Neighbourhoods in Milwaukee and London

Toby Cadoux

27 April 2020

1 Introduction

1.1 Background

In this work I will be comparing and contrasting neighbourhoods in two cities. I come from London while my wife comes from Milwaukee. When we visit it would be nice to know which neighbourhoods within her city are similar to the one in which she was raised. It would also be nice to know which neighbourhoods in Milwaukee are the similar to the one in which we live in London. Finally, we are considering moving (within London) fairly soon and so a categorisation of localities would potential enable us to narrow down our search for a new home.

1.2 Approach

When thinking about neighbourhoods there are as many definitions of borders and extent as there are people who know the areas. I decided to use a different definition of 'neighbourhood' for each city.

There were 76 neighbourhoods in Milwaukee that were officially recognised by the city council in 1990. The most recent city council definitions I could find, from 2000, had 190 neighbourhoods listed[1][2]. Milwaukee is not a large city compared with London and this number of neighbourhoods seemed excessive. In the end I decided to use a list prepared by the University of Wisconsin – Milwaukee[3]. This lists 48 neighbourhoods and roughly matches the list on Wikipedia[4] but with its more historic focus it is likely to match my wife's and her family's notions more

[1] The Milwaukee DCD gives ArcGIS neighbourhood definitions.
[2] The city also has a map of neighbourhoods.
[3] Found here.
[4] Wikipedia has a List of Milwaukee neighbourhoods.

closely. Many of the newer neighbourhoods are a long way outside what most would consider 'Milwaukee proper'.

London is divided into 32 boroughs and the City of London (the City of Westminster is also a borough). The boroughs are diverse and consequently, especially considering London's size, I decided not to use these division to represent neighbourhoods. Instead I used the much smaller postcode areas (SW19, E1 etc.) This is a fairly common way that Londoners reference individual neighbourhoods, moreover it conveniently separates central and inner London from he outer suburbs. There are 119 London postcode areas[5], which all have higher populations than any of the neighbourhoods in Milwaukee.

To measure similarity I decided attempt to use two main metrics:

- The first is the mix of consumer businesses, amenities and leisure venues as compiled by Foursquare. The Foursquare information is present in both London and Milwaukee.
- The second is to look at average house prices for each neighbourhood (relative to the overall house price average). This will be simple enough in London — Land Registry records in London all have a postcode attached. Real estate sales in Milwaukee are listed by address so, for this to work, I will need a way to apportion addresses in Milwaukee to the correct neighbourhoods. Despite a number of small value (<$500) properties and a smaller number of high value properties (>£20m), I will use a mean average price as I am looking at property values as a proxy for affluence.

A successful outcome will be successfully clustering neighbourhoods in both cities to find the most similar locations in terms of amenities and (relative) house prices.

2 Data

I will be using the following data sources

- A list of historic, mainly central, neighbourhoods of the city of Milwaukee. This can be found here. I simply scraped the names of the neighbourhoods directly.
- Their locations were obtained from Open Street Map's Nominatim service[6] through the Python Geocoder package.

[5] See the Wikipedia page for the London postcode areas.

[6] © OpenStreetMap contributors. Full licence details here.

- A CSV file of real estate transactions in Milwaukee City from the City Council. I used the data from 2018 as this was the most recent data available for both cities.
- To enable allocation of each address to its neighbourhood I had to obtain the latitude and longitude which was accomplished by leveraging the United States Census Bureau through its *Geocoder* API.

  Allocating sales to postcodes is easy in London as every property has an associated postcode in
- HM Land Registry CSV list of property transactions in England and Wales[7].
- To obtain the centroids of postcode areas I used the Ordnance Survey's list of postcode unit locations, part of Code-Point Open. This is available as a free CSV download but requires an email address to receive a download link[8].
- Foursquare's *Places* API.
- The British Geological Survey's web service to convert Eastings/Northings to Latitudes/Longitudes.

3 Methodology

3.1 Data collection and preparation for Milwaukee

I initially scraped the UWM website mentioned above to give a list of Milwaukee neighbourhoods. This list was particularly focused on the older, more historic and more central neighbourhoods that I prefer. There are other lists but this struck a balance between sizes of neighbourhoods and preferred locations.

Location data were given in more than one format which I invariably converted to Latitude/Longitude before use.

I found the location of each neighbourhood using the OpenStreetMap Nominatim API which gave most a location although I had to omit three neighbourhoods that neither I nor OSM knew. I suspect these are historic neighbourhoods that no longer exist. Another neighborhood, Granville Station, is not considered to be part of the city by most and so was also removed.

[7] Contains HM Land Registry data © Crown copyright and database right 2020.
This data is licensed under the Open Government Licence v3.0

[8] © Ordnance Survey Limited, 2020.
This data is licensed under the Open Government Licence v3.0.

Property prices were obtained from the City council, an example record is shown in Table 1.

PropType	Address	Sale_price
Residential	7144 N 38TH ST	129000

Table 1: Milwaukee real estate sales

The 2018 data was the most recent that was available and validated for both cities. I felt it unlikely that neighbourhoods would change significantly in under 24 months.

I needed to allocate each property to a neighbourhood which meant finding the latitude and longitude for each sale. The US Census Bureau provides and API for just this purpose. Once I had the distance to the nearest centre for every sale I excluded those over 1.5 km from the centre of their neighbourhood. This was because sales in outlying areas would sometimes be allocated to neighbourhoods up to 8km away of which they are obviously not part. In Figure 1 the green dots are sales and the blue circles are neighbourhood centres. It should be clear that sales near, for example, Brown Deer are not part of any of the neighbourhoods despite having a closest centre.

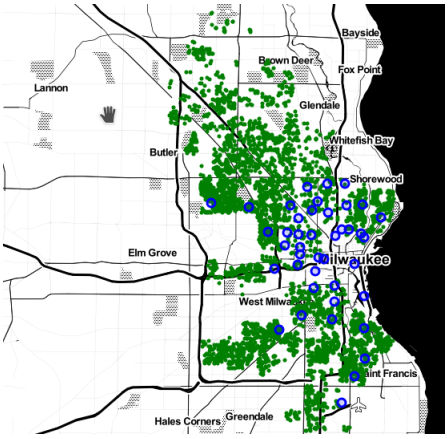


Figure 1: Locations of Milwaukee real estate sales

The neighbourhoods do not cover large parts of the city - but do cover all the more central areas. There were few sales in some areas because these are primarily industrial zones in the modern city and the very centre of the city is dominated by rental properties, flats condos whose prices cannot be easily extracted from the city's data which is about land sales.

I decided to use mean house prices in neighbourhoods despite a small number of properties with very low values e.g. $100. Three neighbourhoods had fewer than 3 sales but I let the values stand as none was ludicrous. Given time I would have gathered sales data from, say 2017–2020, to ensure all neighbourhoods were given more robust average property values.

A 'venue' can be anything from a supermarket to a river to a theatre to a sports-ground. I used the FourSquare API to pull data for the top 100 venues in each neighbourhood (most had less than 100) within 750m of the centre. Each venue has a category of what type of place it is. I extracted the categories and one-hot encoded them for each neighbourhood, taking a mean to get proportions of venue categories in each. Find clusters of similar properties

3.2 Data collection and preparation for London

London postcodes are remarkably regular and just knowing the eight district prefixes and how many postcode areas there are (from Wikipedia) I was able to generate them all. I did not include non-geographic postcode areas, nor E20 — the Olympic Park area.

The Ordnance Survey provides a list of postcodes units with their locations. I took the mean of the Eastings and Northings for each postcode unit and use the British Geographic Society's web servuce to convert these to latitude/longitude format. This meant I had the centre of each postcode area as shown in Figure 2. Although not a geometric centre it should be good enough.

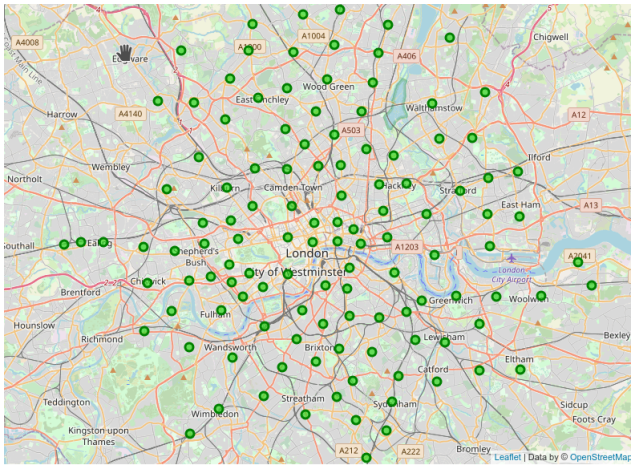


Figure 2: Centres of postcode areas in London

HM Land Registry's list of property sales in England and Wales is a large file.

An example entry could is shown in Table 2 with the irrelevant fields omitted.

Price	Postcode	PropType	ValueType
745000	SW19 2BZ	T	A

Table 2: London property sales

I excluded commercial properties (`PropType==O`) and properties were the true value was not paid (`ValueType==B`) then dropped these fields. I then allocated each postcode area the average of property values sold within it.

As with Milwaukee I pulled in the Foursquare venue list within 750 m, extracted the categories, one-hot encoded them and took means to find the proportions in each postcode area.

Looking at London property prices I realised they were very skewed, as shown in Figure 3.

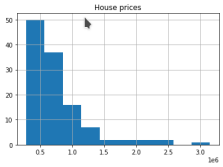


Figure 3: Average house prices in london neighbourhoods (£)

The usual way to fix skew is to transform the data by a transformation from

$$\{\ldots, x^3, x^2, \sqrt{x}, \log x, \frac{1}{\sqrt{x}}, \frac{1}{x}, \frac{1}{x^2}, \ldots\}$$

It turned out in this case that $\log x$ gave the most symmetric distribution as shown in Figure 4.

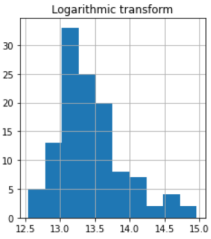


Figure 4: Log of neighbourhood house prices in London

I went back to consider Milwaukee prices and they suffered the same problem, shown in Figure 5 (where house prices were scaled so the average across the city was 1.0). This had the same solution — a logarithmic transformation which resulted in Figure 6.

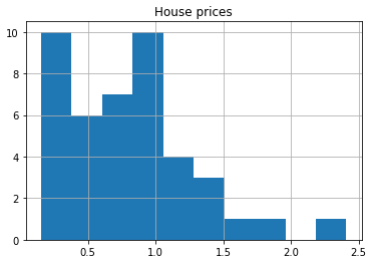


Figure 5: Scaled neighbourhood house prices in Milwaukee

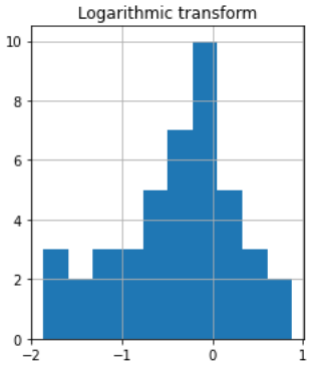


Figure 6: Log of scaled neighbourhood house prices in Milwaukee

3.3 Looking for similar neighbourhoods in each city

I used the same approach in each city:

- Use k-means clustering over a range of values of k.
- Choose the values of k that split the data most evenly with the least number of singleton clusters.
- Map the clusters.
- Produce a silhouette plot to help assess the quality of clustering.

3.4 Comparing all neighbourhoods

In order to compare neighbourhoods from both cities I had to standardise each separately take account of the differences in house values and currencies. I then concatenated the two sets using an inner join so that only the categories they both had at least one of would be used for comparisons. This meant that 39 categories unique to Milwaukee were unused and also 205 items from London were

not compared. Given that there were still nearly 200 categories I felt this would be enough. Nevertheless I was conscious that many neighbourhoods might have 0 of some category in common and removing the category removes that aspect of similarity.

Following this I followed the same procedure as for the individual cities - trying a range of values of k and choosing the ones that split the neighborhoods more evenly, with as few singleton clusters as possible. This time I could obviously not map both sets of neighbourhoods but could still produce silhouette plots.

Having identified the most likely sets of groupings of clusters I combined these to find places similar to where I live or would like to live in both cities.

4 Results

4.1 Milwaukee

I used a black and white map to allow coloured clusters to be easily seen and distinguished. The best number of clusters for Milwaukee was 8 which is mapped in Figure 7. However the silhouette plot in Figure 8 showed that these clusters are

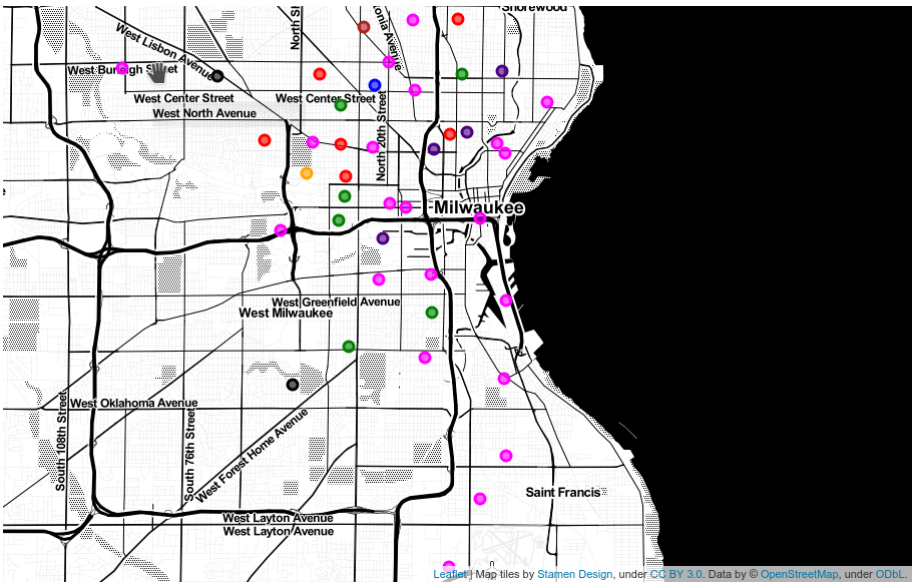


Figure 7: Similar neighbourhoods in Milwaukee

not terribly well-defined. This was backed up by the fact that although usually similar, clustering with random initial positions, occasionally gave very different groups. The ones shown in Figure 7 were quite typical.

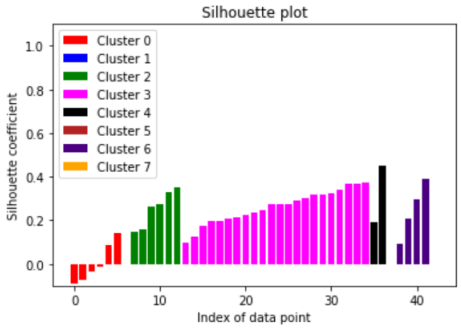


Figure 8: Quality of clusters in Milwaukee

4.2 London

The picture for London was, if anything, worse. The optimum number of clusters seemed to be 12 as shown in Figure 9. While the silhouette plot in Figure 10

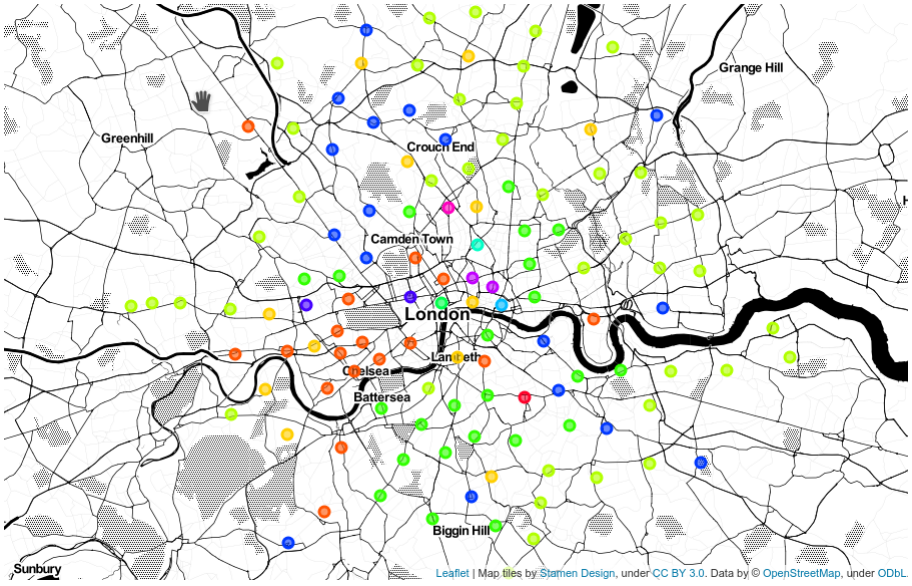


Figure 9: Similar neighbourhoods in London

showed that only one cluster was close to being well-defined.

On the other hand some of these clusters do tally with the postcodes that many might think are similar:

Clusters 1 and 9 are the expensive West London postcodes together with a couple of other affluent areas such as the Docklands(E14) and Wimbledon (SW19).

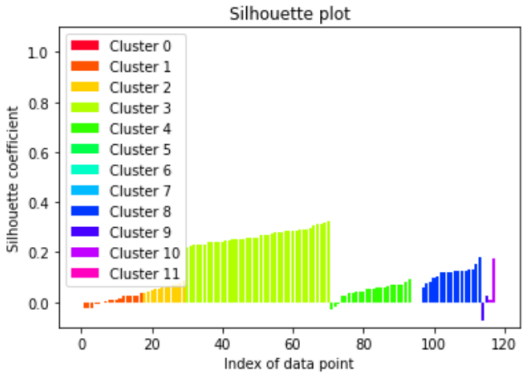


Figure 10: Quality of clustering in London

Cluster 2 are islands of affluence in London especially in the South-West.

Cluster 4 is the less affluent but trendier areas of South and East London.

Cluster 3 almost entirely consists of average residential areas outside travel Zone 3.

Central London is a mix of small and singleton clusters which makes sense as these areas are very different from each other and the rest of London.

Cluster 8 has nothing obvious tying its members together — it may be spurious.

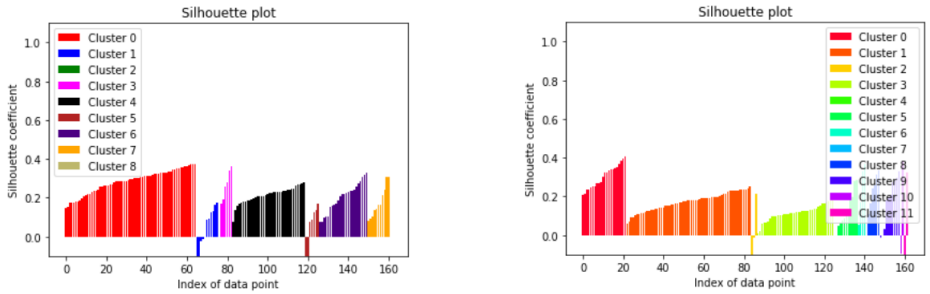
4.3 Both cities

When looking for similar neighbourhoods I had considered clustering on one city and then using a classification algorithm classify neighbourhoods in the second city. The problem with this is the sparse feature matrix across 200 dimensions would make it extremely unreliable. Instead I looked at clustering all neighbourhoods in both concatenated together. This suffers from the same issue of dimensionality but hopefully to a lesser extent.

I again dropped latitudes and longitudes before comparing as they would have certainly made a difference!

Repeatedly looking for good values of k led me to a clustering where k=9 and another where k=12. The silhoutte plots for these are in Figure 11. Neither clustering is stellar but k=9 appears to have better defined clusters than k=12 (which is to be expected).

One interesting is that, in both cases, most clusters included only neighbourhoods from one of the cities.



(a) k = 9 (b) k = 12

Figure 11: Silhoutte plots for k-means clustering of all neighbourhoods

5 Discussion

The scheme used here resulted in a sparse feature matrix for both cities. This is problematic for clustering as *feature = 0* in two neighbourhoods does not really mean they have anything in common other than their *lack* of an amenity. The only consistently non-zero attribute is the average house price. This means that much of the distinctions between clusters might be based more heavily on this than on the other attributes as the other attributes will generally be zero for both neighbourhoods (similar) and only very occasionally non-zero/zero (different).

I am also unsure as to whether London/Milwaukee differences are caused by difference in classification. Is a 'Burger joint' in Milwaukee likely to be categorised as an 'American restaurant' in London etc? There is no easy way to tell. Moreover some places are categorised very specifically for example *Greek restaurant* while others are just *Food*. What is really needed is fewer, broader categories.

With these caveats I there are still some interesting associations when looking at the lists of similar neighbourhoods for example E4 is in a singleton group in both cases. Perhaps there is something very particular about that area of the Lea Valley.

That clusters of all neighbourhoods tended to be from one city or the other suggests that the neighbourhoods are highly dissimilar, even based on features the cities have in common.

The k = 12 clustering in London in Figure 9 was particularly interesting as it separated the city into 5 broad groups:

- Rich West Inner London
- Edgy Inner South and East
- Affluent Suburban (often closer to the centre)

- Regular Suburban (often further from the centre)
- Central London/CBD

The last of these is actually a diverse group of singleton/small clusters geographically close together.

We are now looking at some areas in cluster 4 [9] as places to move next.

6 Conclusion

Milwaukee and London are sufficiently different that it is not possible to compare neighbourhoods between them. Milwaukee neighbourhoods are more similar to each other than London postcodes and hence are more difficult to categorise.

I am not sure that the approach here is likely to work well and is not easily savleagable. What is needed is fewer categories with more eveny dispersed values - crime rates, tax levels, pollution levels, etc. Foursquare's feature data is not amenable to clustering.

However some personally useful information has been gleaned and that has certainly made it worthwhile.

[9] SW2, SW4, SW9, SW12, SW17, SE4, SE5, SE8, SE24, E1, E2, E8, E9 and N16