Urban Livability for CUSP graduates in NY & NJ

Zhiao Zhou, Lingyi Zhang

1. Abstract

As current NYU CUSP students are going to graduate in no more than a year, some of them would be busy finding a new livable place where there might well be relatively more data science-related job positions, which has not been investigated before. In this study, income per capita, the time taken to travel to work, unemployment rate, related occupation rate and median age by Block Group from U.S. Census of the calendar year 2011-2015 are used as the 5 most important livability indicators for those students, equally weighted. In addition, Moran's I global and local autocorrelation techniques in ArcGIS and Geoda could be used to find clusters of data science career oriented livability (which would be stated as "livability" below) in New York and New Jersey. Consequently, from the outputs of a global Moran's I, it turns out that the livability is significantly and highly autocorrelated in NY and NJ, based on a p-value of 0.001 and a Moran's I statistics of 0.482429. At last, this leads to a local Moran's I which shows that high livability areas tend to be located in New York and especially across the coastline of Long Island and long beach island and around lower Manhattan and low ones tend to be scattered in New Jersey.

2. Data and methodology

2.1 Datasets

In this study, census data on the block group level is used, from the 2011-2015 American Community Survey (ACS) 5-year estimates (Branch, 2017). Block Groups (BGs) are statistical divisions of census tracts, are generally defined to contain between 600 and 3,000 people. A block group consists of clusters of blocks within the same census tract that have the same first digit of their four-digit census block number (Branch, 2017). It is assumed in this study that block groups could reflect more clustering information than census tracts could do as the former are smaller geographic units. The data is open to download as geodatabase containing a file with geo-information, a metadata table with a short name and a full description of each data element and tables of each element.

2.2 Methodology

2.2.1 Determine 5 variables

Wang, J., Su, M., Chen, B., Chen, S., & Liang, C. (2011) have defined the urban livability based on three criteria, including social progress, living level, and the environmental quality. The criteria layer further divided into factor layer, and then the indicator layer. In a similar way, the livability in this study is defined based on one criterion -- whether it's good for career development or not. Then 5 most important variables are chosen based on based on face-to-face interviews with several CUSP students as is shown in Table 1. As it's hard to validate which factor is more important in our analysis, each indicator is weighted equally here.

Table 1. Indicator system for urban livability

Definition of urban livability	Criteria layer	Factor layer	Indicator layer	Weights
	Good for Career Development	income	Income per capita	0.2 (+)
		commuting time	Time taken to travel to work / min	0.2 (-)
		unemployment rate	Unemployment rate	0.2 (-)
		occupation	Related occupation rate	0.2 (+)
		age	Median age	0.2 (-)

Note: (+) means positive influence, (-) means negative influence.

2.2.2 Data Processing

First, the data is added in ArcGIS from datasets as in Table 2 then “Delete Field” toolset in ArcToolBox is used in order to drop useless fields in each table of elements besides the 5 variables chosen. Then they are merged together and exported as csv format. Additionally, Python is used to first convert all data into density and normalize the 5 factors using Student's t-statistic and then calculate the livability. Then the geo-information data of NY and NJ are merged using the “Merge” toolset in ArcToolBox and then joined again with the data of 5 variables and livability. The shapefile of New York has quite a few islands. However, after the data of New York and New Jersey are merged, no islands exist anymore as the areas around the islands are also filled. At last, the final datasets is exported as shapefile for further analysis in GeoDa in that based on former experiences, autocorrelation analysis in ArcGIS could be very slow and inefficient, sometimes even coming up with unknown errors.

Table 2. Data processing index

Indicator	Dataset name	Census table short name	Census table full name
Income per capita	X19_INCOME	B19025e1	AGGREGATE HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2015 INFLATION-ADJUSTED DOLLARS): Total: Households -- (Estimate)
	X01_AGE_AND_SEX	B01003e1	TOTAL POPULATION: Total: Total population -- (Estimate)
Time taken to travel to work / min	X08_COMMUTING	B08134e1	MEANS OF TRANSPORTATION TO WORK BY TRAVEL TIME TO WORK: Total: Workers 16 years and over who did not work at home -- (Estimate)
Unemployment rate	X23_EMPLOYMENT_STATUS	B23025e5	EMPLOYMENT STATUS FOR THE POPULATION 16 YEARS AND OVER: In labor force: Civilian labor force: Unemployed: Population 16 years and over -- (Estimate)
	X01_AGE_AND_SEX	B01003e1	TOTAL POPULATION: Total: Total population -- (Estimate)
Related occupation rate	X24_INDUSTRY_OCCUPATION	C24010e3	SEX BY OCCUPATION FOR THE CIVILIAN EMPLOYED POPULATION 16 YEARS AND OVER: Male: Management, business, science, and arts occupations: Civilian employed population 16 years and over -- (Estimate)
	X24_INDUSTRY_OCCUPATION	C24010e39	SEX BY OCCUPATION FOR THE CIVILIAN EMPLOYED POPULATION 16 YEARS AND OVER: Female: Management, business, science, and arts occupations: Civilian employed population 16 years and over -- (Estimate)
	X01_AGE_AND_SEX	B01003e1	TOTAL POPULATION: Total: Total population -- (Estimate)
Median age	X01_AGE_AND_SEX	B01002e1	MEDIAN AGE BY SEX: Total: Total population -- (Estimate)

3. Spatial autocorrelation test on each indicator

Firstly, instead of jumping into autocorrelation analysis on livability, the test on each indicator is done at first in order for future comparison. For both of the global and local analysis, in terms of weights, rook contiguity weights are used which means the weight for two polygons sharing one edge would be set as 1. A global Moran's I is used to check if a factor is significantly autocorrelated in an area and if yes, a local Moran's I could be used to cluster. Except for the data of age, the other four data's distributions are all skewed right (see Appendix 3.), meaning there are extremely large values in our data. Because Moran's I are strongly affected by extreme values, we decided to utilize the Geary's c to double check our results.

As the result shows, income and occupation both are presenting strongly positive spatial autocorrelation. Commute, unemployment, and age are presenting median positive spatial autocorrelation.

3.1 Global test

Table 3. Global test result

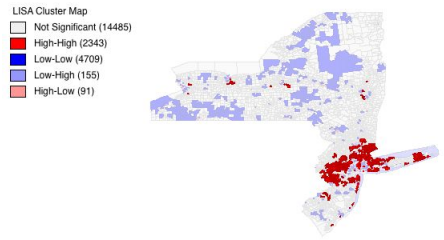
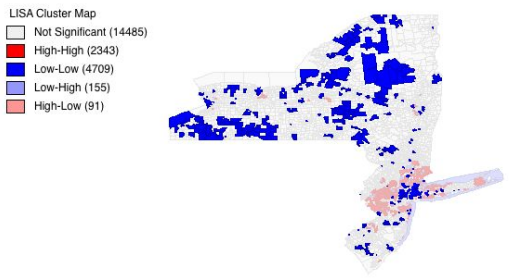
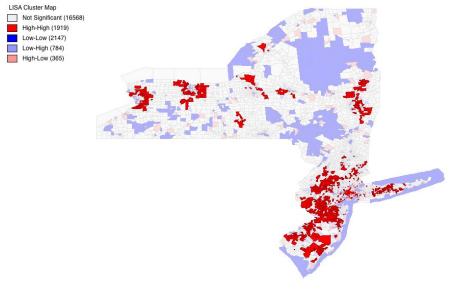
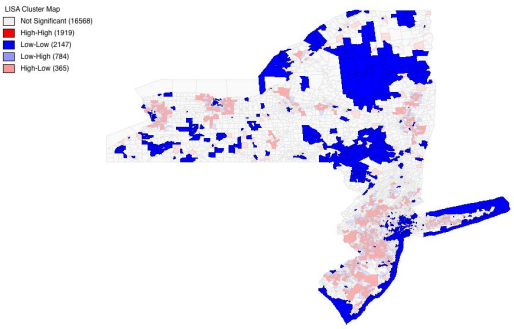
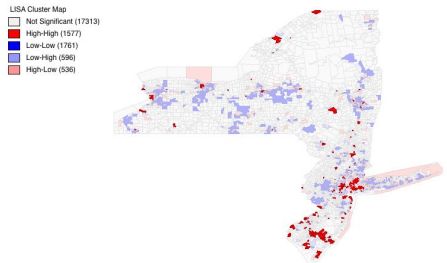
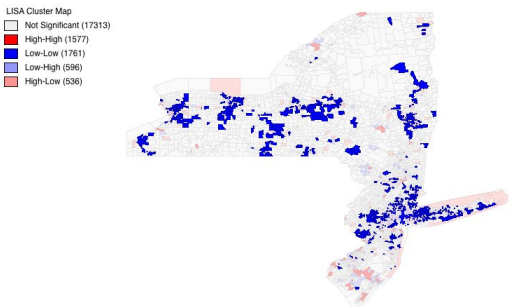
Indicator	Moran's I (ArcGIS)/(Pysal 'B')	Moran's I (GeoDa)/(Pysal 'r')	Geary's c (Pysal 'B')
Income	0.661788	0.698335	0.306055
Commute	0.280576	0.273462	0.821954
Unemployment	0.205453	0.211429	0.780983
Occupation	0.566248	0.601823	0.401723
Age	0.332803	0.346671	0.682077

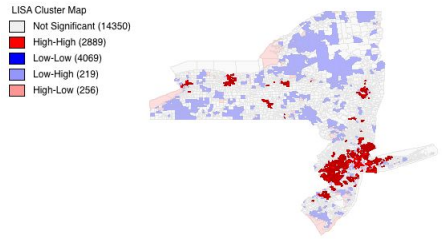
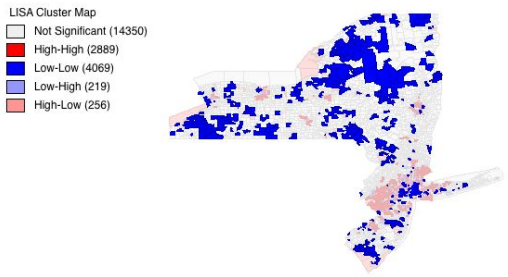
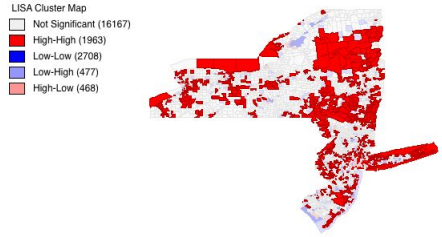
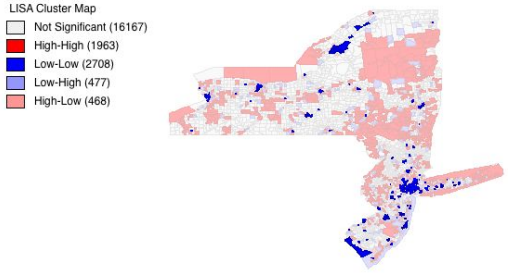
Note: The result differences between ArcGIS and GeoDa come from the default methods of weights transformation. ArcGIS' default is 'binary', while 'row-standardized' in GeoDa.

3.2 Local test

Based on the local Moran's I test result, significant HH and LL clusters are often located around cities (see Appendix 5.). As for income and occupation, large proportions of HH locate at areas around NYC. As for the commute, LL clusters are observed in cities' downtown areas and the wilderness (e.g. West Canada Lake Wilderness, Big Indian Wilderness). As for age, HH clusters are loosely located in the whole space, while LL mainly around cities.

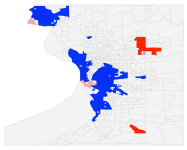
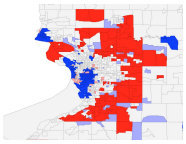
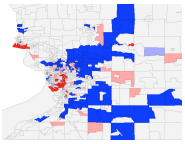
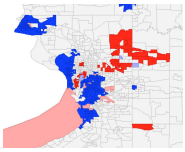
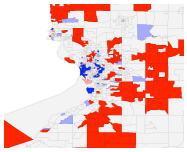
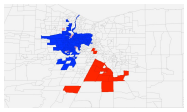
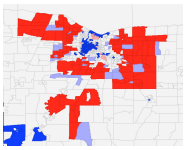
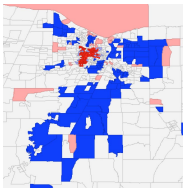
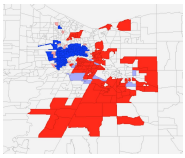
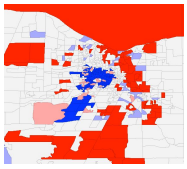
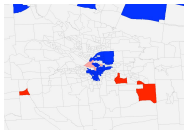
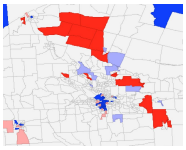
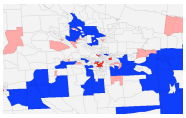
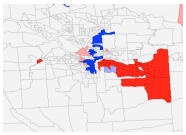
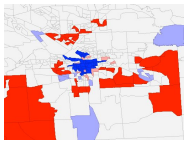
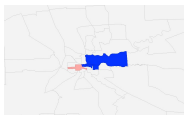
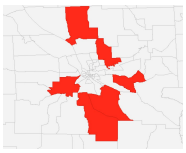
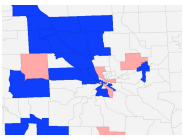
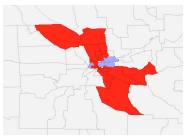
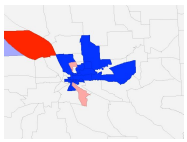
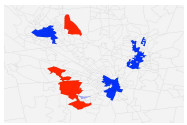
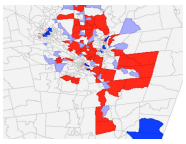
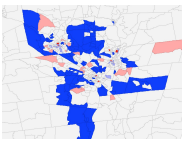
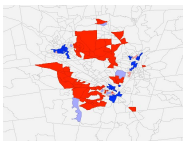
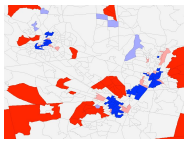
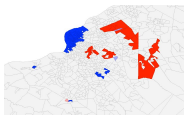
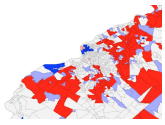
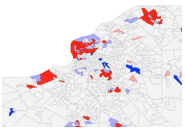
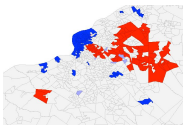
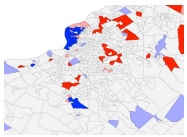
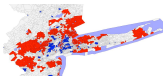
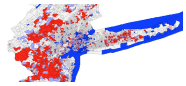
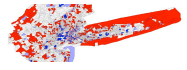
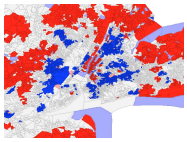
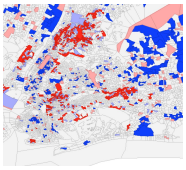
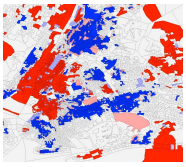
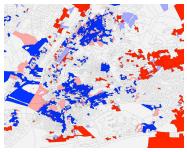
Table 4. HH and LL clusters distribution

Indicator	HH	LL
Income		
Commute		
Unemployment		

Occupation		
Age		

Some common clustering areas of those five indicators we observed are Buffalo, Rochester, Syracuse, Ithaca, Albany, Philadelphia, and New York area. In terms of income, generally, the LL are in downtown while HH in rural areas. One exception is NYC, with HH in the center of the city. Commute follows another pattern. With mainly LL in the center, HH and LH scatter around the center. The HH of the occupation related to management, business, science, and arts are highly concentrated around NYC.

Table 5. Patterns around major cities in NY and NJ

Urban area	Income	Commute	Unemployment	Occupation	Age
Buffalo					
Rochester					
Syracuse					
Ithaca					
Albany					
Philadelphia					
New York area					
NYC					

4. Spatial autocorrelation test on livability

4.1 Global Moran's I

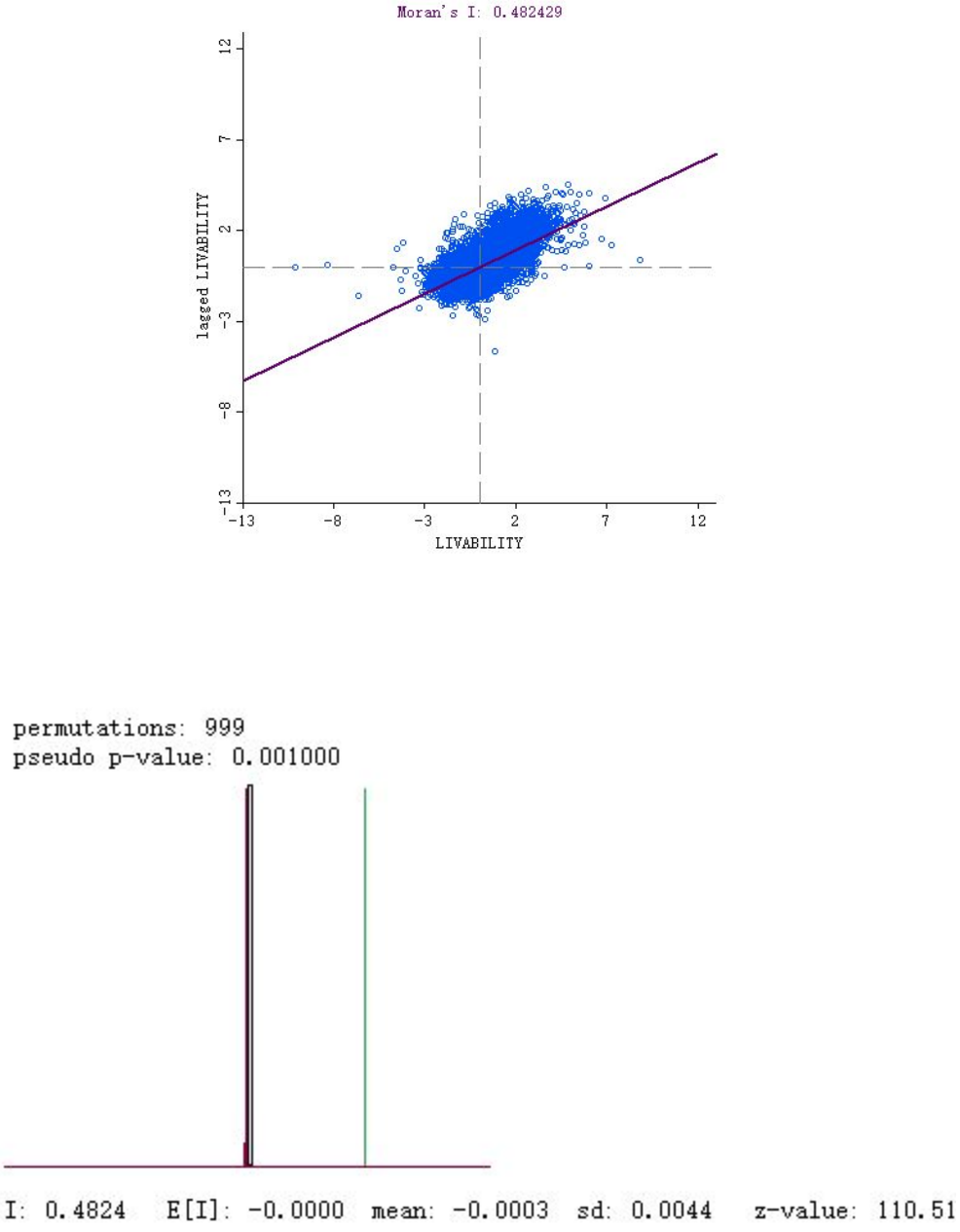


Figure 1. Global Moran's I on livability

From the result of Figure 1, it shows that livability is significantly autocorrelated as the z-value is bigger than the critical value 1.96 and the p-value < 0.01. Additionally, the Moran's I is 0.4824 which indicates that livability is highly autocorrelated in NJ and NY. Accordingly, a local Moran's I could be done to see the specific clustering.

4.2 Local Moran's I

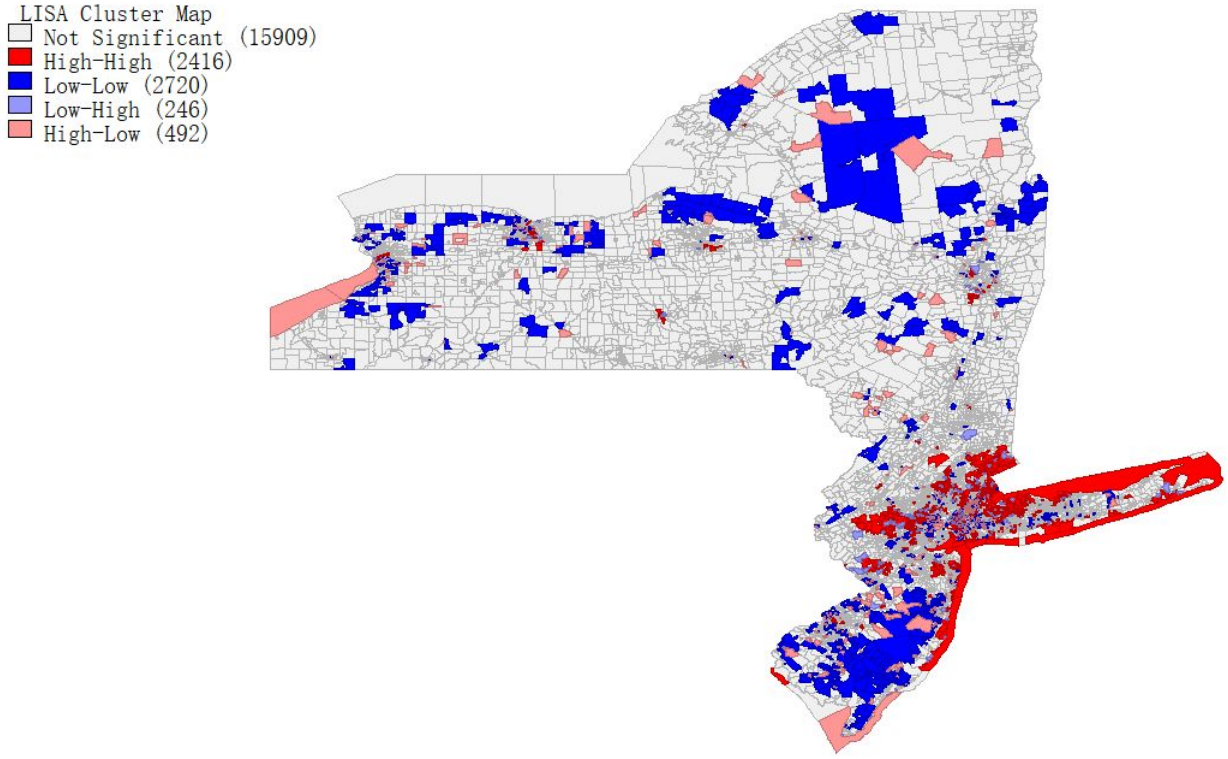


Figure2. Local Moran's I on livability

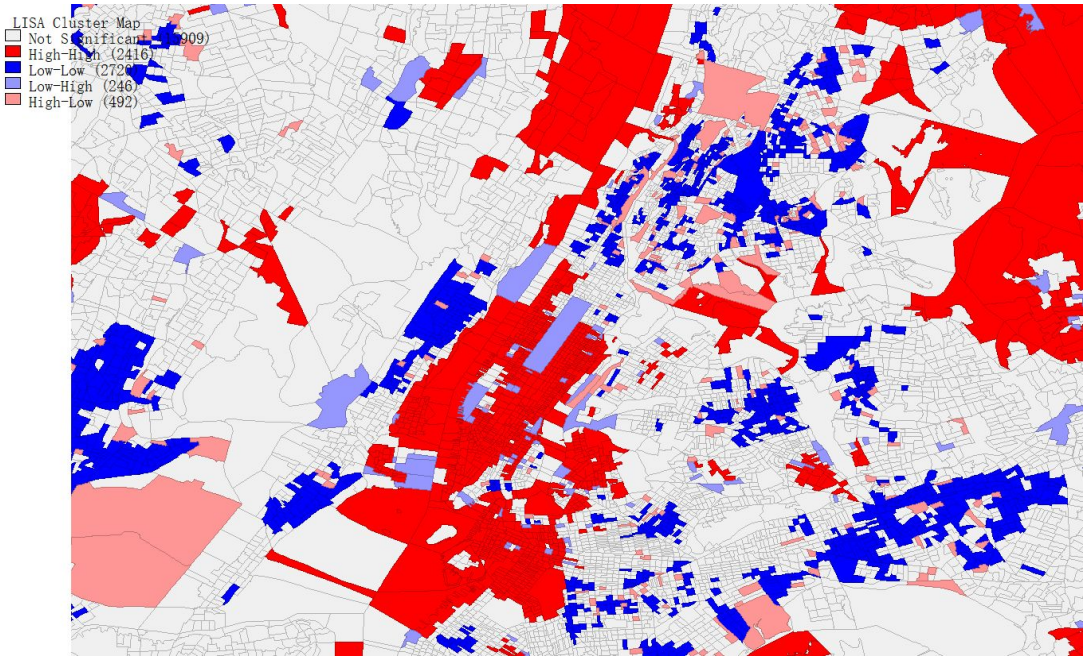


Figure 2 shows the result of a cluster map of livability based on a local Moran's I. It could be seen that high livability areas tend to be located in New York and especially across the coastline of Long Island and long beach island and around lower Manhattan. Low ones tend to be scattered in New Jersey.

5. Conclusion

Based on the livability results above, when we try to find a new place to live which may be beneficial to our career, New York City is still the best choice for us. After graduation, CUSP students could choose areas around lower Manhattan, east Newport, downtown Brooklyn, and along the coastline of Long Island. However, this conclusion is built on an important hypothesis that we new graduates are not sensitive to rent (in fact most of us are highly sensitive to it). If we further consider fair-price housing, areas in NYC which are currently not marked as HH might be our potential choices as well.

In addition, as these factors and livability all show a significant autocorrelation in these areas, in the future when analyzing them using regression in urban science, the autocorrelation effect should be accounted for or eliminated in case the autocorrelation of our variable would have a huge impact on the models.

6. Lesson we learn

6.1 How to assign the weights

In this study, weights are determined very naturally due to the fact that there is no related literature and the factors are treated as equally important. However, another way could have been conducted which is that firstly do a local Moran's I on each variable and then based on respective result such as HH, LL, LH and HL, a corresponding weight could be assigned to each result for each geo feature by the factor. In this way, the result might have been more meaningful.

6.2 Study design

We found that study design is crucial in order to process smoothly and get a meaningful result. In our study, a mighty important factor ---- housing price is ignored in our study design phase, which has caused a big problem when we tried to interpret our result. The neighborhoods good for career development we found are actually quite expensive in rent. But a rich neighborhood is obviously not livable for a majority of fresh graduates.

6.3 Data collection

When doing data collecting, carefully reading and getting familiar with the metadata file is necessary. Since the description of each column in the census metadata is quite concise. At first, we had a misunderstanding when finding the column ‘total population’, which led to mistakenly choosing the total household.

6.4 Methodology

It turns out if it comes to a big volume of shapefile, GeoDa could be way more effective than ArcGIS, which shows a good compatibility for shapefile analysis, though. Errors have emerged for quite a few times when local Moran’s I was done in this study using ArcGIS.

References:

1. Branch, G. (2017). TIGER/Line® with Selected Demographic and Economic Data - Geography - U.S. Census Bureau. [online] Census.gov. Available at: https://www.census.gov/geo/maps-data/data/tiger-data.html [Accessed 26 Nov. 2017].
2. Branch, G. (2017). 2010 Geographic Terms and Concepts - Block Groups - Geography - U.S. Census Bureau. [online] Census.gov. Available at: https://www.census.gov/geo/reference/gtc/gtc_bg.html [Accessed 26 Nov. 2017].
3. Wang, J., Su, M., Chen, B., Chen, S., & Liang, C. (2011). A comparative study of Beijing and three global cities: A perspective on urban livability. Frontiers of Earth Science, 5(3), 323-329.

Appendix

1. Team roles

Data collecting and columns truncating: New Jersey (Zhiao Zhou), New York (Lingyi Zhang)

Data cleaning and merging: Zhiao Zhou

Study design: group discussion

Autocorrelation test: ArcGIS and GeoDa (equal contribution), PySAL (Lingyi Zhang)

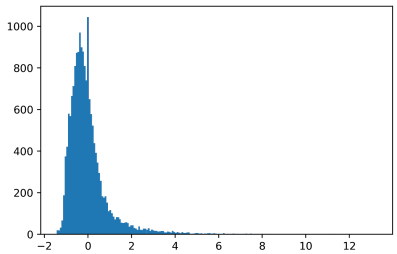
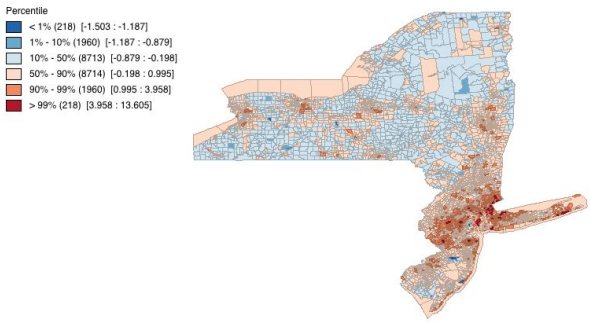
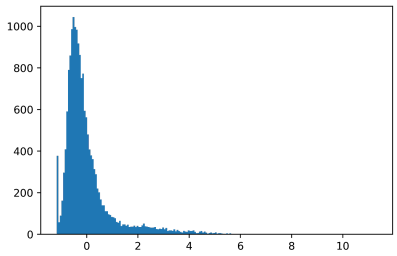
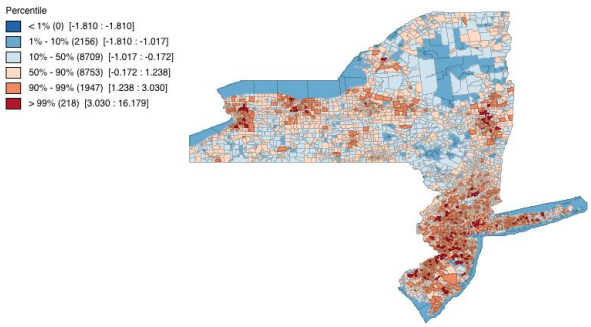
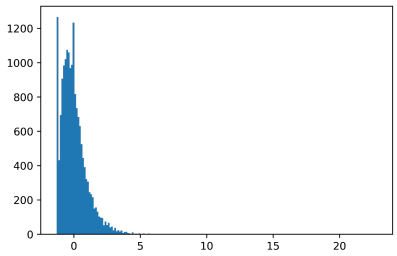
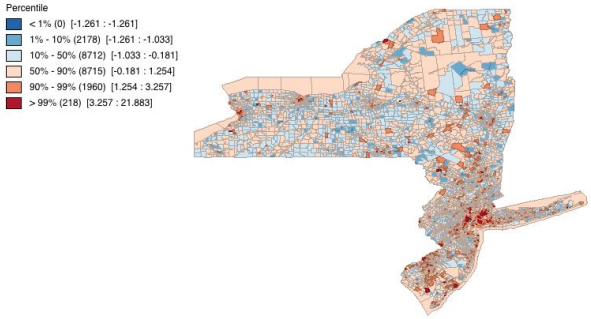
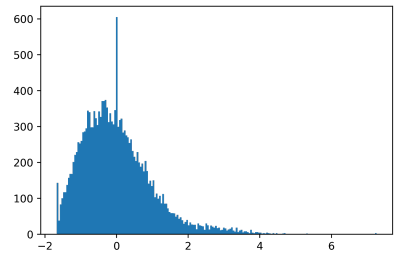
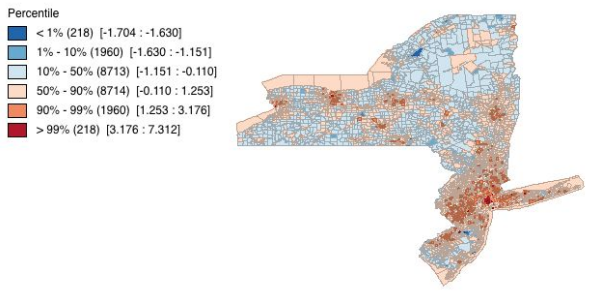
Report writing: equal contribution

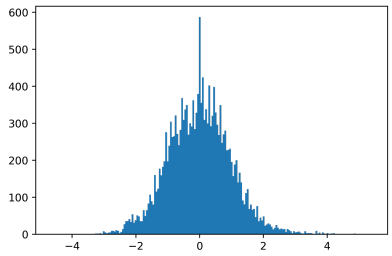
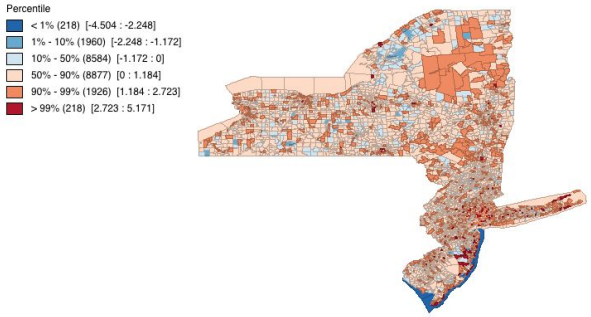
2. Data source:

[U.S. Census Bureau American Community Survey (ACS)]

https://www.census.gov/geo/maps-data/data/tiger-data.html

3. Data descriptive

Indicator	Data distribution	Value percentile
Income		
Commute		
Unemployment		
Occupation		

Age		

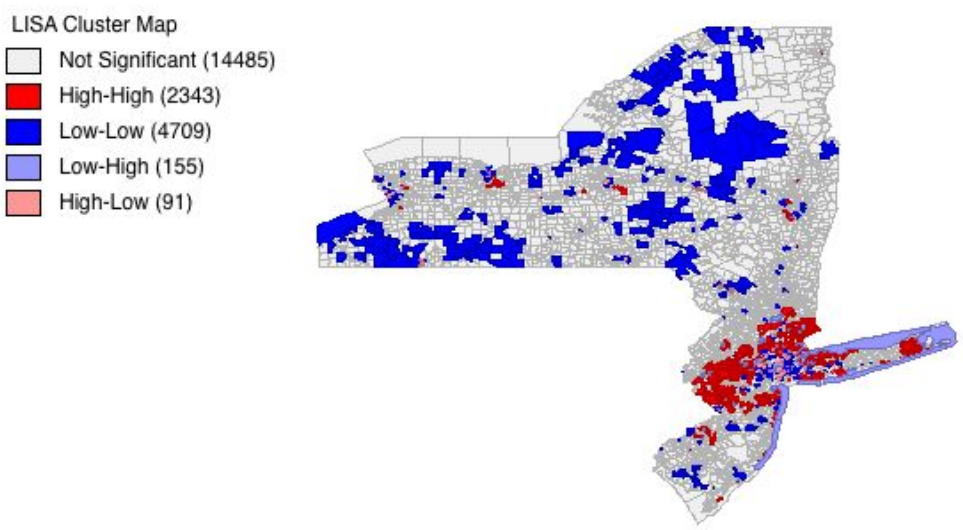
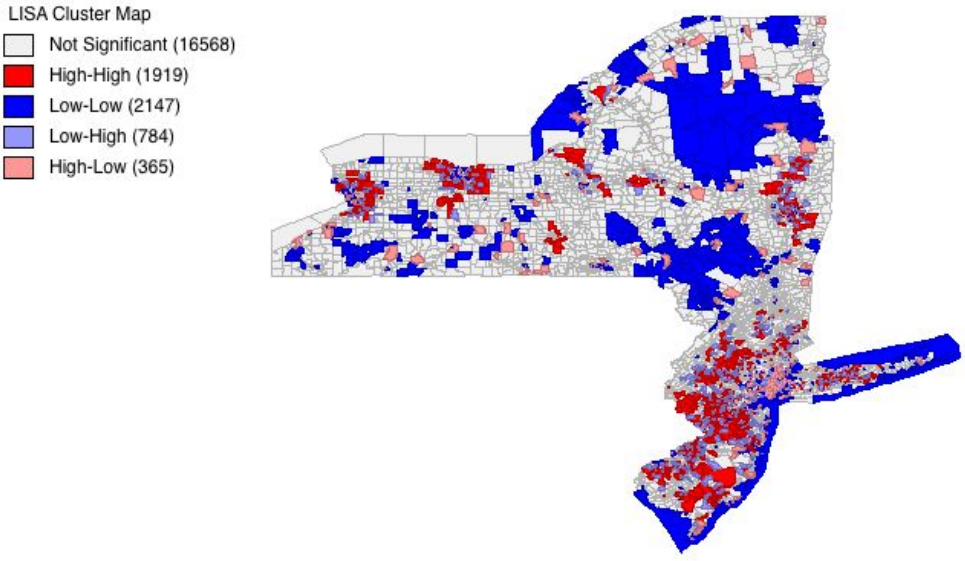
4. Global Moran's I result (ArcGIS)

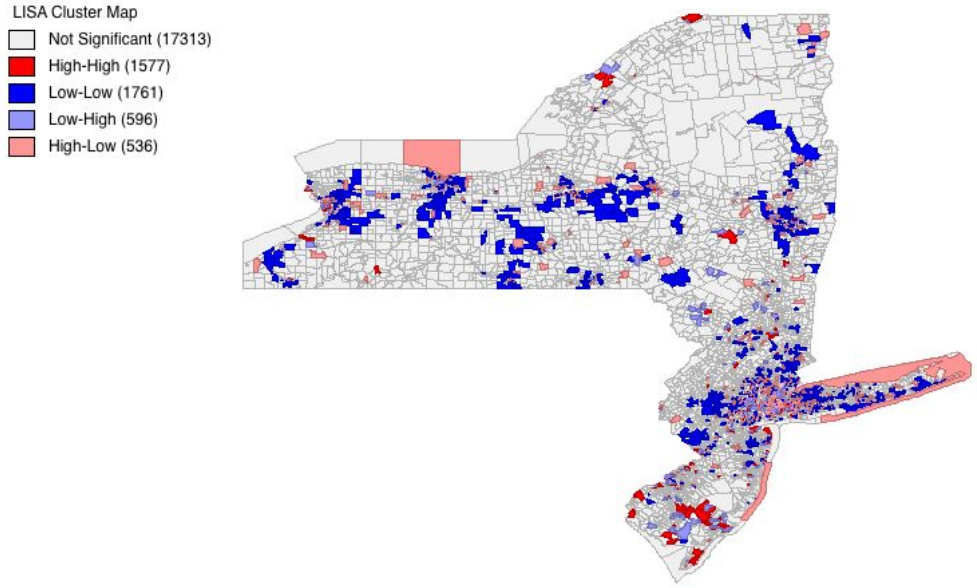
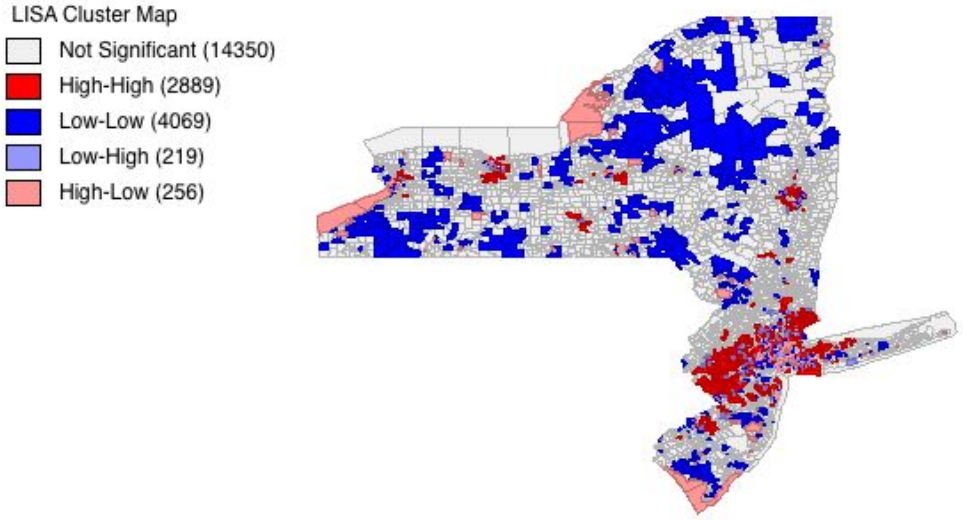
Table 6. Result of global Moran's I (assumption: normality)

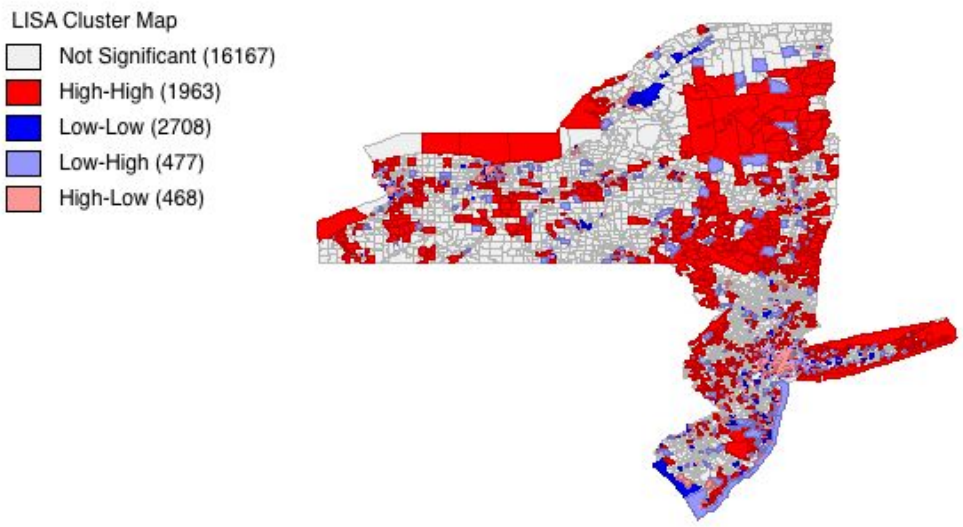
Indicator	Moran's Index	expected	Std dev	z-score	p-value
Income	0.661788	-0.000046	0.000017	161.406211	0.000000***
Commute	0.280576	-0.000046	0.000017	68.425656	0.000000***
Unemployment	0.205453	-0.000046	0.000017	50.118069	0.000000***
Occupation	0.566248	-0.000046	0.000017	138.052929	0.000000***
Age	0.332803	-0.000046	0.000017	81.141757	0.000000***

Note: *** 0.01, **0.05, *0.1

5. Local Moran's I result (GeoDa)

Indicator	Local Moran's I
Income	
Commute	

Unemployment	
Occupation	

Age	 LISA Cluster Map Not Significant (16167) High-High (1963) Low-Low (2708) Low-High (477) High-Low (468)