Watkins, Kari E <kari.watkins@ce.gatech.edu>

Fri 11/8/2019 4:24 PM

The text below gives some explanation of how things changed as clusters increased. The table you need is attached though – columns for 11 clusters and 5 clusters plus my weird 8 groups (not actually how they clustered at 8).

Let me know if you want to chat about it.

Best,

Kari

Hi all,

I was in vacation last week so am jumping a bit late on this.

Although most of the panel doesn’t see it, I think that clusters have the advantage of comparing metros that have structural similarities beyond just their sizes. Also, on a personal note, it would be really cool if the clustering mechanism we just published on could be used for the TCRP study. At the same time, when Greg starts fitting the data, it will probably become apparent which clusters make the most the sense.

I like Kari’s idea of 11 clusters to keep metros of different sizes separate while differentiating between density profiles and transit network sizes. I would put (9 and 7) together and (6 and 8) together instead of Kari’s (6) and (7, 8, and 9). In my mind, Clusters 7 and 9 are mostly sprawling sunbelt cities, whereas 6 and 8 are mostly legacy rust-belt systems. But I think at some point it gets subjective.

Note: I am referring to cluster numbers as shown on the PDF, not on the excel spreadsheet. I believe that Kari was too.

Cordially,

—  
Dr. Simon J Berrebi  
404.663.0341

[http://berrebi.net](https://nam04.safelinks.protection.outlook.com/?url=http%3A%2F%2Fberrebi.net&data=02%7C01%7Cgreg.erhardt%40uky.edu%7C47b524eb9dce4a1b20c508d7025cd704%7C2b30530b69b64457b818481cb53d42ae%7C0%7C0%7C636980470100798262&sdata=J9Onkm10uu1sDXU1wDyeuwEpt4g7z7BdPJ6KgPmULWU%3D&reserved=0)

**From:** Ederer, David J <davidederer@gatech.edu>   
**Sent:** Tuesday, July 2, 2019 12:27 PM  
**To:** Watkins, Kari E <kari.watkins@ce.gatech.edu>  
**Subject:** Updated cluster solutions with rail dummy

Hi Kari,

I ran some cluster analyses using presence of rail as a dummy variable. I think it's a good improvement as it helps different the bus systems more effectively. Clusters with a rail system are all group together. Without this, there are many similarities between systems that might not be informative (e.g. population, density, and zero vehicle households could create similar groups with relatively little weight for system characteristics. Adding the rail dummy basically differentiates the system).

One thing that is not surprising is the LA is singular. If I remember correctly, we cut the clusters off at 5 so that LA would have peers. In this iteration, I think we should leave it as it's own group. LA is just too big and too dense to compare in a reasonable manner.

I created the table below to show you how things break down as we change the number of clusters. I think it is a little easier to see how clustering changes as we change the number of groups in the table below. Understanding what happens when we change the number of clusters depends on the method used. In prior iterations, we used an agglomerative method that basically grabs the most similar metro(s) and builds a cluster based on similarity. In other words, everyone starts in their own cluster and grabs the nearest neighbor to build up. Thus, when increasing the number of clusters groups will split off from one cluster and (most likely) go into another cluster. Metros will not split evenly into different groups, they all go together based on similarity. Does that make sense?

In the table below, you can see that process at work. As we increase the number of clusters, groups are chipped away and appended to other clusters. As we go from 5 to 10 clusters, cluster 1 remains pretty stable. As we increase above 5 clusters, LA is alone. No matter the number of clusters, it will remain by itself. I try to explain each iteration below. The typical mechanical tests for determining clusters (scree plots, gap stats) aren't all that helpful here.

**5 to 6 clusters**

This is the cut point for LA. They "leave" the other large dense metros and form cluster 6. In each iteration, they are the highest numbered cluster.

**6 to 7 clusters**

5 metros break off from cluster 4 (B'more, Buffalo, Cleveland, Pittsburgh, San Juan) and form their own cluster. In terms of interpretation, this is a good addition. These are legacy systems that are different from the newer systems in cluster 4 (sprawling areas like Atlanta, Sacramento, Charlotte).

**7 to 8 clusters**

Here, the cluster breaks off cluster 3, which contains a wide range of cities (e.g. Nashville, Charleston, Orlando, Birmingham, Greenville)**.** This forms cluster 6, which feature some of the smaller metros from cluster 3 that  spend more money on transit operations than one would expect (Kalamazoo, Carbondale)**.**

**8 to 9 clusters**

Cluster 2 loses 10 metros, which form the new cluster 8. Cluster 2 originally had smaller cities (Santa Rosa, Oxnard, Stockton). Cluster 8 consists of the 10 big cities in that group (Detroit, San Antonio, Tampa).

**9 to 10 clusters**

This takes 11 metros from cluster 4 to create another cluster (again the 8th cluster is new). The new cluster 8 consists of large newer cities with rail, and relatively recent growth (Denver, Charlotte, Portland, Phoenix). The three reamining areas in cluster 4 are Atlanta, Houston, and Dallas (where sprawl is king!).

**10 to 11 clusters**

This iteration breaks down cluster 3 by density.The 28 most dense metros remain (about 15% of tracts are considered "dense"), while the 60 less dense areas form a new cluster.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **# Clusters** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| **Five** | 41 | 47 | 139 | 19 | 8 | - | - | - | - | - | - |
| **Six** | 41 | 47 | 139 | 19 | 7 | 1 | - | - | - | - | - |
| **Seven** | 41 | 47 | 139 | 14 | 5 | 7 | 1 | - | - | - | - |
| **Eight** | 41 | 47 | 88 | 14 | 5 | 51 | 7 | 1 | - | - | - |
| **Nine** | 41 | 37 | 88 | 14 | 5 | 51 | 7 | 10 | 1 | - | - |
| **Ten** | 41 | 37 | 88 | 3 | 5 | 51 | 7 | 11 | 10 | 1 | - |
| **Eleven** | 41 | 37 | 28 | 60 | 3 | 5 | 51 | 7 | 11 | 10 | 1 |

I've attached a copy of an excel sheet with 11 clusters should you like to take a look. I've also included a few pdfs of the dendrograms. The font is small so you'll need to zoom in. I know that you mentioned 7 as a possible number of clusters. Frankly I wouldn't do fewer than 8. That means some of the clusters are too small for any sort of conclusions based on regressions, but I think they begin to make more sense once we cross that threshold.

Best,

Dave

**David J. Ederer**

Ph.D. Candidate | Urban Transportation Information Lab

Transportation Systems Engineering  
Georgia Institute of Technology  
716-348-9944