A SPATIAL ANALYSIS ON THE ACCESSIBILITY AND SOCIOECONOMIC DIFFERENCES OF US AIR TRANSPORT ACROSS COUNTIES

Yigong Wang
GIS Final Project
12-8-2019
yigong@uchicago.edu

I. Introduction

The United States has the most advanced aviation industry in the world, in terms of both commercial and civilian aviation. There are more than 19 thousand aircraft landing facilities (airports, airstrips and helipads) in the US and almost every moderate town has its own airport. For big metropolitan areas, one could easily find over 10 airports serving the area. This scale of aviation infrastructure is not seen anywhere else in the world, and the massive span of aviation also gives us the opportunity to look at the differences in usage of aviation in communities with different socioeconomic and demographic profiles.

How are the distribution of airports in the contiguous United States related to the population structure on the county level? In this research, we aim to answer how aviation usage is different in communities (county level) with different geographic and demographic background, and how this could be indicators on how well a job the government has done in improving access to remote communities and how socioeconomic background could also contribute to the differences in usage of aviation.

II. Data Collection

For this project, I used data from several different sources: the main data set comes from the *Homeland Infrastructure Foundation - Level Data (HIFLD)* open data source (hifld-geoplatform.opendata.arcgis.com), from where I download the *"Aircraft Landing Facilities"* dataset, from which I select data only for the contiguous US (48 states). The data is for the year 2018. This dataset is useful as it contains very detailed information regarding US airports (annual operations, passenger number, based aircraft number, detailed geographical information including county and street level information), and it is a comprehensive collection of information regarding all US aircraft landing facilities. On the demographic information side, I then collect data on household income, per capita income, population, racial structure (percentage of each race in county) for all contiguous 48 states. This data comes from census.gov and we choose the 2010 census data.

The aircraft landing facilities dataset contains 19803 entries(meaning that there is 19803 airports/ helipads/airstrips, and the census dataset contains 3221 entries(meaning that there is 3221 counties/ independent cities in the United States.) We then use Excel to match these two datasets (primarily using VLOOKUP function) and aggregate results from the aviation dataset. For our end results, we record 3085 counties in the contiguous US (extensive data cleaning and county matching from both datasets are completed in project 2, in which details on how these procedures are conducted are noted.) Here I quote an excerpt from project 2 to show some examples of how I merge and aggregate the two datasets into the final dataset:

> "Trying to merge the two data sets proved to be quite a challenge. First, although both data sets contain county information, the department of transportation does not stick to conventions when it comes to counties. The airport data and the US Homicide data does not notate county names the same way, as one dataset leaves in all the space, cases, symbols in county names and the other does not. The US Homicides also omits a lot of the "non-contiguous but also a part of the contiguous US" counties such as San Juan county in Washington. There are also independent cities and communities at the county level which the US Homicide dataset completely ignore, which I had to manually refit into the data structure. In addition, when an airport is on a state or county border, it often goes misclassified to the

neighboring area, and over the years, some counties have disappeared and some new ones have emerged, which took me quite a bit of due diligence and research to manually correct the data.

The completed dataset that I use for this project now has the following variables:
`PLOY_ID`: a GeoDa generated integer ID for each entrance (county);
`County`: county name of each entrance;
`State`: state name of each entrance;
`PCA_IN`: per capita income;
`M_HH_IN`: median household income;
`M_FAM_IN`: median family income;
`Population`: population number;
`#HH`: number of households;
`%WHITE`: percentage of population in the county that is White;
`%BLACK`: percentage of population in the county that is Black;
`%ASIAN`: percentage of population in the county that is Asian;
`%AINDAL`: percentage of population in the county that is native American and Alaskan;
`%PACIS`: percentage of population in the county that is Pacific Islander;
`%LATIN`: percentage of population in the county that is Hispanic or Latino;
`%OTHER`: percentage of population in the county that is other races;
** The racial percentage data might add up to more than 100% as in the census, an individual can indicate more than one race;*
`APTCNT`: number of airports in the county;
`BASED_ACFT`: number of aircrafts based at airports in the county;
`ANNUAL_OPS`: annual operations of aircrafts (takeoffs and landings) in the airports in the county;
`EMPLANEMENT`: number of passengers/pilots taking flying trips annually in the county;
`FREIGHTLBS`: freight processed (in pounds) annually in the county's airports;
`MAILLBS`: mail processed (in pounds) annually in the county's airports;
`aptpo`: number of airports per capita;
`basepo`: number of aircrafts based at airports in the county per capita;
`opspo`: number of operations of aircrafts (takeoffs and landings) in the airports in the county.

This dataset gives us good information regarding income, racial structure of population, and aviation usage in counties. Variables that I concentrate on are `PCA_IN Population %WHITE %BLACK %ASIAN %LATIN aptpo basepo opspo`.
** The aviation variables are *not* spatially extensive because even though all of them are aggregated occurrences (otherwise know as "counts"), they are not spatially dependent because even though California counties vary in sizes, each county has similar amount of population (with outlier which I discuss later) and aviation operation is generally spatially independent of county sizes as we discuss later.

We also focus on the State of California for this analysis in the later stages due to multiple reasons: first, GeoDa's capability of cluster analysis in not scaling well for datasets with large datasets at this stage of development, hence limiting us to a smaller geographical area; with that in mind, California has one of the most complex geographical forms (with counties located on high mountains, around water, and with vastly sizes and population structures), and it is one of the largest states in the US (population is number one in the US): this complexity will ensure the validity of our methods and will make it easier to scale this study to the entire country in the future.

III. Exploratory Data Analyses

Using scatter plots and histograms, we find that the income variables are highly related, hence using just one of the variables would be sufficient, yet the three airport variables are not highly related, that means a community that has high number of airports per capita might not have a lot of aircrafts based in the airports (aircraft ownership) etc. However, `basepo` and `opspo` are highly related. (This reduction in dimensions would also be further discussed.) We also find some preliminary associations across different variables: counties with high percentage of white populations generally have higher income, and counties with high percentage of black population generally have lower income levels. In terms of aviation, however, this racial trend is weaker: counties with higher white population and lower black population generally have more aviation activities, but the variance of the data is a lot higher. The high aviation operations counties are almost exclusively white counties (this is consistent with the traditional image of old white men piloting general aviation airplanes), yet the low aviation operations counties have no set racial demographic characteristics. The more interesting *disassociation* occurs in the *income* side: there is no clear association between having high income and having high aviation usage, as a matter of fact, the scatter plots show that the counties on the extreme high side of aviation operation are actually low income counties. This might be due to that mountainous areas have higher usage of aviation as other means of transport are generally less accessible to these communities, and high mountain areas are associated with lower income levels.

Some useful scatter plots are combined into ***figure group 1****.*

Plotting simple choropleth maps (6 natural breaks univariate maps), we first look at aviation related variables across the contiguous US counties (these variables are `aptpo basepo opspo`), and we can see a couple of characteristics of aviation facilities:

- The number of aircraft landing facilities and population number is generally consistent across the country - the more people a county have, the more airports it have;
- On a per capita bases, the trend is the opposite: the less people a county have, the more airports this county have on a per capita basis;
- Low income areas have higher aircraft operations: (see figure 3 plots on the second column);
- High white population areas have higher number of aircraft ownership: (see figure 3 plots on the third column);
- Rural areas have higher aviation usage on the per capita scale (see figure 2 and figure 3). This is shown in figure 2 in the central mountainous regions of the US (especially in the Idaho - Wyoming - Utah regions); in California, this is shown in the Sierra-Nevada Mountain regions and Mount Shasta regions (east and central-north regions of the state);
- There are also some interesting findings regarding California counties themselves, albeit not strictly related to this project, but they are meaningful regardless:
    - California counties vary hugely in geographical area, but generally the population in each county is similar;
    - There is an exception: Los Angeles County (and surrounding areas), as it has way more population than the other counties.
    - The Bay Area has significantly higher income level compared to other areas of the state.

These maps are show in ***figure group 2*** *and* ***figure group 3****.*

IV. Advanced Analyses

In order to statistically confirm some of the findings outlined in section III, we conduct some spatial-statistical techniques in this section. Due to the limitations of GeoDa mentioned in section II, I will only use California as my area of analysis in this section.

First, we need to create some spatial weights. For this project, we explore various different methods in spatial weights, including queen and rook contiguity weight, nearest neighbor weight, and distance band weights. After I test out these weights with some exploratory spatial autocorrelation analyses, we decide to use *4-nearest-neighbor weight* for the majority of this section. This is due to the unique geography of California, especially in the Bay Area. The counties of San Francisco Bay Area are located around the San Francisco Bay, which makes them not arrayed together like conventional counties, hence dramatically reducing their spatial contiguities. Since the Bay Area is also a very important area in our analysis, we need to circumvent this anomaly by using distance weight, and we find that using 4-nearst-neighbor is the best distance weight method.

a. Dimension Reduction

There are 24 variables in this dataset, which is a large amount. We need to pick out relevant information regarding aviation and demographic data. We can see that many of the variables are similar or highly related to each other, such as racial data where the percentage of white population in a community is inversely related to black population percentage, and we can also see that the various income variables, such as income per capita, and median household income and the other variables are also highly related, and the population number is also highly similar to the number of households in a community. Hence we need to select out some of these similar variables, in order to make our analyses a little bit less cumbersome. After drawing many scatter plots, LISA univariate analyses, and other exploration, I decided on two to three variables in each of the indicator categories: `PCA_IN` and `M_HH_IN` in the income category, and `%WHITE`, `%LATINO`, and `%ASIAN` for the racial variables, and all three of the aviation population adjusted variables `aptpo basepo opspo`. These will be sufficient in running the analyses and provide a comprehensive variation in each of the aspects: each of these variables are different in their own ways (the two income variables show both mean and median, and the three aviation variables are each different, and the variables reflect the typical demography on our geographic area of analysis: California, where I select the three main races in the state.

We then run PCA with SVD to determine the influence of each variable, which turned out to be surprising. The first two principle components explain close to 60% of the variation, and the first four principle components explain almost 90% of the variation in the dataset. One other interesting finding is that it is not until the third and fourth component does the aviation variables start to play any role of importance. This analysis means that the traditional socioeconomic and racial indicators play a lot more important role in the explanation of spatial variance in California compared to aviation, which is not surprising as aviation is just one aspect of the society. This does not affect our analysis, though, because our research question is not how important is aviation compared to racial variation in communities, but we simply want to find how these two might be related. It is surprising, however, to see that aviation still played a somewhat impactful role in the variations in this dataset as the aviation variables are significant in the third and fourth principle component.

We save the results of PCA analysis (saving the first four principle components.) Looking at PC1 to PC2 scatter plot, we can see that the slope is zero with a close to circular distribution of points, that means that the PCs are independent, which is a good thing. We also look at a Moran local

autocorrelation analysis of PC1, together with the PCP of four most impactful variables in this PC, and we find that the low-low clusters follow the same directional shape in the PCP, but the high-highs do not.

*These graphs are shown in **figure group 4**.*

b. Cluster Analysis

For cluster analysis, we use the six variables mentioned in the previous section to conduct different means of cluster analyses:

Looking at the cluster map of unconstrained K-Means classic clustering with 4 clusters (KMeans++) The four clusters are dispersed in these relative geographic areas:

- LA and Bay Area
- Central/Southern California
- Rural Coastal and Northern California
- Sierra Mountains and Mt. Shasta.

These clusters are very consistent with the geographical and demographical distributions.

I then amend the K means clustering with a spatial component (geometric centroid, with geometric weight of 0.625.) The clusters then turn to a very uniform north-south cluster. However, the Bay Area still remained its own cluster, and this strong conference of the Bay Area continues into the following clusters as well.

We then look at the skater mechanism. Using the default Standardize (Z) transformation, we can see that even when we specify four clusters, we almost only get two: Bay Area vs. the rest of the state. Hence I then tested different transformations. (In the graphs, they are Standardize (Z), Raw, and Standardize (MAD) from left to right.) We find that the northern parts of California increasingly falls into a cluster as we change the transformation methods.

Finally, we test max-p, which is yet another spatially constrained clustering method. We are using greedy local search for this clustering. One interesting discovery I make is that changing the minimum bound variable has significant impact on the clusters. For this dataset, setting different seeds does not make much of a difference in the results. We do two variations of the max-p clusters. The first one uses airports count and the second one uses per capita income as the minimum bounds, and both were set to 10%. In both methods, both C2 (northern California) has the highest within cluster sum of squares. Finally, we compare the ratio of between to total SS.

Looking at the cluster statistics, we can see that Max-p with the proper selection of minimum bound (Income) has the closest ratio compared to the unconstrained K Mean.

*The graphs of cluster analyses and each clusters' statistic are presented in **figure group 5**.*

c. Spatial Autocorrelation Analysis

With all the previous information and analyses, we conduct some local spatial autocorrelation analyses to verify and explore further detail through the clusters. We test different local spatial autocorrelation techniques using all 9999 permutations, 0.05 significance level, but with different methods: we test

Moran's I, Getis-Ord, and Local Geary. It turns out that all three methods return similar results, hence we will stick with the classic Moran's I LISA in this report.

We first run LISA on income and population data, and we find that the Bay Area has a high-high cluster, meaning that it has significantly higher income levels compared to the rest of the state. We then see that the northern part of the state has a low-low cluster of income levels. The low-low income level area is almost identical to the cluster map created using skater(raw) algorithm in the previous section.

The second LISA cluster plot is of the population in counties. We see that there are almost no significant areas with the exception of LA county and some Sierra-Nevada mountain counties. This is consistent with our discovery that the California population is largely similar across counties, with the outlier of LA county.

We then look at four LISA cluster maps of the three aviation variables `aptpo basepo opspo`. One new discovery that we make is that the southern part of the state has pretty low aircraft operations per capita. This could simply be due to the large number of population in Southern California, or it has something to do with another variable which we will find out about in the next part of the analysis. (Proportion of Latino population.) The other LISA plots large confirm the discoveries made in previous sections. We even run a Bivariate Moran's I LISA and the results are still in the lines of previous discoveries.

*The graphs are shown in **figure group 6**.*

d. Regression Analysis

Just out of curiosity, we played with the regression functionality in GeoDa on my dataset, since we do want to find out the association between the aviation variables and the socioeconomic/demographic variables. I run a classic linear regression (OLS) with `opspo` as the dependent variable. The independent variables include `M_HH_IN %WHITE %ASIAN %LATIN`. The regression results are within the intuitive understanding and are attached below:

```
Dependent Variable  :       opspo  Number of Observations:   58
Mean dependent var  :    0.678444  Number of Variables   :    5
S.D. dependent var  :    0.684889  Degrees of Freedom    :   53

R-squared           :    0.225256  F-statistic           :     3.85243
Adjusted R-squared  :    0.166785  Prob(F-statistic)     :  0.00805156
Sum squared residual:     21.0779  Log likelihood        :    -52.9441
Sigma-square        :    0.397696  Akaike info criterion :     115.888
S.E. of regression  :    0.630631  Schwarz criterion     :      126.19
Sigma-square ML     :    0.363412
S.E of regression ML:    0.602836

------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------
          CONSTANT       1.97941        1.45043         1.3647     0.17811
           M_HH_IN      -1.29843      0.0782111       -1.66016     0.10279
             %WHITE    -0.018749      0.0151415      -0.123825     0.90192
             %LATIN    -0.112885     0.00672003       -1.67983     0.09888
             %ASIAN    -0.0159469      0.0226296      -0.704691     0.48409
------------------------------------------------------------------------
```

The regression results show that these variables do have association to the aviation data. The F-test result is very ideal, yet R squared result is less than ideal. Looking at individual variables, M_HH_IN and %LATIN come close to statistical significance, where as %WHITE has almost no association with airplane operations. The coefficients show that the less the median income, the more aviation operation, and the less the percentage of latino population, the more the aviation operation. The association on %LATIN explains the low aircraft operations per capita in Southern California discovered in the previous section.

V. Conclusion

Throughout this analysis, we look at different demographic and socioeconomic indicators in understanding the spatial heterogeneity of aviation in the US, and particularly in the state of California. We find that airport usage (number of based aircrafts and aircraft operations) are somewhat inversely related to income, and positively rated to race (especially proportions of white population). In the state of California, we find that:

a. Bay Area stand out as a high income, low per capita aviation area;
b. LA Area stand out as a high population, low per capita aviation area;
c. Rural and mountainous parts of the state have high amount of aviation activities (per capita).

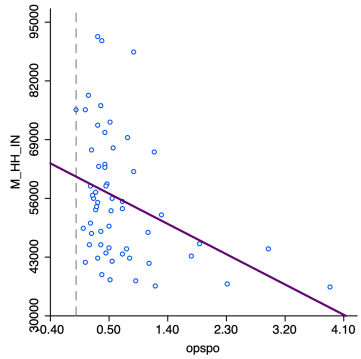
Aviation is a dynamic field, and many aspects come into play when it comes to aviation usage. It is comforting to see that the rural areas have higher participation in aviation, hence social equality is somewhat preserved through this means of transportation. Although we do not gather much statistically significant information, we still find some interesting pattern and have a better understanding of how aviation helps rural communities.

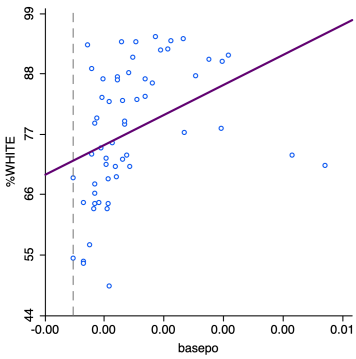
VI. Discussions and Future Analyses

There are many factors in the aviation industry because it is highly dynamic, hence it is not surprising that we do not find obviously significant results. We need further analyses to pick out the noises and enhance our dataset to better run the statistical tests. We find that a lot of the aviation results seem to have something to do with elevations of the airports, which is one thing we can develop on in the future. For this project, elevation data is too hard to compute as an algorithm in computing mean elevation of airports in a community is complicated.

Why do I not discuss gender? Aviation is considered to be a highly macho industry, and there are not a lot of women flying. But this data is also very hard to collect on a county level as almost all counties almost all have equal number of male and female. This is another aspect future analyses could develop on.

Appendix: Graphs and Figures



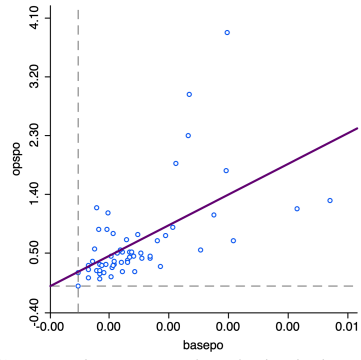


'obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value
58	0.133	60596.296	2438.182	24.853	0	-7426.229	2529.149	-2.936	0.005
0	0	0	0	0	0	0	0	0	0
58	0.133	60596.296	2438.182	24.853	0	-7426.229	2529.149	-2.936	0.005

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

#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	0.147	72.148	2.289	31.526	0	4201.943	1352.271	3.107	0.003
0	0	0	0	0	0	0	0	0	0
58	0.147	72.148	2.289	31.526	0	4201.943	1352.271	3.107	0.003

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



#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	0.340	0.214	0.114	1.878	0.066	361.894	67.369	5.372	0.000
0	0	0	0	0	0	0	0	0	0
58	0.340	0.214	0.114	1.878	0.066	361.894	67.369	5.372	0.000

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

figure group 1 - Scatter Plots

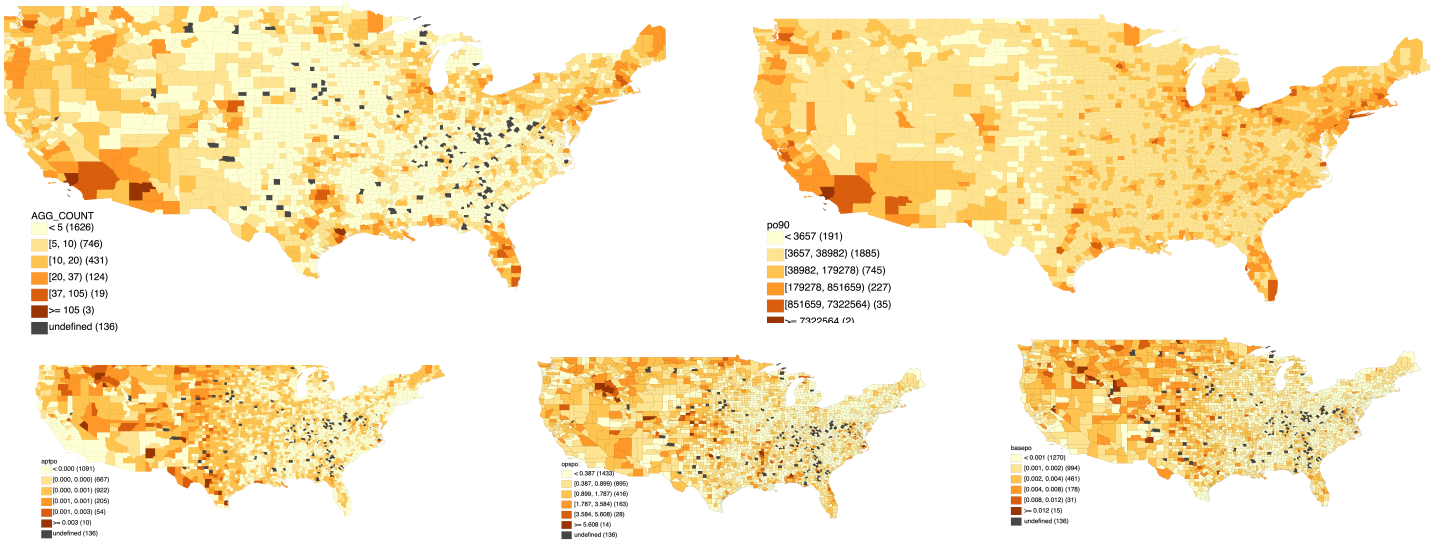


figure group 2 - Choropleth maps of aviation and socioeconomic/demographic data in contiguous US

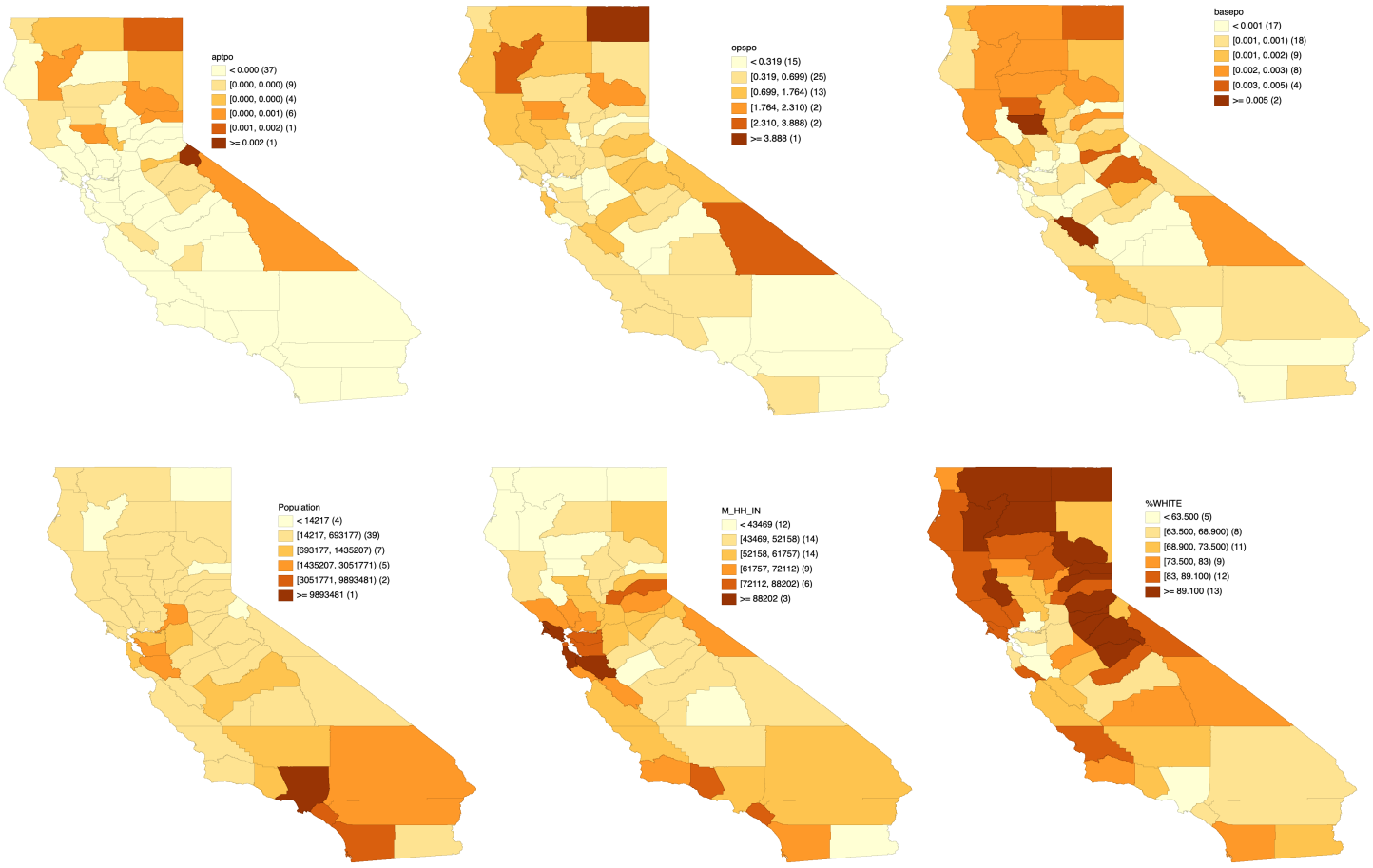


figure group 3 - Choropleth maps of aviation and socioeconomic/demographic data in California

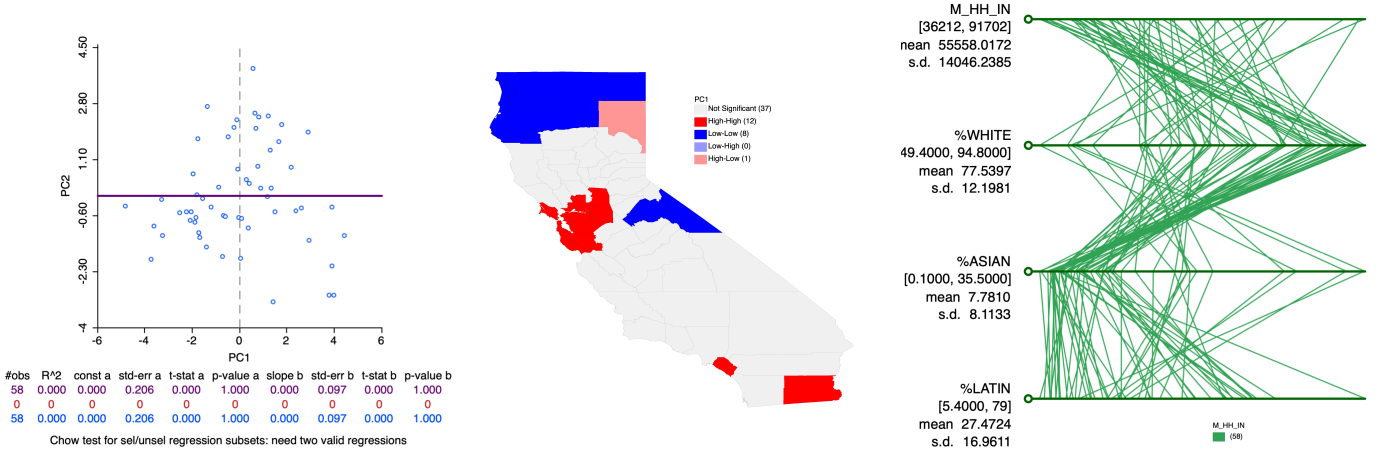
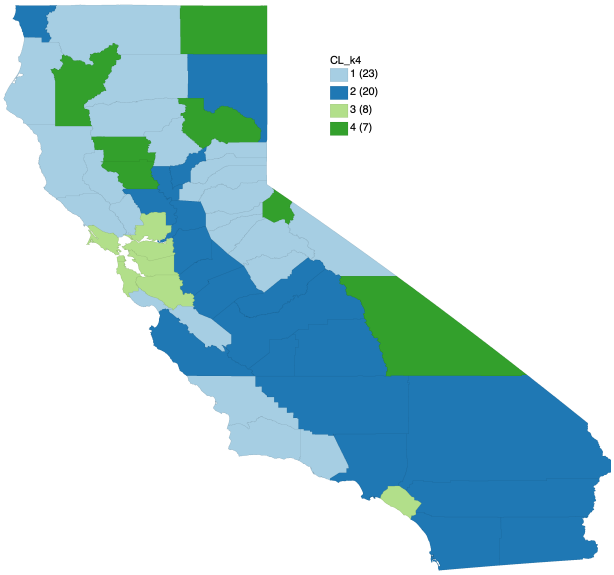
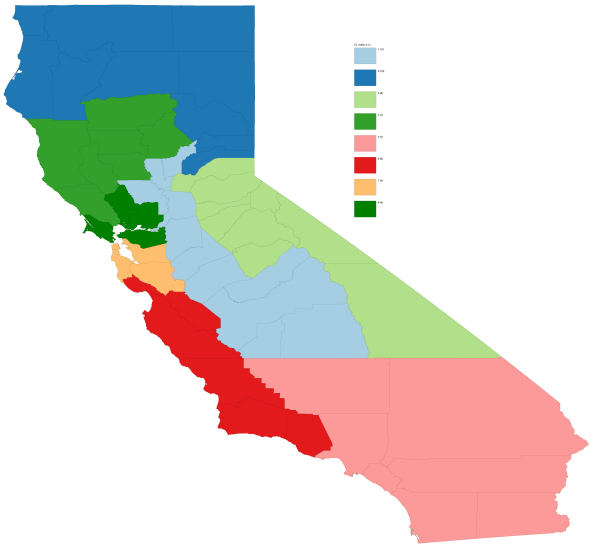


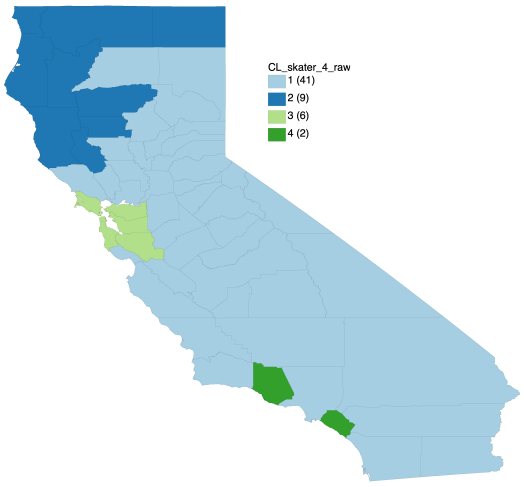
figure group 4 - Principle component analysis with PC1vsPC2 scatter plot, Local Moran's I Cluster PC1, and associated poly-coordinate plot



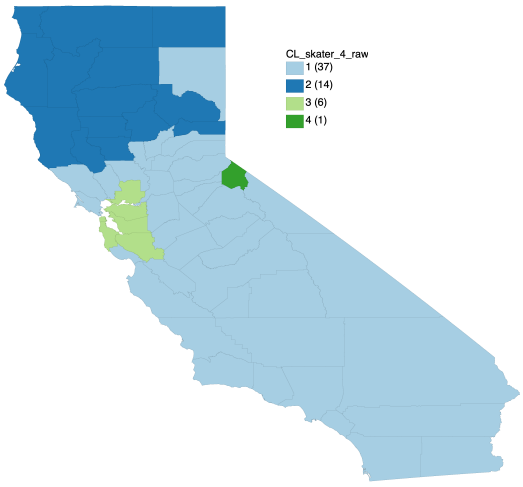
K Means



max-p (Income)



Skater (Raw)



Skater (MAD)

Clustering Method	Between SS/Total SS
K Means	0.559
K Means (spatially constrained)	0.288
Skater (Z)	0.395
Skater (Raw)	0.671
Skater (MAD)	0.443
Max-p (Airport)	0.441
Max-p (Income)	0.507

figure group 5 - Clusters generated with different methods and statistics for all methods tested.

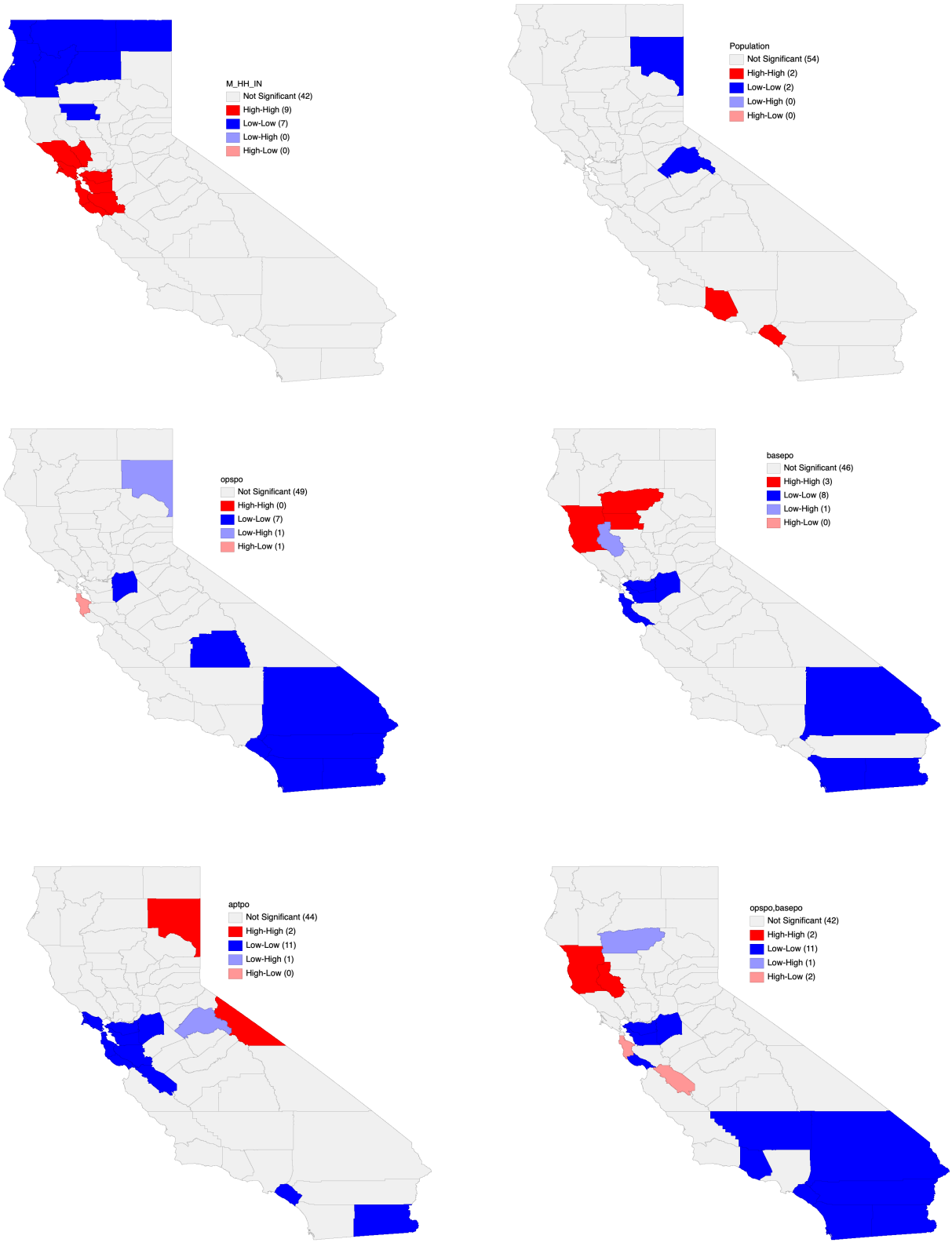


figure group 6 - LISA plots for analyses in section IV c.