***Healthy dollar store models and community demographics***

***Research proposal***

PI: Dr. Jerry Shannon, University of Georgia

Co-reseacher: Maya Rao, University of Georgia

**Overview**

The purpose of this research is to identify where two major dollar store chains, Dollar General and Family Dollar, offer store models with enhanced access to fresh produce and assess whether there is a difference in the sociodemographic characteristics of the neighborhoods in which these dollar store chains offer vs. do not offer fresh produce. This research builds on existing research examining dollar store locational strategies and their relationship to neighborhood sociodemographic characteristics and the broader landscape of food retail.

**Research Questions**

Overarching question: Is there a difference in the sociodemographic characteristics of the neighborhoods in which dollar store chains offer vs. do not offer fresh produce?

***Sub-questions***

1. Where are the hot and cold spots for census tracts with comparatively high access to dollar stores offering produce for each chain?
2. What are the demographic characteristics of census tracts in these hot/cold spots and how do they differ?

**Data**

Retailer listings would be obtained from Safegraph, a POI provider who also includes information on store traffic patterns based on cell phone data. For Dollar General, stores offering fresh produce includes both Dollar General Market and retailers identified on the corporate website as providing fresh produce. For Family Dollar, it is only the latter. Dollar General Market retailers would be identified via name from these data.

The most difficult aspect of this analysis is obtaining the list of retailers offering fresh produce. To do this, we use the RSelenium package in R to scrape the amenities offered by each retail location on the corporate websites. This tool opens a new browser window within the R environment and navigates to a directed location, in this case the retailer websites. It then automatically pulls information from these pages, such as store information and listed availability of fresh produce. In hand-testing the reliability of this method in three Georgia cities, we found 100% accuracy.

Demographic information will be obtained from the U.S. Census American Community Survey, downloaded through the Census API.

**Methods**

***Question 1: Where are the hot and cold spots for census tracts with comparatively high access to dollar stores offering produce for each chain?***

This section of analysis is based on a calculation of store proximity, measured through travel times. To calculate this figure, we will use the Mapbox Directions API to identify drive times from census tract centroids to the closest and third closest dollar retailers from each chain as well as the same two figures for retailers that also feature fresh produce. We will obtain these numbers by first identifying the five retailers with closest Euclidean proximity (straight-line distance) via computational methods and then querying the travel time to each retailer using the API. This is a cost effective approach that has proved effective in past research.

We will use these two figures to set up a proximity ratio (PR):

Here, PR is the proximity ratio, dtp is the drive time to a retailer from a given chain including fresh produce, and dta is the drive time to any retailer in the chain. This number will approach 1 if all retailers offer fresh produce, but a higher figure will indicate relatively greater distances to retailers with these items. We use both the closest and third closest retailer to assess density as well as basic proximity. Each of these two PRs (closest and third-closest store) will be standardized to a z-score, where the mean value is 0 and each one-unit change is equivalent to a standard deviation from the mean. This aids in map interpretation.

We will first use exploratory analysis to assess patterns in the composite PR, creating maps to visualize the overall pattern and calculating descriptive statistics. For analytical purposes, we will use the Getis-Ord Gi\* statistic to identify tracts in hot and cold spots for these values. Getis-Ord Gi\* is a statistic that compares the mean of a variable in a given areal unit (census tracts in this case) combined with the value of its neighbors to the result of an entire entire region. Areas with statistically significant high or low values, identified by converting these means to Z-scores, are labeled as hot and cold spots respectively. We will create these hot spots maps for both proximity ratio measures (closest store and third closest store).

See the sample map in Figure 1 below for an example of results from this approach. In this case, the map shows hot and cold spots for teen birth rates by county in the United States. The cold spots (blue) identify areas where the value of a county AND the value of its neighbors are significantly lower than the mean for all counties. Likewise, the the hot spots (red) show areas where a county and its neighbors are significantly higher. The values in the legend refer to Z-scores, and so the cutoff for these identifications reflects statistically significant cut-offs: 1.65 for 90% confidence, 1.96 for 95%, etc.

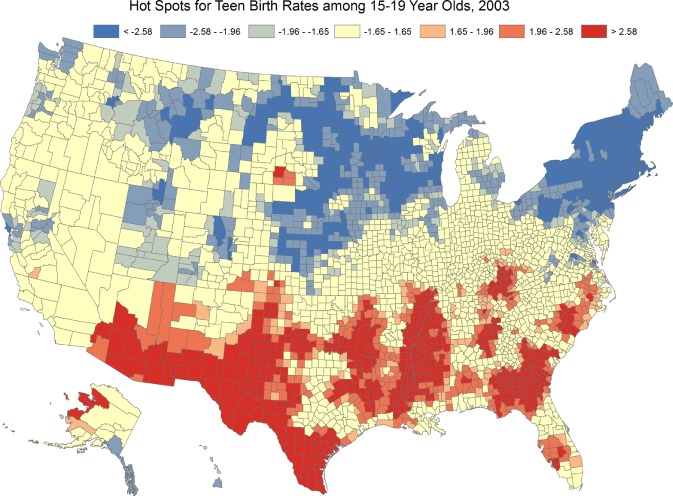


Figure 1: Sample Getis-Ord Gi\* hotspot map. From Khan, D., Rossen, L. M., Hamilton, B. E., He, Y., Wei, R., & Dienes, E. (2017). Hot spots, cluster detection and spatial outlier analysis of teen birth rates in the US, 2003–2012. *Spatial and spatio-temporal epidemiology*, *21*, 67-75.

***Question 2: What are the demographic characteristics of census tracts in these hot/cold spots and how do they differ?***

We will use the American Community Survey pooled five-year sample for 2018-22 to download data on several socioeconomic variables include household income, poverty rate, racial identification, car ownership, and urban/rural status. The mean/median values of each of the continuous variables will be calculated for identified hot and cold spots, with a t-test used to assess statistical difference. We will also create a regression model with the standardized PR as the dependent variable and the demographic data as the independent variables to see what factors are associated with greater relative access. Additional variables may also be included after exploratory analysis if warranted.

**Preliminary findings**

Based on an initial scrape of all stores within Georgia, we created the maps shown below in Figure 2, with stores offering produce shown in green. Dollar General Market locations, which are a different designation than those with fresh produce, are not included but will be in the final analysis.

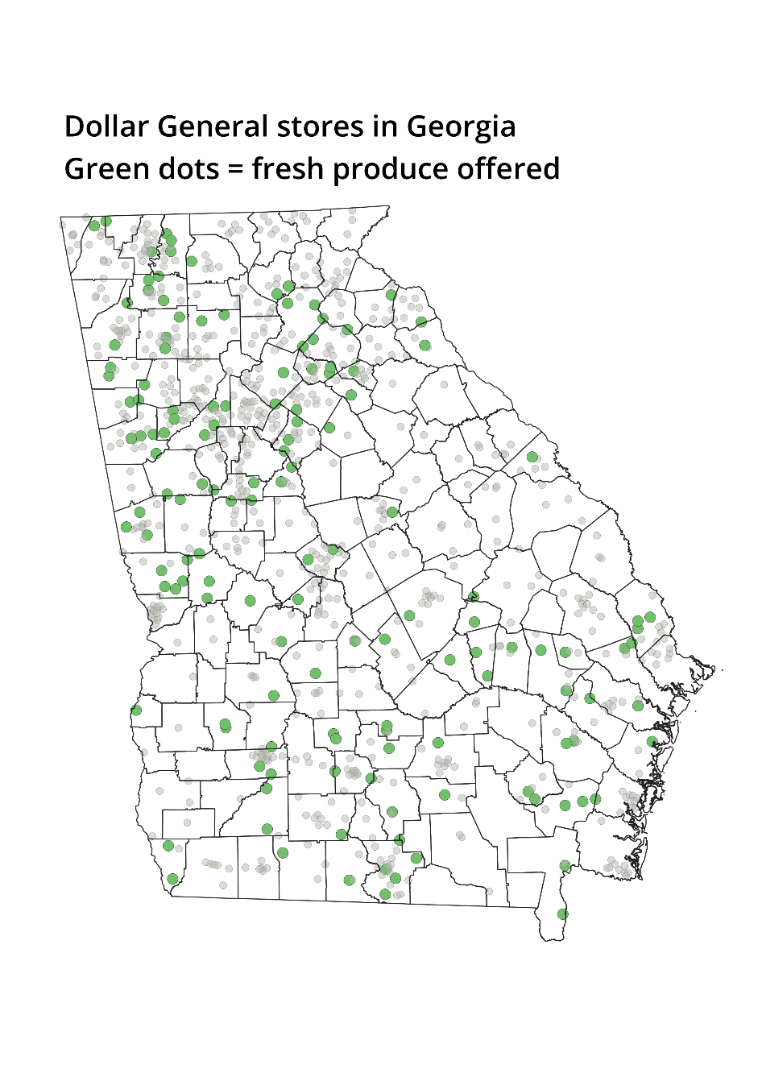
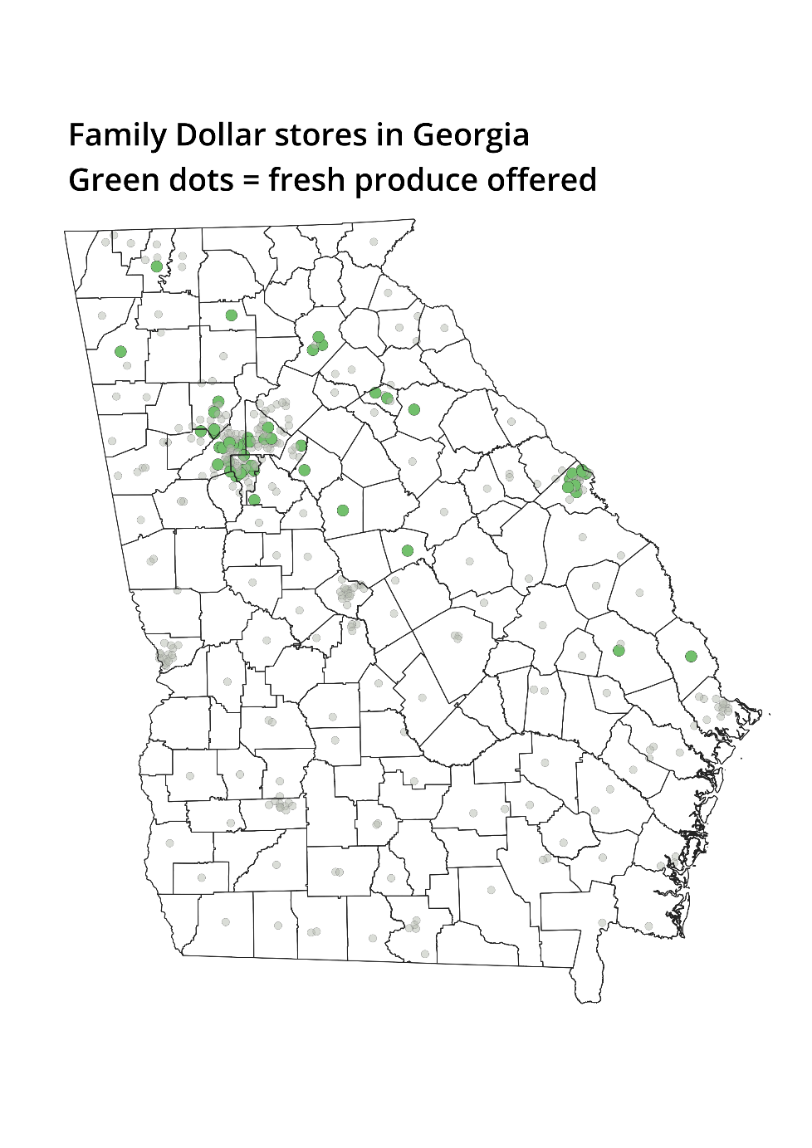


Figure 2: Dollar General and Family Dollar stores in Georgia based on website store listings

Based on this initial data, we calculate the proximity ratio for census tracts just in the core part of Atlanta (Figure 3). The map below shows Dollar General stores, with those offering produce filled in with green. The darker colors represent high proximity ratio values, where locations with fresh produce have much longer travel times than the nearest location of any store. The map on the left shows a ratio based on travel times to the third closest stores in both groups. The geographic pattern in both maps is similar, but much more pronounced for the third closest store, with clusters of high values in parts the east, south, and northwest metro.

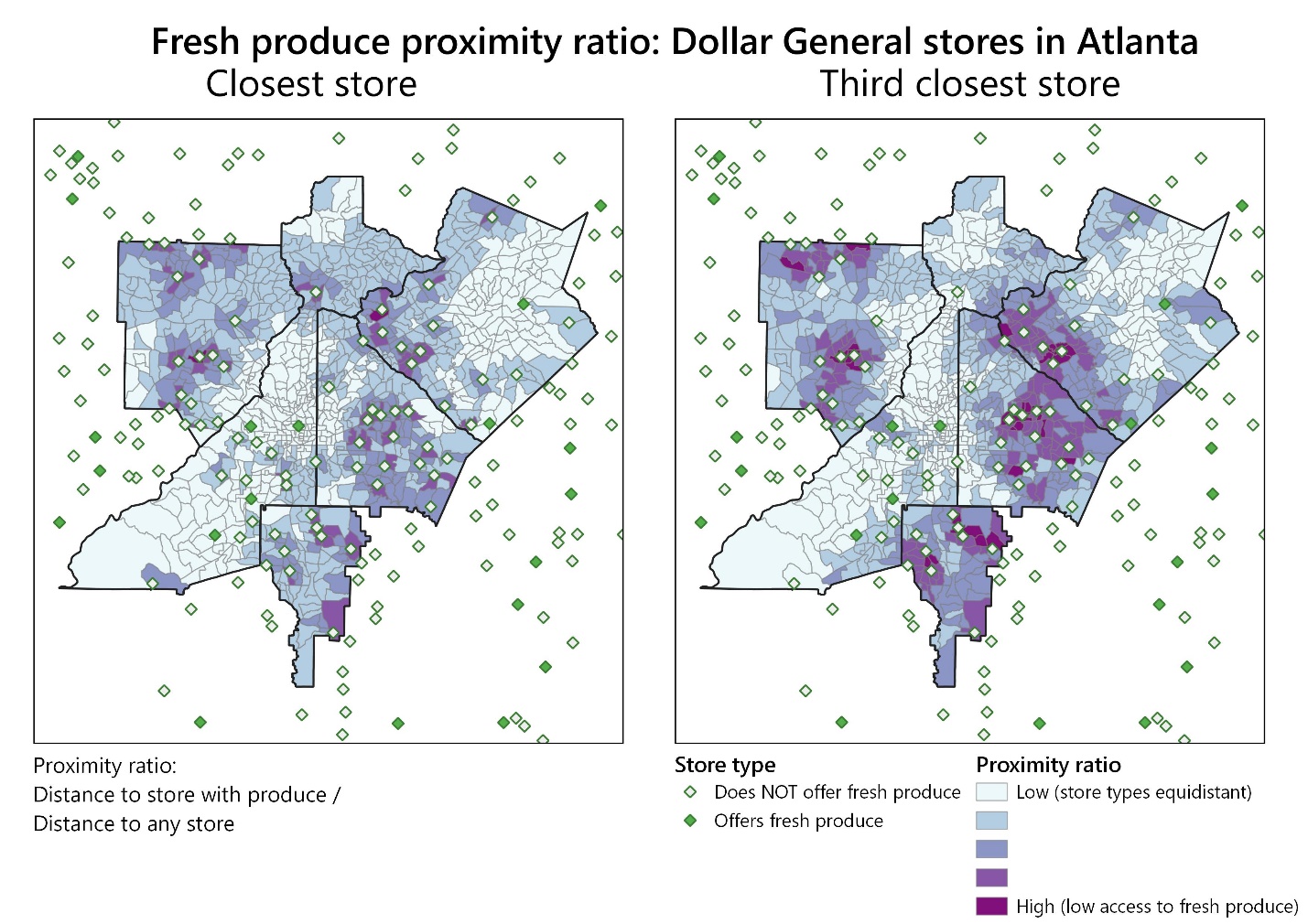


Figure 3: Proximity ratio for the core Atlanta metro for Dollar General.

**Timeline**

Start date: 5/15/23

Data collection complete: 6/15/23

Analysis complete by 7/15/23

Final report submission: 10/1/23

**Costs**

Safegraph

* **$150** for one month subscription to download data

Mapbox Directions API (for drive times):

* 85,000 census tracts \* 5 queries per retailer type \* 5 retailer types (DT/FD with/without produce + supermarkets) = 2,125,000 queries. ([pricing guide](https://www.mapbox.com/pricing))
* **$3,000** total for these queries

Remote server for data collection

* Three month subscription to [Posit Cloud](https://posit.cloud/plans): $75 \* 3 = **$225**

Consultant payroll

* Dr. Jerry Shannon (~40 hours @ $175/hour): **$7,000**
* Maya Rao (~60 hours @ $60/hr): **$3,600**

Total funding requested: **$13,375**