Analysis Of Township Comprehensive Development In Connecticut By Clustering And Spatial Map

by

M.S., Haoxi Ma

University of Connecticut, 2020

StudentID:2833925

ABSTRACT

Township distribution and development show varieties from different states in the US, which is important for the government to formulate policies wisely and appropriately. In this project, firstly, we will focus on the pattern of human settlements' distribution in Connecticut using spatial maps. Then, when doing analysis of township comprehensive development, towns will be classified into three levels by clustering based on six main factors: Total crime number, Per capita income, Median household income, Median family income, Population and Number of households.

Conclusion shows that, in Connecticut, human settlements are distributed as polycentric settlements with the city as the center. And almost all cities belong to the third level while most towns are second level. Besides, there is a "Gold Coast" (*Wikipedia, The Free Encyclopedia*) in the left bottom of Connecticut composed of several first level towns.

INTRODUCTION

1. Description of data

The data provided by crdata.org containing the crime index for individual towns in CT from 2010 to 2017 can be downloaded from http://data.ctdata.org/dataset/ucr-crime-index.

This crime dataset has 29920 rows and 7 columns corresponding to 7 variables:

Table 1: Variable summary for crime

Variables	Description
Town	The name of the towns
FIPS	Federal Information Processing Standards
Year	From 2010 to 2017
Crime.Type	Crime index like Total crime, Total Violent Crime and so on
Measure.Type	Number or Rate (per 100,000)
Variable	One level -- "Crime index"
Value	Value corresponding to the Crime.Type and Measure.Type

Besides, We retrieve a table named income from:

https://en.wikipedia.org/wiki/List_of_Connecticut_locations_by_per_capita_income

which contains data from the 2010 United States Census and the 2011-2015 American Community Survey 5-Year Estimates.

This table has 179 rows and 8 columns corresponding to 8 variables:

Table 2: Variable summary for income

Variables	Description
Town	The name of the towns
Type	3 levels factor: Town, Borough or City
County	Administrative region of a country
Per capita income	Average income earned per person
Median household income	Median income of household
Median family income	Median income of family
Population	Number of residents
Number of households	Number of households

2. Related works

In order to construct a spatial map, finding longitude and latitude for each Town are necessary. Therefore, we need to register a Google API key and activate it in R.

What's more, because we are not familiar with Connecticut towns development, asking a local American friend and looking for some articles to acknowledge related information is a good choice, which can help combine our analysis with reality.

RESULTS AND CONCLUSIONS

We use *qmap* function to construct a density plot for township distribution and then locate all cities:

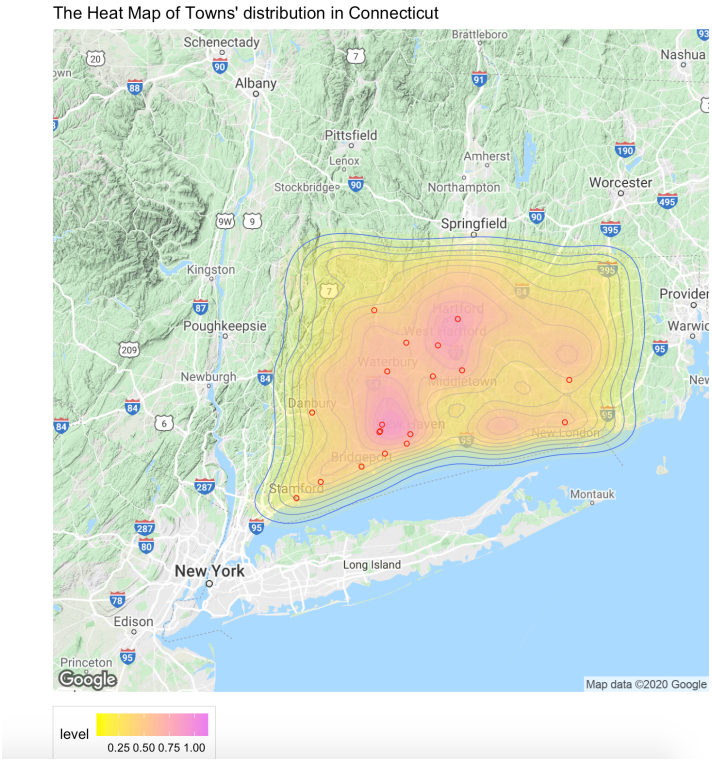


Figure 1: Density plot for township distribution

In Figure 1, it seems that almost all of the cities lie in or near the region with high density. Therefore, we can conclude that human settlements distribute as polycentric settlements with the city as the center, which accord with Central Place Theory (Walter Christaller, 1933).

Then we do 3-group clustering based on the six features mentioned above and show the result in a graphic way (shown in Appendix-F-1). First of all, we make a table for three cluster versus settlements' type (Code result is shown in Appendix-F-2):

Table 3: Cluster vs Settlements' type

	Cluster 1	**Cluster 2**	**Cluster 3**
Borough	0	3	0
City	0	3	16
Town	15	119	11

According to Table 3, we find that almost all of cities belong to cluster 3. In order to

identify the property of different clusters, we should summarize those six features grouped by different cluster (Code result is shown in Appendix-F-3):

Table 4: Six features within each cluster

Cluster	Crime	*PCI**	*MHI**	*MFI**	Population	*NOH**
1	943	*$76544*	*$151080*	*$190646*	23888	8489
2	1300	*$38158*	*$79826*	*$92862*	11307	4332
3	3109	*$30244*	*$59046*	*$72325*	65600	25391

Note: the table shows the mean value of each variable within each cluster

**PCI: Per capita income; MHI: Median household income; MFI: Median family income*

**NOH: Number of households*

According to Table 4, we know that, contrary to cluster 3, cluster 1 corresponds to the lowest crime rate and highest per/median income. And cluster 2 is in the middle. Therefore, we can label the towns in cluster 1 as the first level, which means great comprehensive development. Label criteria is the same for cluster 2 and 3.

Then we plot these three types towns in a spatial map:

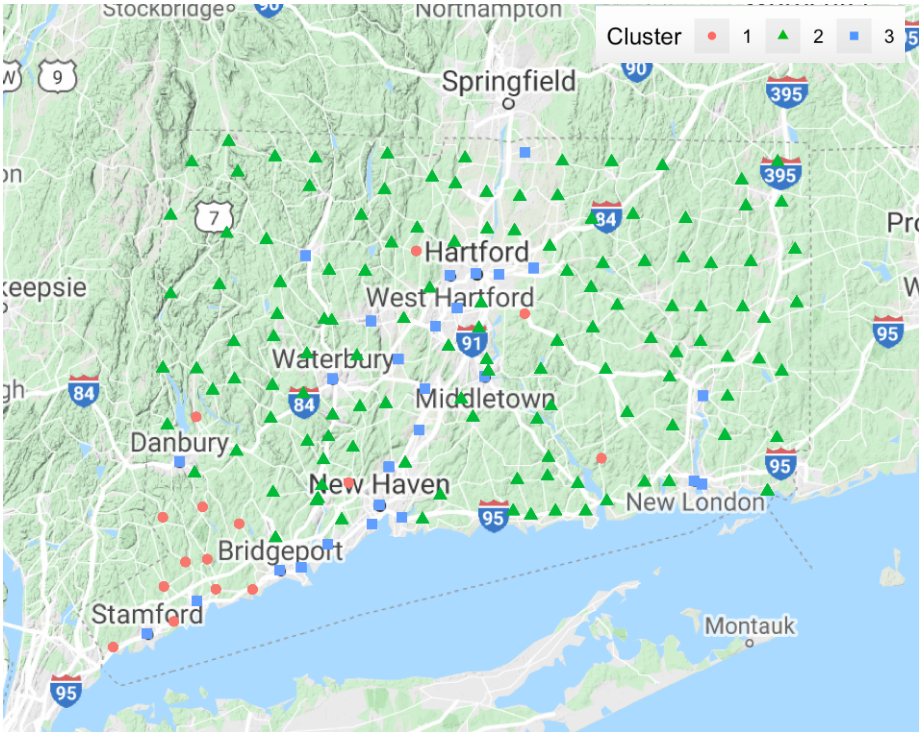


Figure 2: Township distribution classified cluster

From the Figure 2, we have some conclusions:

(i) Blue Square represents the third level towns corresponding to highest crime rate and lowest per/median income. We can see that almost all of the cities belong to this level.

Maybe the reason for this is that cities have large populations and gaps between rich and poor. Although the total GDP of a city may be high, GDP per capita can be low.

(ii) Green rectangular represents the second level towns. This kind of town makes up a large proportion and is widely distributed.

(iii) Red point represents the first level towns, corresponding to highest per/median income and lowest crime rate. The towns in this level are rare. It's worth mentioning that almost all first-level towns are concentrated in the left bottom in Connecticut. This region is called “Gold Coast”, home to many wealthy Manhattan business executives.

METHOD

1. Retrieve a table from website

First, we download a HTML file from the URL. Then, use $htmlParse$ function to parse the HTML file and $getNodeSet$ function to retrieve all tables. Finally, find the third table containing income information by $readHTMLTable$ function. All functions mentioned above are set in XML package.

2. Refine income table

We delete some useless columns and rows. Then, change the columns' name and rows' name to get an appropriate income data frame.

3. Geocode to get Lat and Lon

In this part, we focus on the crime dataset. Because there are lots of crime index corresponding to a single town, we use pipe function to link two steps $filter$ function with str_detect function to retain total crime rate for each town. After that, in order to reduce the API usage, we delete duplicates in the Town variable. Then, after using $Paste$ function to transform "Andover" into "Andover, CT, US", we retrieve longitude and latitude by $geocode$ function. At last, we obtain a new dataset, *crime.lat*, by $merge$ function. All functions mentioned above are set in $dplyr, stringr\ and\ ggmap$ packages.

4. Merge income and crime.lat dataset

We merge income data frame and crime.lat data frame by Town variable generating a new data frame called *crime.income*. Notice that there are redundancy levels in County and Type variables and we should delete them. At last, check data frame's structure and missing value by $md.pattern$ function in $mice$ package.

5. Draw density plot

Now, after getting map of Connecticut by $qmap$ function, we can draw a density plot for human settlements distribution by $geom\_density2d$ function. Then, select all city-type regions and locate them in the density plot by $geom\_point$ function. All functions mentioned above are set in $ggplot2$ package.

6. Summarize clustering data frame

In this part, we need to summarize a data frame for clustering. First, we transform all income form "$66,862" to 66862 by $gsub\ and\ as.numeric$ functions. Besides, because the information in the income table is collected by 2010 United States Census and 2011-

2015 American Community Survey 5-Year Estimates, we use the mean of total crimes from 2010-2015 as Total crime number for each town. We calculate this by $group\_by\ and\ mutate$ functions. Then, scale all numeric data preparing for clustering by $scale$ function. Now, we get the *data.cluster* data frame.

7. Do clustering

Here, we choose to use $k-means$ method to do clustering on *data.cluster* data frame. We are not going to introduce $k-means$ method and you can check it at https://www.sciencedirect.com/science/article/pii/S0031320302000602?casa_token=vtC-qpU1BdEAAAAA:dV8UPV3oX_Ha7I3j5igZ4KsawJLeGndEbosyVUXbIT0e9Z4zRgw3H8fNJrjh3EHcRGW4MuDmtG4.

The function in R doing $k-means$ clustering is $kmeans$. And we can show the result graphically by $fviz\_cluster$ function. These two functions are all set in $factoextra$ package.

8. Retrieve group index

First, we construct a new vector to retrieve clustering index in Step 7. Then, merge it into *crime.income* data frame. What's more, we make a table for six factors versus 3 cluster by $group\_by\ and\ summarize$ functions linked by pipe function.

9. Plot for classified towns

Finally, we can draw a scatter plot of classified towns in Connecticut.

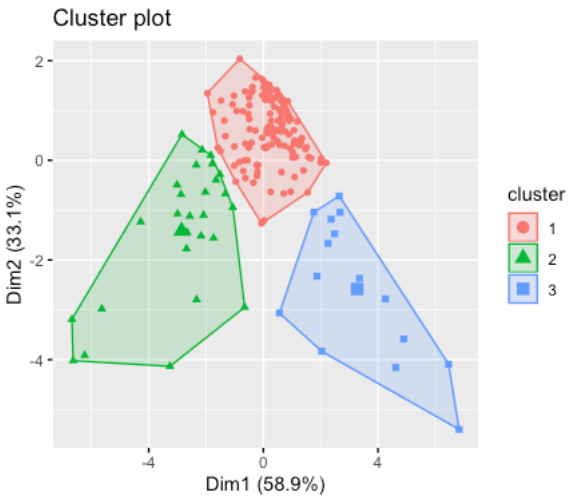
Reference

Wikipedia contributors, Gold Coast (Connecticut). 8 February 2020 05:16 UTC. https://en.wikipedia.org/w/index.php?title=Gold_Coast_(Connecticut)&oldid=939707930

Wikipedia contributors, Central place theory. 20 April 2020 11:30 UTC. https://en.wikipedia.org/w/index.php?title=Central_place_theory&oldid=952064612

Appendix

F-1:



F-2:

	1	2	3
Borough	0	3	0
City	0	3	16
Town	15	119	11

F-3:

```
# A tibble: 3 x 7
  Cluster mean.crime mean.Per.income mean.M.H. mean.M.F mean.Popu mean.N.H
  <fct>        <dbl>           <dbl>     <dbl>    <dbl>     <dbl>    <dbl>
1 1             943.          76544.   151080.  190646.    23888.    8489.
2 2            1300.          38158.    79826.   92862.    11307.    4332.
3 3            3109.          30244.    59046.   72325.    65600.   25391.
```