queen contiguity weight:



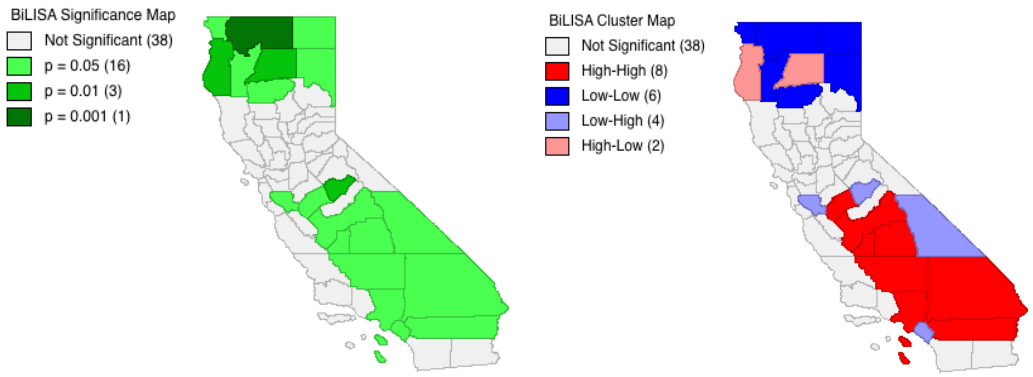
Compared to its counterpart in 2010, the number of high-high clusters decreases from 5 to 2 and the two high-high clusters near the Bay Area still exist while the other three high-high clusters used to be in the south California disappear. 3 low-low clusters remain with one low-low cluster on the east boundary unchanged but the other two low-low clusters shift to the northeast. The spatial pattern of local clusters in 2015 can be summarized as following: all low-low clusters are located on the east boundary of California and the only two high-high clusters are still in the Bay Area while all the high-high clusters used to be in South California now disappear.

Next we examine the local cluster pattern for infectious disease rates in 2015 using the distance based weight:

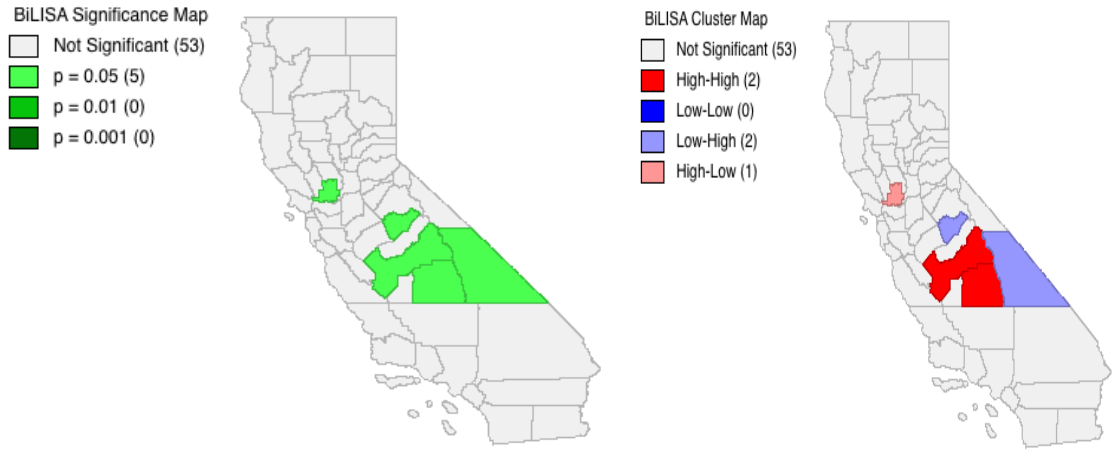


Compared to the local cluster map using queen contiguity weight, the one using distance based weight has more significant counties at level $p<0.05$ the number of which increases from 9 to 14. The number of low-low clusters increases from 3 to 4 and high-high clusters increases from 2 to 3, and both types of outliers increase as well. Again all low-low clusters are located in east boundary of California while all high-high clusters are still located in Central California but now include the San Francisco County.

Then we carry out a bivariate local cluster analysis between the first principle component/the pollutant burden score and the infectious disease rates in 2015:



Compared to its counterpart in 2010, now the number of significant counties at level $p<0.05$ increases from 16 to 20, which indicates a stronger correlation between the infectious disease rates and the first principle components in 2015. The spatial pattern of this local cluster map can be summarized as following: all low-low clusters are located in North California while all high-high clusters are in Central and South California. Similarly, we carry out a bivariate local cluster analysis for the second principle component/the population characteristic score and the infectious disease rates in 2015:



We can see from the maps above that the number of significant counties decreases to 5, all at level $p<0.05$ but none at level $p<0.01$. Compared to both the local cluster

map using the first principle component and the local cluster map counterpart in 2010, the number of significant counties drops, which indicates the correlation between the infectious disease rates and the second principle component further decreases in 2015. Low-low clusters no longer exist and only two high-high clusters remain there.

5. Conclusion

After all the exploratory data analysis with geo-visualization in Section 3 and spatial autocorrelation analysis in Section 4, we have identified the following spatial patterns of infectious disease rates in California with temporal differences between 2010 and 2015:

1. The inner counties in California have significantly higher infectious disease rates than counties located on the state boundary in both 2010 and 2015.
2. Counties in North California generally have lower infectious disease rates than counties in South California, and counties in Central California have the highest infectious disease rates in both 2010 and 2015.
3. The inner counties in California come with higher correlation between the infectious disease rate and the first two principle components (the pollutant burden score and population characteristic score) than counties on the state.
4. The spatial patterns of the 2010 infectious disease rates and the 2015 infectious disease rates are significantly correlated with high spatial overlap.

5. The correlation between the infectious disease rates and the pollutant burden scores is stronger than the correlation between the infectious disease rate and the population characteristic score in both 2010 and 2015.
6. The correlation between the infectious disease rates and the pollutant burden scores further increases in 2015 compared to the one in 2010, and the correlation between the disease rate and the population characteristic score further decreases in 2015 compared to the one in 2010.