Urban Clusters

When you see a picture of a city, does instinct immediately bring you to guess where it was taken? It's an urge I can't quite suppress. I find that even if I can't recognize the city, I can almost always still guess the continent. Even without signpost textual giveaways or inhabitants' facial features, many subtle urban features can clue you in. The roof architecture, street food, make of the buildings, road paving - these all help distinguish the continental origins of a city. Though there is incredible urban diversity between countries within continents, it still seemed to me that cities of one continent had more in common with each other than they did with cities of other continents.

I was staring at an aerial photo of Dhaka when I decided I wanted to numerically prove this hypothesis. Surely there could be some city metrics that featured more variation within continents than between them - the old ANOVA test from classical statistics. And so I set about gathering as much data as I could.

The headaches started immediately. This project almost certainly couldn't have been possible at any scale 5 years ago. It will certainly get easier in 5 years. There aren't really any comprehensive worldwide cities database. There are no standards of definitions and though efforts are being done to correct this (including a proposal for ISO for cities), nothing has been widely adopted. Even basic definitions of area and population are frustratingly inconsistent, with significant discrepancies of where to draw borders. Here though discrepancies are also continental in nature - American cities go by strict legal districting, whereas Asian cities often redefine their borders to match their urban sprawl. I had begun with dreams of finding creative metrics such as the average sidewalk width or the % of restaurants open after midnight, but soon realized I'd have to settle for what I could find.

Acquiescing that this would not be an exact science, I began with using a base dataset from the World Cities Cultural Forum (WCCF), who collected such interesting metrics as Art Exhibits daily visits and Rare & Secondhand Bookshops for 25 core cities. Their data was not without proven flaws (there was clearly a lack of consistent methodology) resulting in some questionable figures (Berlin has 4 rare bookshops and Johannesburg has 943?). I would fact check strange results and often manually make changes after vetting through data collection methods. It might seem like modern society is swimming in big data, but estimates for international tourists in Hong Kong ranged from 60 million to 27 million because the government doesn't have a method for identifying what a tourist is.

Ultimately I created a dataset of 42 cities (10 from Asia, 19 from Europe, 6 from North America, 7 from elsewhere) with 24 metrics. These included Number of Concert Halls and Median Weekly Earnings from the WCCF supplemented with data I could find on metro systems (length of rail, annual ridership and % usage), number of Starbucks, CO_2 emissions, number of airport runways and the number of international firms (calculated by McKinsey). There was plenty of missing data, which I imputed with the metric mean. Then, on every possible permutation combining 6 metrics, I ran a K-means

algorithm. I then analyzed the resulting clusters and found the combination that best matched reality, putting over 70% of the cities together with other cities from the same continent. There were a few “best combinations” and the one I’ve chosen to display in this application is Foreign Born %, Number of Cinemas (per capita), Metro Usage, Number of Restaurants (per capita), Working Age Population (as a % of total population) and Number of International Service Firms (per capita). Some takeaways from this combination:

- American cities have by far the highest foreign born %, followed by European cities. Most cities in Asia, Africa and South America have close to 0 foreign born %. Singapore, being an exception, was actually clustered with the North American cities by the algorithm.
- European cities have a lot more cinema screens per capita than other cities.
- Asian cities have way more restaurants per capita (although this statistic is hard to measure)
- Asian cities also have a large % of working age population, with American cities at the other extreme. To be honest, this one doesn’t quite make sense. I do think you see more elderly working in Asia - and when it’s a little octogenarian pushing trash uphill, in heartbreaking public ways - but I don’t think that’s captured in the accounting methods here. More likely we have vastly different population denominators between methodologies.
- Predictably, most international service firms are European or American and thus cities from those continents have much higher firms per capita. This statistic is pretty biased but I think it might have some effect on how a city looks and feels, as a proxy for how many familiar logos one sees.
- It wasn’t clear to me what to do with the cities outside these 3 continents in my database, including 3 South American cities, Istanbul, Mumbai, Johannesburg and 2 Australian cities. I labeled them all as Other, but the algorithm clustered the Australian cities with the Europeans, which meets the eye test.

The app has an interactive map (built in Leaflet) with all the 42 cities plotted and colored by their cluster. The color legend labels each color by the continent most associated with the cluster - this results in cities with a label not matching their true continent. Rome is colored as an Asian city - that just means it is closer to the Asian cluster than any other, even though I am fully aware that Rome is in Europe. No geographic information is included in the clustering algorithm.

The app lets you click on a city to have its data popup. You can also see 6 tabs of on the left which show density plots for each metric, split by the clusters. The idea is that the density plots will look rather distinct for each cluster. An orange line then shows where the metric of that exact city fits into the density plot. For cities missing data for a given metric, no orange line is shown.

There are plenty of flaws in my methodology and data, but ones that can be improved over time with more and better data. I believe there are many metrics that can reveal interesting urban planning or sociological distinctions between the continents - essentially the data that I saw helped confirm my thesis to me. Understanding the underlying reasons behind these distinctions can drive interdisciplinary conversations.

Personally this was also an important project for me. It was the major impetus driving my data science training, giving me a goal to work towards that necessitated me learning about data merging, language encoding, data standardization, clustering/classification algorithms and web application development. I even talked about the project in my final interview with GE.

The project is hosted on the free shinyapps.io server at https://cal65.shinyapps.io/Cities/ . This minimally viable approach is slow and won’t work when my home laptop is turned off (!). For users of R, a better user experience is available by downloading the Shiny package and running **runGitHub("Cities", “cal65")** . All my (sloppy) code is up on Github, and I’m happy to collaborate with people to improve this project.