Access to low-cost food in the Washington, DC region

A. Introduction

Problem considered

Observers have noted that persons living in low-income areas may have comparatively low access to supermarkets, and instead have to rely on higher-cost and lower-quality food options such as fast food and convenience stores:

> The presence of supermarkets/grocery stores in a neighbourhood is associated with buying and consuming healthier food. Grocery stores typically sell healthier food items at affordable prices compared with convenience stores and fast-food outlets. Wide ranges of fresh food choices, with relatively lower price points, frequent availability and visually appealing presentation, are associated with choosing healthier food options and healthier diets.[1]

This analysis explored the ability of Foursquare data to shed light on whether those types of food access issues are present in the Washington, DC region.

Significance

Given the obvious social and health importance of access to affordable food, the results from this types of analysis could have public policy significance for local governments and/or food programs. The results in theory also may be of value for the private sector by identifying underserved markets.

B. Data and processing techniques

Overview

The analysis used Foursquare venue data to examine the distribution of low-cost food options across the Washington, DC region, in conjunction with clustering techniques to more rigorously compare food access across local areas.

Specific elements of the project included: (i) defining geographic limits; (ii) extracting and cleaning Foursquare data to reflect select categories of low-cost food vendors; (iii) geographic grouping of the resulting vendors; and (iv) clustering the geographic zones to examine the local availability of low-cost food options.

Extracting and cleaning Foursquare data

Data extraction

Foursquare venue data was extracted using a grid search over the District of Columbia and suburban Maryland and Virginia.[2] Data was obtained regarding four types of low-cost food vendors – supermarkets, convenience stores, fast-food restaurants and dollar stores. In each case, data initially was extracted based on the applicable Foursquare categories, and then was cleaned to more accurately target those categories.[3]

---

1 Abeykoon, Stringer and Muhajarine, "Health-related outcomes of new grocery store interventions: a systematic review" (footnotes omitted), Public Health Nutrition: 20(12), 2236-48 (available at https://www.researchgate.net/profile/Nazeem_Muhajarine/publication/317287012_Health-related_outcomes_of_new_grocery_store_interventions_A_systematic_review/links/5c2cdb87299bf12be3a83e48/Health-related-outcomes-of-new-grocery-store-interventions-A-systematic-review.pdf).

2 The notebook for creating the grid squares, and the resulting csv file, respectively are found at:

https://github.com/kansjs/IBM_applied_capstone_share/blob/master/DMV%20grid%202.15.20.ipynb
https://raw.githubusercontent.com/kansjs/IBM_applied_capstone_share/master/DMV_grid2.csv

3 Following clean-up, the data contained a total of 303 supermarkets, 1096 fast food restaurants, 950 convenience stores and 115 dollar stores.

The notebooks for extracting the raw data for the supermarket, convenience store, fast-food restaurant and dollar stores categories respectively are found at:
https://github.com/kansjs/IBM_applied_capstone_share/blob/master/Foursquare%20supermarkets%202.15.20.ipynb
https://github.com/kansjs/IBM_applied_capstone_share/blob/master/Foursquare%20convenience%202.15.20.ipynb
https://github.com/kansjs/IBM_applied_capstone_share/blob/master/Foursquare%20fast%20food%202.15.20.ipynb

Geographic groupings, and clustering based on food availability

The resulting 2464 venues were grouped into 50 local zones, using K-means clustering based on distances between venues.

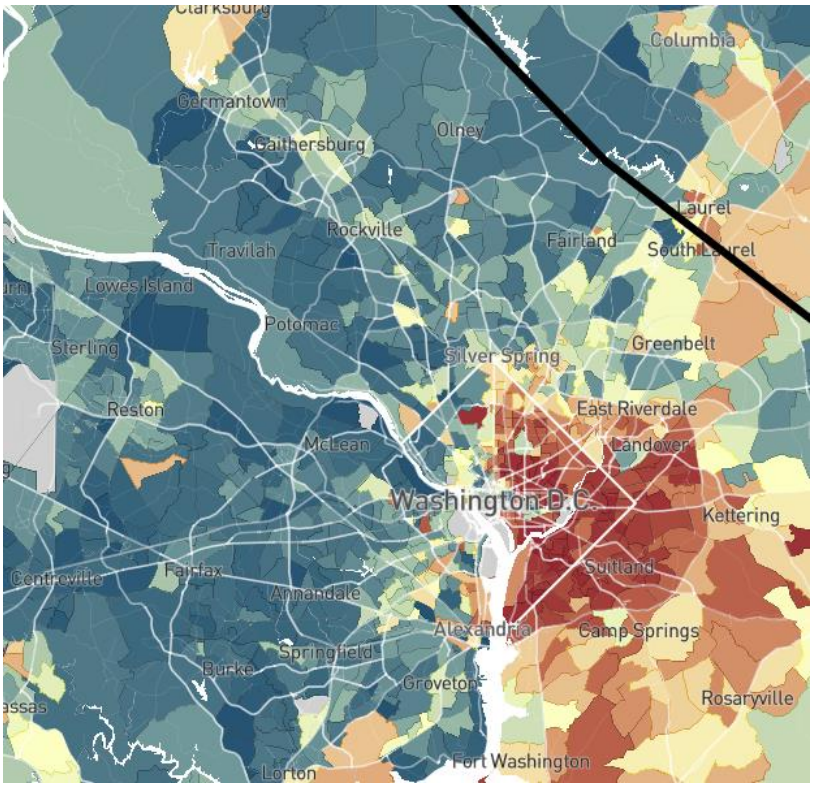
For each local zones, the analysis calculated the percentage of venues that were supermarkets, fast-food restaurants, convenience stores and dollar stores. Based on those relative percentages, the zones were clustered using K-means techniques.[4]

C. Results and significance

Distribution of food venues

Figure 1 depicts external mapping of income levels in the Washington, DC area.

Figure 1 – Relative income in the Washington, DC area



Source: The Opportunity Atlas (https://www.opportunityatlas.org/)

https://github.com/kansjs/IBM_applied_capstone_share/blob/master/Foursquare%20discount%202.15.20.ipynb

The associated csv files respectively are found at:
https://raw.githubusercontent.com/kansjs/IBM_applied_capstone_share/master/Supermarkets.2.8.20.csv
https://raw.githubusercontent.com/kansjs/IBM_applied_capstone_share/master/Convenience.2.9.20.csv
https://raw.githubusercontent.com/kansjs/IBM_applied_capstone_share/master/Fast_food.2.8.20.csv
https://raw.githubusercontent.com/kansjs/IBM_applied_capstone_share/master/Discount.2.9.20.csv

The notebooks for the cleaning and combining of the data, and the associated csv file, respectively are found at:
https://github.com/kansjs/IBM_applied_capstone_share/blob/master/Combined%20DMV%202.15.20%20B.ipynb
https://raw.githubusercontent.com/kansjs/IBM_applied_capstone_share/master/DMV_food%202.9.20.csv

[4] The notebook for the geographic grouping and the clustering is found at:
https://github.com/kansjs/IBM_applied_capstone_share/blob/master/DMV%20processed%202.16.20%20B.ipynb

Lower-income levels are red-maroon, and higher-income levels are blue-green. The question at hand is whether the higher-income areas have a comparatively higher preponderance of supermarkets and a comparatively lower preponderance of lower-quality food vendors than the lower-income areas.

Figures 2 through 5 are heatmaps, derived from Foursquare venue data, reflecting the distribution of supermarkets, fast-food restaurants, convenience stores and dollar stores. Those figures suggest that there are differences in the comparative distribution of those types of food vendors, but the figures alone are not sufficiently clear-cut to draw conclusions regarding the comparative availability of low-cost food vendors. In part, the issue is complicated by the fact that the locations of fast food restaurants and convenience stores appear to be influenced by the locations of office and tourist areas.

Figure 2 – Supermarket distribution

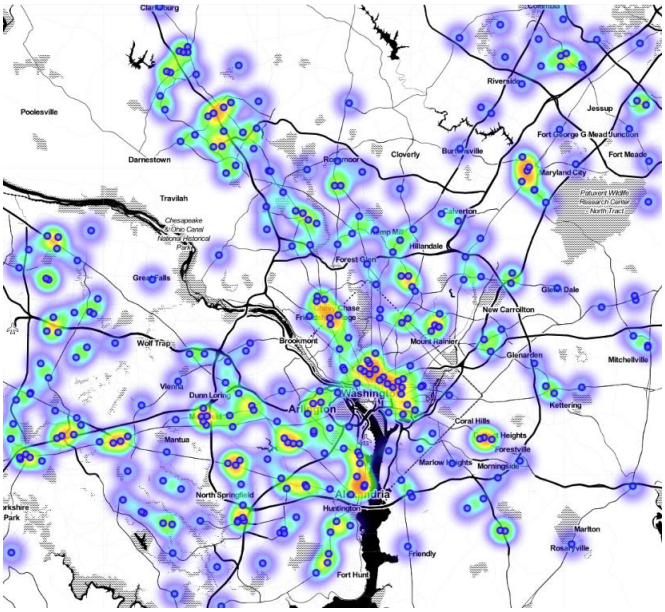


Figure 3 – Fast food distribution

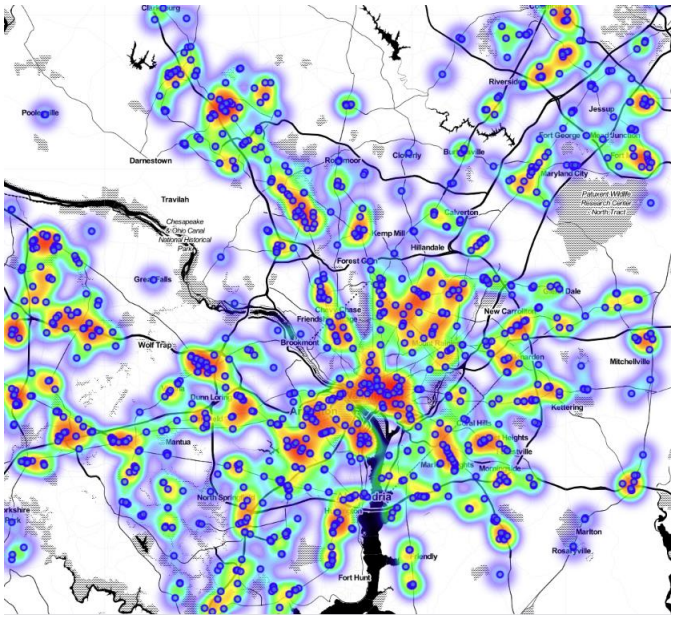


Figure 4 – Convenience store distribution

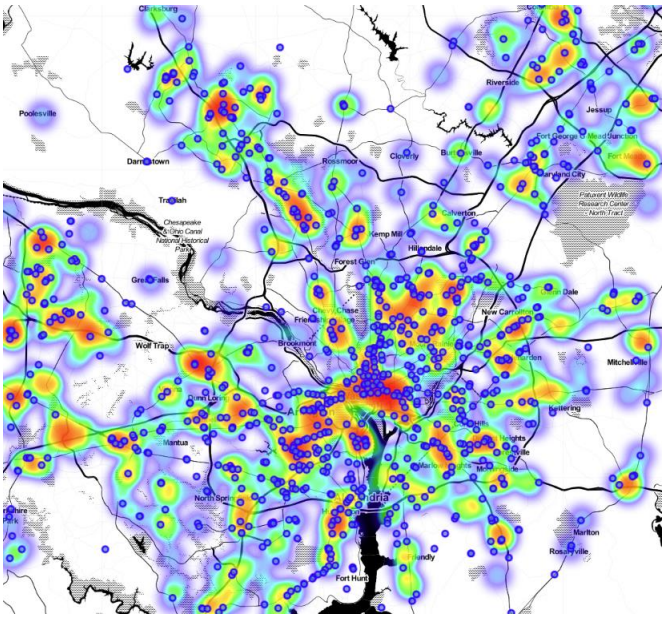
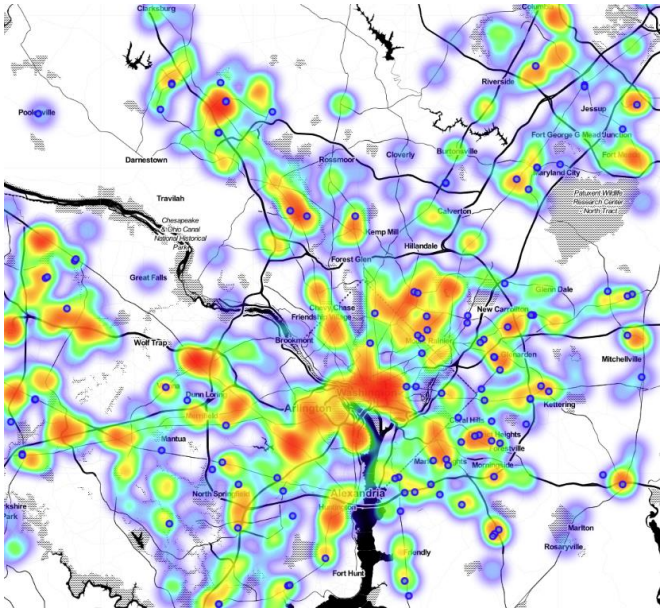


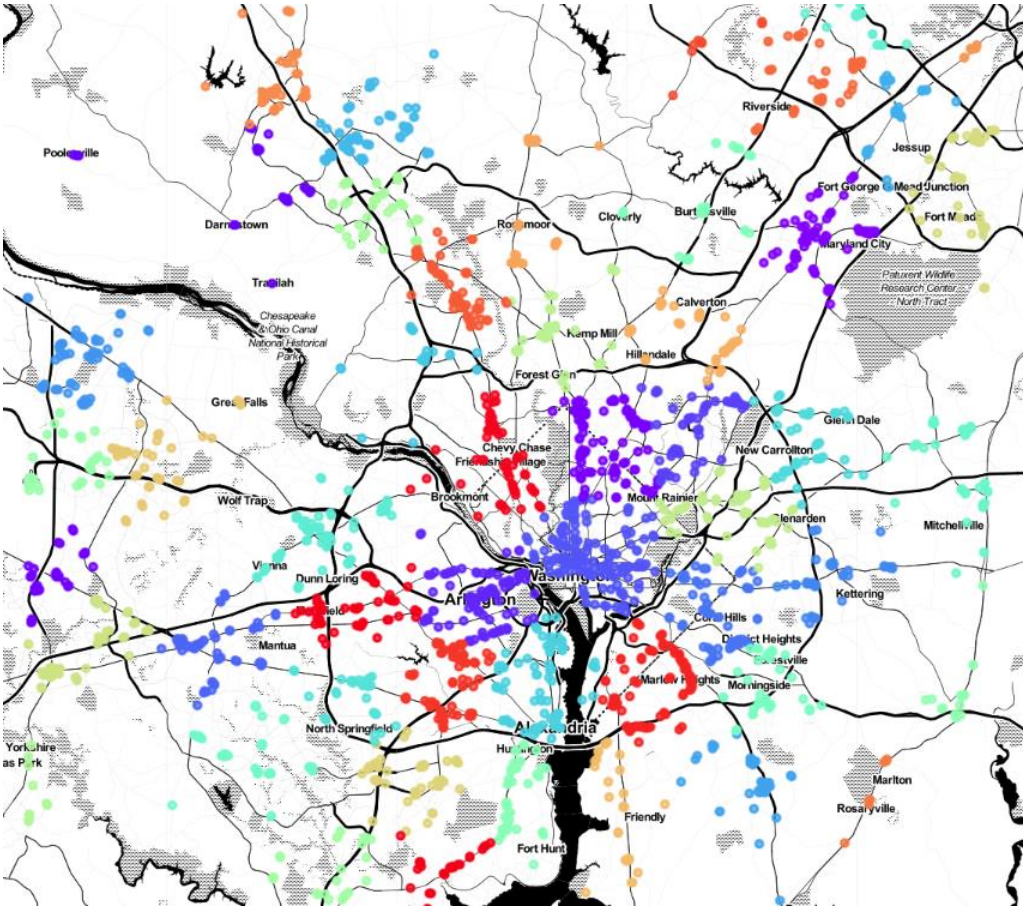
Figure 5 – Dollar store distribution



Grouping and clustering

The 2464 venues were allocated across a total of 50 local zones, with the view toward identifying groupings that reflect reasonably convenient food options for persons in the Washington, DC region. Those zones are depicted in Figure 6.

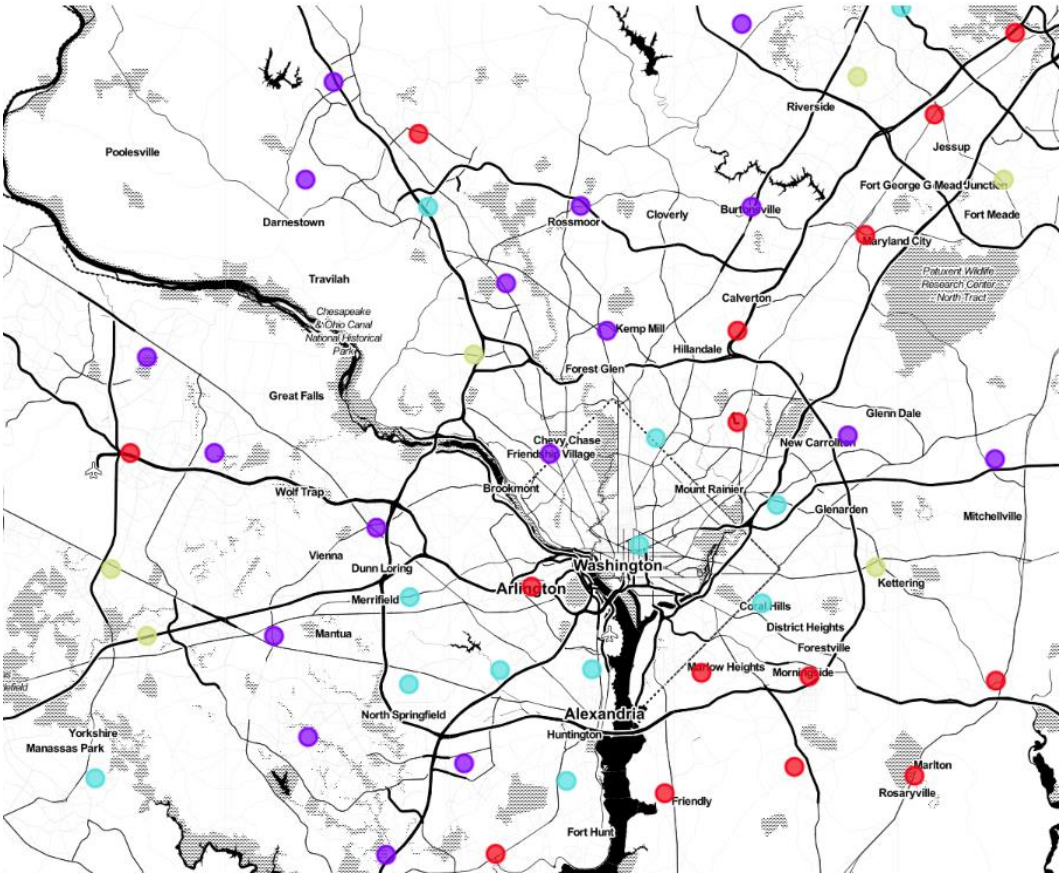
Figure 6 – Geographic groupings



The geographic zones were based on K-Means clustering of the physical distance among venues – a method that was sub-optimal in that it led to uneven cluster sizes. Attempts to rebalance the geographic clusters showed some promise, but were not completed.

For each geographic zone, the analysis then calculated the relative portion of supermarkets, fast food restaurants, convenience stores and dollar stores. Based on those relative portions, K-means clustering was used to allocate the geographic groupings across four clusters, in an attempt to identify regions that have similar access to low-cost food options. Figure 7 shows the resulting clustering.

Figure 7 – Clustering based on availability of food options



Cluster 0 – red; cluster 1 – purple; cluster 2 – cyan; cluster 3 – light green

In general, clusters 0 and 2 tend to correspond with the lower income areas depicted in Figure 1, while clusters 1 and 3 tend to correspond with the higher income areas – but, significantly, there are exceptions in both directions. Table 1 provides further information regarding those clusters.

Table 1 – Food availability clusters

	Cluster 0		Cluster 1		Cluster 2		Cluster 3	
	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev
Convenience	0.425	0.030	0.309	0.034	0.485	0.041	0.197	0.026
Dollar	0.054	0.043	0.044	0.038	0.041	0.041	0.049	0.042
Fast Food	0.441	0.029	0.503	0.030	0.350	0.030	0.602	0.043
Supermarket	0.080	0.028	0.144	0.039	0.124	0.049	0.152	0.053

Application of Student's t-test suggests that the proportion of supermarkets in Cluster 0 has a statistically different mean from the other three clusters. On the other hand, application of that test indicates that we cannot reject the hypothesis that the proportion of supermarkets across Clusters 1, 2 and/or 3 have the same mean.

D. Conclusions and next steps

This preliminary analysis suggests that Foursquare venue data may serve as a useful tool for examining the availability of low-cost food vendors among various local areas in the Washington, DC region. Several enhancements may be expected to improve the analysis, including: (i) developing refined geographic groupings that are more precisely linked to locality data regarding income (as well as other relevant criteria, including health-related criteria); (ii) accounting for the fact that the distribution of fast food restaurants and convenience stores reflects the distribution of offices and tourist areas as well as the distribution of residences; and (iii) if practicable, accounting for other types of food vendors.